The Unofficial LinkedIn Algorithm Guide

Not endorsed, approved, or reviewed by LinkedIn!

Q1 2026 Edition

TRUST INSIGHTS

The Unofficial LinkedIn Algorithm Guide, Q1 2026 Edition

Introduction

If you've spent more than five minutes on LinkedIn in the last year, you have likely seen one or more "gurus" making definitive claims that they've "cracked the new algorithm." They'll tell you the magic number of comments to leave, the exact time to post, or the one type of content that gets "10x reach." Comment on their post within the first hour, they promise, and they'll sell you the secret to boosting your performance on LinkedIn.

For a long time, that advice, while often oversimplified, pointed in the right direction. It rested on the idea of a complex, multi-stage pipeline of machine learning models that processed signals. A like served as a signal. A comment carried more weight. A keyword functioned as a signal. The game centered on sending the best signals to a sophisticated but ultimately mechanical system.

That's not how LinkedIn works anymore.

LinkedIn's March 2026 Architecture Announcement: What Changed and Why It Matters

In March 2026, LinkedIn's Engineering Blog published a landmark announcement: the company was rolling out "a new advanced ranking system, powered by LLMs and GPUs, that better understands what a post is actually about and how it relates to a member's evolving interests and career goals." Written by engineering lead Hristo Danchev and supported by simultaneous academic papers from LinkedIn's AI researchers, this announcement marked the public reveal of an architecture transformation years in the making.

LinkedIn had quietly replaced its traditional feed infrastructure — a "heterogeneous architecture" of trending content systems, collaborative filtering, historical feed indices, topic-based retrieval, and embedding-based systems, each maintaining separate infrastructure — with two new AI systems designed to work in concert. As Danchev explained: "After some experimentation we converged onto a hybrid complementary relevance solution — a unified retrieval system leveraging advances in LLMs to generate a high-quality representation of our members and content, and a sequential ranking model that captures how professionals engage with content over time."

The traditional approach had two fundamental limitations. For retrieval, separate specialized systems created engineering complexity and made holistic optimization difficult. For ranking, the prior model "treated each impression independently, missing the sequential patterns in how professionals actually consume content over time." The new architecture addressed both limitations simultaneously.

What This Means for Different Users

The practical impact of this architectural shift varies significantly depending on your LinkedIn profile:

- **Members with established engagement histories** experience a feed that reflects their professional trajectory — topics they've been consistently engaging with shape what surfaces, not just what's fresh. The sequential model explicitly captures "how you engage with content over time" rather than treating each visit independently.
- **Newer members and those with fewer connections** benefit most from LLM-powered retrieval. Where prior systems required extensive interaction history, the Causal LLM can "assess latent interests from world knowledge" — inferring relevant content from your profile alone, even at cold start. LinkedIn's retrieval research shows a +1.17% increase in Daily Unique Professional Interactions for these users.
- **All members** benefit from unified retrieval: a single embedding system that replaced five or more separate content sources, reducing noise and improving semantic precision across the entire feed.

This architecture operates within FishDB's hard 30-day content window for connection-based content — that retrieval system cannot surface posts older than 30 days regardless of relevance score. The Causal LLM out-of-network path operates under different constraints.

LinkedIn hasn't just upgraded its engine; it has rebuilt its recommendation architecture around two complementary AI systems: an LLM-powered retrieval system and a sequential transformer ranking system. Together, they function as the new intelligence layer of the platform. This guide, synthesized from LinkedIn's own engineering research — including the March 2026 Engineering Blog announcement and two landmark publications from early 2026 — explains exactly how these systems work and what they mean for your strategy.

The New Premise: Your Language and Your Behavior Both Determine Your Success

This fundamental shift changes everything for marketers, creators, and professionals on the platform. The old game of sending numerical signals to a mechanical system is over. The new game has two dimensions:

Dimension 1: Language quality determines retrieval. The Causal LLM reads your profile and content as text. How clearly and precisely you express your professional expertise — in your headline, your About section, your experience descriptions, your posts — determines which content opportunities the system surfaces for you, and which audiences your content reaches. Your prose is a direct input to the retrieval engine.

Dimension 2: Engagement patterns determine ranking. The Generative Recommender doesn't read your posts as text — it observes the behavioral patterns your content generates. What people engage with, how deeply, and whether those engagements come from the right professional communities — these patterns are what the sequential ranker learns from. Your engagement history, and the engagement history of your readers, is the data the ranking model processes.

Two core capabilities power this new reality:

- **Zero-Shot Reasoning (Retrieval Stage):** The Causal LLM can understand and reason about concepts that LinkedIn's engineers never explicitly trained it on. Because it understands language, it can see a new job title like "Chief Metaverse Officer" and infer its likely seniority, relevant skills, and industry context, even if it has never encountered that exact title before. The system no longer needs a feature called job_title_id = 9875.

- **Sequential Pattern Learning (Ranking Stage):** The Generative Recommender processes your interaction history as a chronological sequence, ordered from oldest to newest, with candidates appended at the end for scoring. The model uses a training-time recency weighting mechanism (exponential loss decay) that gives the most recent position in the sequence full weight while progressively down-weighting earlier positions — meaning recent interactions carry substantially more influence over what ranks next in your feed. The practical effect: recent, topically consistent engagement shapes what rises to the top of your feed. When you consistently engage with posts about a specific domain, that behavioral pattern becomes the ranker's primary signal about your professional interests. Your engagement habits are a form of continuous communication with the ranking system.

With this understanding, we can establish the new guiding principles for success on LinkedIn. Think of your presence on the platform as operating across two dimensions simultaneously: language that the retrieval system reads, and behavior that the ranking system learns from.

- **Your Profile is the Dossier's Executive Summary.** The Causal LLM reads it as text. Your headline is not just a collection of keywords; it is the opening statement of your professional story. Your About section is no longer optional filler; it is the abstract that provides essential narrative and context for everything else. Your Experience descriptions are the evidence, the case studies that prove your expertise. The prose you use to describe your accomplishments is a direct, primary input to the retrieval engine. Additionally, a fine-tuned Qwen3 0.6B model converts your profile into embeddings that the ranking model also uses — making profile quality doubly important, affecting both retrieval eligibility and ranking context.

- **Your Content is the Case Study.** The Causal LLM evaluates each post you create during retrieval for topical relevance and semantic coherence. It then compares your post's embedding against member embeddings across the network. A well-written, insightful post that clearly articulates a unique perspective creates a sharper embedding — one that matches more precisely with the right professional communities. Beyond retrieval, your content's ability to generate sustained, genuine engagement determines how the sequential ranker learns from the behavioral patterns it produces.

- **Your Engagement is the Behavioral Curriculum.** Your likes, comments, and shares are no longer just simple positive or negative signals. They are the interaction sequence that the Generative Recommender processes. When you consistently engage with expert-level content on a specific topic, you are building a behavioral pattern that the sequential model assigns increasing weight over time. Recent engagements matter most — the model's training explicitly up-weights the most recent interactions in the sequence and progressively down-weights older ones through exponential decay. Your engagement choices actively shape the behavioral fingerprint the ranking model uses to determine what appears in your feed.

How We Know What We Know

In this totally unofficial guide, which no one at LinkedIn has endorsed, we have synthesized a body of engineering research, academic papers, and conference presentations from LinkedIn's own AI researchers. We've used generative AI to synthesize sources that describe this paradigm shift in technical detail.

LinkedIn engineers themselves published the three most significant sources for this Spring 2026 edition:

- **arXiv:2510.14223v1** (Ramanujam et al., October 2025): "Large Scale Retrieval for the LinkedIn Feed using Causal Language Models" — the Causal LLM retrieval system in full technical detail.
- **arXiv:2602.12354v1** (Hertel, Srivastava, et al., February 2026): "An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking" — the Feed SR / Generative Recommender paper, describing the production feed ranker, including the explicit evaluation and rejection of text-prompt LLM ranking.
- **"Engineering the next generation of LinkedIn's Feed"** (Danchev, March 12, 2026): LinkedIn's Engineering Blog public announcement naming the Generative Recommender (GR) as the production ranking model.

LinkedIn has not endorsed or approved this guide, and we at Trust Insights — not LinkedIn — hold and express the views in this guide. We ground every technical claim in this guide in these published research papers and engineering blog posts, which we link in our References section.

The three toolkits — comprehensive checklists covering your profile, content strategy, and engagement approach — are available directly after the Start Here navigation page, or immediately following the technical walkthrough in Sections 2–4. The Start Here page will help you decide which path fits your goals and available time. You can copy and paste all three checklists into your favorite generative AI tool, and we have completely re-framed each one to align with this new, two-dimensional paradigm:

The LinkedIn Profile Checklist

Use this toolkit to transform your profile from a list of data points into a compelling professional narrative — one that communicates your expertise clearly to the retrieval LLM and provides rich contextual signal for the ranking model's profile embeddings.

The LinkedIn Content Pre-Launch Checklist

Use this toolkit to craft content that goes beyond keyword optimization to achieve semantic clarity, well-reasoned arguments, and conversation starters — the qualities that both help your content pass the retrieval gate and generate the genuine engagement that the sequential ranker learns from.

The LinkedIn Engagement Checklist

Use this toolkit to guide your daily and weekly activities, helping you build a topically consistent behavioral sequence that communicates your professional identity through the patterns of your engagement — the data the Generative Recommender processes most directly.

The age of chasing algorithmic hacks is over. The age of clear, compelling, and valuable communication — paired with intentional, strategically consistent engagement — has begun. This guide will show you how to thrive in it.

Got Questions?

This guide comes with a NotebookLM instance you can interactively ask questions from: https://notebooklm.google.com/notebook/f6e5efc0-bbdb-492f-a97e-4fc9e675ea46

Our Additional Resources

If you want to deepen your AI marketing capabilities, we offer several ways to work together:

Learn On Your Own:

- *Almost Timeless: 48 Foundation Principles of Generative AI*: Cofounder and Chief Data Scientist Christopher Penn wrote this non-technical AI book to help you think about AI and apply it to your organization.

Learn With Us:

- AI-Ready Strategist: CMOs and C-Suite leaders learn frameworks and methods for developing, deploying, and managing AI at any scale, from the smallest NGO to the largest enterprises, with an emphasis on people, process, and governance.
- GEO 101 for Marketers: A short 90 minute course on how GEO works and how to structure your marketing strategy to work with AI, not against it.
- Generative AI Use Cases for Marketers: Explore the 7 major use case categories for generative AI in marketing through 21 different hands-on exercises, with all data and prompts provided.
- Mastering Prompt Engineering for Marketing: Build the foundation skills you need to succeed with generative AI, including 3 major prompt frameworks, advanced prompting techniques, and how to choose different kinds of prompts based on the task and tool.

Let Us Help You:

- Customized consulting: If you value the promise of analytics, data science, and AI but prefer not to handle the heavy lifting — from data governance to agentic AI deployment — we can do it for you. We bring more than a decade of real-world AI implementation (AI existed long before ChatGPT) to your foundational data so you can realize the benefits of AI while your competitors are still figuring out how to prompt.
- Keynote talks and workshops: Invite us to your event. We offer customized keynotes and workshops for conferences, company retreats, executive leadership meetings, annual meetings, and roundtables. We customize every full-fee talk for your event, industry, or company, and you receive the talk recording and materials (transcripts, prompts, data) for your audience to work with and learn from.

Start Here: How to Use This Guide

This guide draws on more than 30 current LinkedIn engineering publications — including arXiv papers from 2025–2026 and LinkedIn's own engineering blog — to document how LinkedIn's algorithm actually works and what it means for your content strategy. Every recommendation in the checklists (Sections 5–7) traces directly back to specific LinkedIn engineering research. You can act on the conclusions now, or read the full technical evidence in Sections 2–4. Either path works. This page helps you choose.

The guide covers the complete LinkedIn AI ecosystem as of Q1 2026: the two-stage pipeline (Causal LLM retrieval plus Generative Recommender ranking), the serving infrastructure that runs both systems at scale, and the semantic embedding model that underlies all of it. We designed the practical checklists in Sections 5, 6, and 7 to be actionable without the technical walkthrough, but each includes a Key Concepts sidebar for readers who want the vocabulary before diving in.

The Three Things This Guide Proves

1. Relevance has displaced recency as the primary ranking logic of LinkedIn's feed — within the hard 30-day retrieval window that FishDB enforces for connection-based content — and optimizing for timing without optimizing for topic coherence is now a misdirected strategy. Why it matters: A post published three weeks ago that aligns with a reader's professional identity will outrank a post published this morning that does not — so the question of when to post is now secondary to the question of what your content signals about your expertise. → Evidence: Section 3 (How the Algorithms Work) → Action: Section 6 (Content Pre-Launch Checklist)

2. LinkedIn's content distribution system operates as a two-stage pipeline in which the retrieval gate eliminates the vast majority of content before any ranking occurs — meaning your profile (and engagement history) determines whether your content enters consideration, while your content determines how it ranks once it does. Why it matters: The Causal LLM retrieval gate and the Generative Recommender (GR) ranking engine are distinct systems with different inputs. Optimizing for ranking signals while neglecting retrieval means the ranking system never evaluates your content at all; both stages require different, distinct optimization inputs. → Evidence: Section 2 (The Four Pillars) → Action: Sections 5 and 6 (Profile Checklist + Content Checklist)

3. Both LinkedIn's retrieval and ranking systems represent your professional identity as a dense vector in a high-dimensional semantic embedding space, which means topic coherence across your profile and content directly determines the audiences your posts reach. Why it matters: You are no longer placing keywords into categories — you are authoring the document that positions you at a specific coordinate in concept-space, and inconsistent or scattered content dilutes that position, reducing semantic match quality

with your target audience. → Evidence: Section 4 (Semantic Positioning) → Action: Sections 5, 6, and 7

Choose your starting point:

→ **"I need quick wins for my content strategy today"** Read this page, then go directly to Section 5 (Profile Checklist), Section 6 (Content Pre-Launch Checklist), and Section 7 (Engagement Checklist). Each checklist includes a Key Concepts sidebar that defines the technical terms — you need no prior reading before you start. Return to Sections 2–4 when you want to understand why each recommendation works.

→ **"I need to understand the system before I can trust the recommendations"** Read Sections 2, 3, and 4 in order. The four-pillar framework (Section 2), the full two-stage pipeline (Section 3), and the embedding space model (Section 4) together give you the complete picture. The checklists in Sections 5–7 will make full sense after the technical walkthrough. The Technical Appendix (Section 9) has full engineering specifications.

→ **"I'm evaluating this guide for my team, clients, or an educational program"** Read this page and Section 10 (Methodology & Sources) first. The source list documents all citations — arXiv papers, LinkedIn engineering blog posts, and our proprietary research — so you can assess the evidentiary basis before recommending the guide. Section 9 is the complete technical reference for expert verification.

How LinkedIn Works: A Big Picture Look at the Core Pillars

Two decisions determine whether your content reaches anyone on LinkedIn: whether the Causal LLM retrieval system selects it as a candidate, and whether the Generative Recommender ranks it high enough to appear in a specific viewer's feed. Every tactic in this guide — how you write your profile, structure your posts, and build your engagement patterns — ultimately reduces to optimizing for those two decisions. This section explains the four technical pillars behind them: the Training, Retrieval, Ranking, and Serving systems that LinkedIn built to power this process, and what each pillar means for how you write, post, and engage.

To think that you are interacting with "an algorithm" is like looking at a brand-new electric vehicle and calling it "a wheel." You are missing the vast, complex, and deeply interconnected ecosystem that makes it possible. There are now **two AI-powered stages** working in sequence, and your content must pass both. We can understand the modern LinkedIn AI through four foundational pillars: Build, Gate, Rank, and Run.

- **The Factory (Training Infrastructure)**: The sophisticated, high-speed manufacturing plant where engineers build, train, and continuously improve the AI models.
- **The Gate (Causal LLM)**: The retrieval system that determines which content even reaches consideration — selecting the most relevant candidates from hundreds of millions.
- **The Ranker (Generative Recommender)**: The sequential pattern engine that evaluates those candidates by learning from your behavioral history, determining what rises to the top of your feed.
- **The Operating System (vLLM and SGLang)**: The high-performance infrastructure that "runs" these engines at global scale, serving real-time predictions to over a billion members.

Understanding these pillars will give you a complete picture of the forces shaping your visibility on the platform. It will move you from guessing at tactics to developing a durable strategy based on the fundamental principles of the entire ecosystem.

Pillar 1: Training — The Factory That Builds the Engine

Before the system can make a single recommendation, engineers must build the AI model. This process operates at immense scale, involving the ingestion and processing of petabytes of data — the digital equivalent of all the books in the Library of Congress, multiplied many times over. LinkedIn's next-generation AI pipeline platform serves as the "factory" where this happens. (Note: LinkedIn Engineering blogs reference this ecosystem,

which LinkedIn built on the Flyte open-source workflow engine, though the specific internal naming conventions may vary.)

To appreciate how significant this is, you have to understand the old factory it replaced. For years, LinkedIn's AI pipelines ran on a legacy system called ProML. While powerful for its time, it became the digital equivalent of a turn-of-the-century assembly line: slow, rigid, and prone to bottlenecks.

LinkedIn's own engineers reported that making a tiny change to a model — tweaking a single parameter — could require a full 15-minute rebuild of the entire workflow. Imagine if a car factory had to shut down and retool the entire assembly line to change the color of the paint. The very tools meant to enable innovation throttled its pace.

OpenConnect is the gleaming, modern gigafactory that replaced it. LinkedIn's engineers built it on principles of speed, reusability, and resilience, and its design directly impacts how quickly the platform's AI can evolve. For you as a marketer or creator, a smarter factory means a smarter AI that learns and adapts faster. Here's how it works:

Core Principle 1: Standardized, Reusable Parts (The Component Hub)

A modern factory doesn't build every single screw and bolt from scratch for every car. It uses standardized, high-quality components from trusted suppliers — a Bosch fuel injector, a Brembo brake system. OpenConnect does the same for AI.

LinkedIn's platform and vertical AI teams have built a comprehensive library of reusable, pre-approved "components." These are standardized pieces of code for common tasks: a component for collecting data, a component for analyzing a model's performance, a component for processing text. Teams can then assemble these trusted components to build their unique AI pipelines.

1. **What this means for you**: This approach ensures quality and consistency across the entire platform. Engineers build the AI component that understands your job title in the context of the feed from the same foundational block as the one that matches your profile to a job description. This shared understanding allows the AI to make more coherent connections about you across different parts of the platform. When LinkedIn improves one component — for example, a better way to understand skills from text — that improvement can ripple across the ecosystem, making the entire system smarter at once.

Core Principle 2: The Need for Speed (Decoupling and Caching)

The biggest problem with the old factory was its speed. The old system tangled everything together. A change in one area required rebuilding everything. OpenConnect solved this with two key innovations.

- **Decoupling**: The new system isolates every component. Changing one part of the pipeline no longer requires rebuilding the whole thing. It's like a modern pit crew: they can change a tire without having to touch the engine.
- **Caching**: The system intelligently saves the results of previous work. Once engineers build a component, the system stores it in a ready-to-use state (in a Docker image or a manifest file). When an engineer wants to run an experiment, the system doesn't rebuild everything from scratch; it pulls the pre-built components off the shelf and runs them immediately.

LinkedIn reports that this new architecture reduced workflow launch times from over 14 minutes to under 30 seconds (LinkedIn Engineering Blog, 2025).

1. **What this means for you**: This dramatic increase in speed for LinkedIn's engineers translates into a faster pace of innovation for you. An engineer who can run 20 experiments a day instead of one is an engineer who can test more ideas, find what works, and improve the AI much faster. This is why the feed, job recommendations, and other AI-driven features seem to be evolving at a breakneck pace. The factory runs at full speed, constantly executing A/B tests and shipping improvements. The "algorithm" is no longer a static target; it's a rapidly evolving system, and this factory is the reason why.

Core Principle 3: Unwavering Reliability (Disruption Readiness)

An AI factory that processes petabytes of data operates under immense strain. Servers need maintenance, networks can have hiccups, and hardware can fail. In the old world, a disruption like a server reboot could kill a multi-day training job, forcing engineers to start over from scratch, wasting days of work and computational resources.

LinkedIn's engineers designed OpenConnect for resilience. It uses a system of active checkpointing — constantly saving its work. During a long training process, the system automatically saves the model's parameters, its progress, and exactly where it was in the dataset. If a disruption occurs, Flyte (the open-source workflow engine at the heart of OpenConnect) restarts the job on a new set of servers, and it picks up from the last checkpoint. LinkedIn states this approach has reduced training failures due to infrastructure disruptions by 90% (LinkedIn Engineering Blog, 2025).

II. **What this means for you**: The platform's core AI operates more robustly and reliably than ever. This stability allows LinkedIn to train even larger, more complex models with confidence, knowing that disruptions won't derail the process. This industrial-grade reliability serves as a prerequisite for building sophisticated AI systems at the scale of the Causal LLM and Generative Recommender.

Pillar 2: Retrieval (Causal LLM) — The Gate That Determines Who Competes

Before the Generative Recommender can rank content, something has to decide which content even reaches the ranking stage. From hundreds of millions of potential posts, only a pool of the most relevant candidates makes it to the "shortlist." This is the job of the Causal LLM Retrieval Engine — and understanding it is critical because **if your content doesn't pass this gate, the Generative Recommender will never rank it, no matter how good it is.**

Think of the Generative Recommender as the final interview panel for a prestigious job. The Causal LLM is the recruiter who decides which resumes even reach the panel from hundreds of millions of applicants.

The Architecture: A Fine-Tuned LLaMA-3

LinkedIn took Meta's LLaMA-3 (a 3-billion parameter model) and fine-tuned it specifically for retrieval using millions of engagement examples. The Causal LLM operates as a "dual encoder" — generating embeddings (vector representations) for both members and content, then using similarity matching to find the best candidates.

How the system processes you:

For each member, the system constructs a detailed text prompt containing:

III. Your profile information (name, headline, summary, industry, skills)
IV. Your job and education history
V. Your recent **positive** engagement history (posts you liked, commented on, shared — not posts you merely scrolled past)

The fine-tuned LLaMA-3 processes this prompt to generate your "member embedding" — a 3,072-dimensional vector that captures your professional identity and current interests.

How the system processes content:

Similarly, the system processes every piece of content to create an "item embedding" capturing its topic, author, and context.

The retrieval process:

When you load your feed, the system performs a cosine similarity search across hundreds of millions of items to identify the most relevant candidates — all in under 50 milliseconds. This search runs on a dedicated GPU-RAR (GPU Retrieval as Ranking) cluster of 72 H100 GPUs (48 for nearline embedding inference, 24 for retrieval).

Embedding Freshness: Two SLAs That Matter

The system maintains two distinct freshness guarantees:

- **New content and new profiles**: The system generates embeddings **within 1 minute** of creation. When you publish a new post or create a new profile, it enters the retrieval system almost immediately.
- **Updates to existing content and members**: When an existing post gains engagement, or when an existing member's activity changes, the pipeline refreshes those embeddings within **30 minutes**.

This distinction matters for your strategy: new posts enter the retrieval pool quickly, but your engagement activity throughout the day continuously updates how the system understands your professional identity.

What This System Replaced

This single LLM-based system consolidates what LinkedIn previously maintained as a complex patchwork of separate retrieval sources:

1. Embedding-based retrieval (Member_EBR)
2. Global Trending indices
3. Trending in Geo
4. Trending in Industry
5. Cohort EBR
6. Additional collaborative filtering systems

Each of these required separate engineering teams and maintenance. The Causal LLM replaces them all with a unified, semantically-aware approach. As LinkedIn's March 2026 Engineering Blog explains: "By replacing multiple separate retrieval sources with a single embedding-based approach, we cut down engineering complexity." (Danchev, 2026)

The Cold-Start Revolution

This system dramatically improves recommendations for new users and those with smaller networks:

Metric	Overall Impact	Low-Connection Users
Revenue	+0.8%	**+3.29%**
Daily Unique Professional Interactors	+0.2%	—
Daily Unique Professional Interactions	—	**+1.17%**

The gains for new and low-connection users are **3–4x the overall gains** (Ramanujam et al., 2025). The old system relied on network-based signals these users don't have, while the new LLM-based system can match semantic interests even without connection data.

The 512-Dimension Option

Through a technique called Matryoshka representation learning, the system can reduce embeddings from 3,072 dimensions to 512 dimensions with minimal recall loss — a validated option that offers significant storage efficiency. Whether LinkedIn deploys 512 or the full 3,072 dimensions in production remains an internal implementation detail; the source research presents 512 dimensions as a finding with minimal recall impact, not a confirmed deployment specification. (Ramanujam et al., 2025)

II. **What this means for you**: The retrieval gate is where most content visibility wins or loses. If your profile text is unclear or your engagement history is scattered, the system generates a "fuzzy" member embedding that matches poorly with content. Similarly, if your content lacks clear topical focus, the system creates an item embedding that won't match well with relevant audiences. **Profile clarity and consistent positive engagement are retrieval optimization — not ranking optimization alone.**

Pillar 3: Ranking (Generative Recommender) — The Sequential Pattern Engine

After the Causal LLM has selected the candidate pool, a second AI system takes over: the **Generative Recommender (GR)**, also referred to as "Feed SR" (Feed Sequential Recommender) in LinkedIn's academic research. This is the production feed ranker as of early 2026, confirmed by both the arXiv paper published February 2026 (Hertel, Srivastava et al.) and LinkedIn's official Engineering Blog published March 12, 2026 (Danchev).

The Generative Recommender is fundamentally different from what many people imagine when they hear "AI ranker." It does not read your profile like a document. It does not evaluate posts through natural-language comprehension. Instead, it reads your **behavioral history as a chronological sequence** — learning the patterns of how you actually engage with content over time — and uses those patterns to predict how you will engage with candidates today.

> **A note on 360Brew**: LinkedIn's research foundation model (360Brew, built on Mixtral 8x22B with approximately 150 billion parameters) is a separate system. Before deploying GR, LinkedIn's engineers built and tested an "LLM-Ranker" system that represented posts as text prompts fed to a large language model — the approach 360Brew represents. The LLM-Ranker "never achieved superior online performance over the existing production model" and "struggled with

network-based recommendations, because it was difficult to encode the strength of network relationships in a text prompt." (Hertel et al., 2026) The Generative Recommender was chosen instead. 360Brew remains active as a research model and may power other LinkedIn surfaces, but it is not the production feed ranker.

The Core Insight: Your Behavior History Is the Input

The production ranker that 360Brew could not be is not a text-reading engine — it's a **behavioral sequence processor**.

Here is what GR actually does: it takes your last 1,000+ LinkedIn interactions — every post you clicked, liked, commented on, shared, or long-dwelled on — and reads them as a chronological sequence. GR represents each item in that sequence as a compact token pair: the post (encoded with its features) paired with the action you took on it. Your entire engagement history becomes a structured timeline of "what you saw + what you did."

This sequence then flows through a decoder-only transformer — the same fundamental architecture behind modern LLMs, but used here for behavioral pattern recognition rather than text generation. The transformer's **causal attention** mechanism gives every position in your history the ability to "see" all the positions that came before it, but not after. The practical consequence: your most recent interactions naturally carry the most context weight, because everything in your history that preceded them is visible to them. Recent behavior genuinely matters more, not as a design choice but as an architectural property.

As LinkedIn's Engineering Blog explains: "Instead of scoring each post in isolation, GR processes more than a thousand of your historical interactions to understand temporal patterns and long-term interests." (Danchev, 2026)

The Architecture: What's Actually Under the Hood

For readers who want the precise technical picture:

The transformer architecture:

III. Decoder-only design with Pre-LayerNorm (Pre-LN) formulation for training stability
IV. Rotary Positional Embeddings (RoPE) to encode position in the behavioral sequence
V. Causal attention mask — each position attends only to previous positions in the sequence
VI. Input: interleaved post representations and action representations, up to 1,000 history items

The profile embedding (Qwen3 0.6B): Separately from the behavioral sequence, the system generates a dense vector representation of your LinkedIn profile using a fine-tuned **Qwen3 0.6B** model. This 600-million parameter model reads your profile — your headline,

skills, work history, and education — and produces a compact professional identity embedding. GR then "late-fuses" this embedding into the model: the system concatenates it to the transformer output *after* the transformer has processed your behavioral sequence, rather than feeding it into the sequence itself.

This late-fusion design means your profile serves as contextual background that informs the ranking decision without adding computational cost to the transformer's core sequence processing. It is especially valuable for newer members or those with sparse behavioral histories, where GR lacks sufficient interaction data. Adding profile embeddings improves Long-Dwell AUC by more than +2% for members with fewer than 10 historical actions. (Hertel et al., 2026)

The prediction head (MMoE): After the transformer and late fusion, the combined representation passes through a **Multi-gate Mixture-of-Experts (MMoE) prediction head** with shared DCNv2 experts. The key design choice: tasks are grouped into two sets, each with specialized gating:

- *Passive tasks*: click, skip, long-dwell
- *Active tasks*: like, comment, share

This architecture allows the model to simultaneously predict multiple engagement types, and different task groups can attend to different aspects of the combined representation via their distinct gating mechanisms.

The A/B result: GR achieved **+2.10% time spent** compared to the previous production DCNv2-based ranker in LinkedIn's online A/B test — a meaningful improvement at the scale of over a billion members. (Hertel et al., 2026)

The serving infrastructure: GR runs on a **custom PyTorch inference server** with a disaggregated CPU/GPU architecture. The system uses a specialized custom CUDA kernel called **GRMIS** (Generative Recommender Multi-Item Scoring; referred to as SRMIS in the academic paper, same kernel) — a Flash Attention variant designed specifically for Feed SR's attention pattern. GRMIS achieves an average 2x speedup over masked standard attention, enabling efficient parallel scoring of all candidates against a member's shared history context. This is distinct infrastructure from SGLang (see Pillar 4).

What This Means for Your Strategy

The Generative Recommender's architecture has direct implications for how you should think about your LinkedIn presence:

1. Your engagement history is your behavioral profile. The Causal LLM uses your positive engagement history to update your retrieval embedding. The Generative Recommender uses that same history as its primary input sequence. Every time you like, comment, share, or dwell on a post, you are adding a data point to the sequence that both systems use to understand your professional identity. This is not metaphorical — it is architectural.

2. Recent behavior carries the most weight. Because causal attention gives recent positions the most contextual visibility, the last few weeks of your engagement activity exert more influence on your ranking than older history. This does not mean old engagement disappears — GR processes up to 1,000 interactions — but the temporal pattern genuinely matters. Consistent, topically coherent engagement over time reinforces who the system understands you to be.

3. Profile quality matters at the ranking stage, not just retrieval. The Qwen3 0.6B profile embedding is a ranking input, not just a retrieval input. A clear, complete, keyword-rich LinkedIn profile generates a sharper professional identity vector that GR uses as a context feature. Profile optimization is both a retrieval strategy and a ranking strategy simultaneously.

4. Behavioral consistency beats volume. GR does not simply count your interactions. It reads the pattern of your engagement — the sequence of topics, authors, and content types you engage with — to understand your professional trajectory. Scattered, incoherent engagement across unrelated topics produces a noisy behavioral signal that is harder for the model to use for accurate predictions than a smaller number of focused, topically coherent interactions.

Pillar 4: Serving (vLLM and SGLang) — The System That Runs the Engines at Scale

Once the OpenConnect factory has built the AI models and they are ready to serve, there is one final, monumental challenge: how do you actually run them for over a billion members? A sophisticated transformer model is an incredible piece of technology, but it's also computationally intensive. Running it for a single user is demanding; running it for hundreds of millions of active users in real-time is an engineering problem of the highest order.

This is where the final pillar comes in: the serving infrastructure. If the Generative Recommender is the engine and the Causal LLM is the precision filter that selects which races you even enter, the serving layer is the race team's entire operations center — the chassis engineers, fuel crew, and telemetry team keeping multiple high-performance systems running simultaneously. For this, LinkedIn relies on a combination of custom-built infrastructure and two powerful open-source frameworks: vLLM and SGLang.

It is worth being precise about which system runs what, because the architecture is more specialized than a single serving layer:

- **Feed ranking (GR/Feed SR)**: Runs on a **custom PyTorch inference server** with the GRMIS Flash Attention kernel (referred to as SRMIS in the academic paper arXiv:2602.12354v1). This is purpose-built infrastructure, not SGLang.

- **AI Job Search ranking**: Runs on **SGLang**, hosting a cross-encoder small language model (SLM).
- **AI People Search ranking**: Runs on **SGLang**.
- **50+ conversational and generative AI applications** (including LinkedIn Hiring Assistant): Run on **vLLM**.
- **AI Job Search query understanding**: Runs on **vLLM**.

Understanding these technical details helps explain why these frameworks scale efficiently. LinkedIn strategically deploys each framework where its strengths matter most.

vLLM: Powering Interactive AI Applications

vLLM serves as the backbone for more than 50 generative AI use cases across LinkedIn. Products like LinkedIn Hiring Assistant — where the system generates rich, detailed responses about job candidates — run on vLLM's infrastructure. Its PagedAttention memory management makes it exceptionally efficient at handling the varied, high-output workloads these applications demand.

AI Job Search illustrates how these frameworks layer within a single product: vLLM powers the **query understanding** component — interpreting a member's free-form search query, extracting structured facets, and generating a rich semantic interpretation of what the member is looking for. This is a generative text production workload. SGLang then handles the underlying LLM-based ranking of job results that follows.

SGLang: Powering Search Ranking Systems

For LLM-based ranking workloads in job search and people search, LinkedIn uses SGLang. The February 2026 Scaling LLM engineering publication (Ramachandran et al.) explicitly confirms: "At LinkedIn, these advancements power AI Job Search and AI People Search to deliver state-of-the-art LLM ranking to millions of members."

SGLang excels at scoring thousands of candidate items against a single query context at scale, without generating any text output — a fundamentally different workload pattern from the conversational applications that vLLM handles.

LinkedIn hasn't adopted these frameworks passively; they've actively improved them. Their engineers contributed **Multi-Item Scoring** to SGLang, a technique that allows the system to score multiple pieces of content in a single model pass rather than one at a time. This contribution alone achieved a **69% reduction in latency** (Shimizu, 2025). They also integrated FlashAttention 3, a cutting-edge attention mechanism that dramatically speeds up the model's core computations.

Perhaps most cleverly, LinkedIn developed the **"Knock-Knock" technique**, which hides prefill latency by intelligently pre-computing parts of the model's work before the full request arrives — specifically, prefilling the member context while candidate retrieval is

still in progress. This innovation reduced overall latency by approximately **38%** (from 520ms to 200ms)(Shimizu, 2025), making the system feel significantly more responsive.

LinkedIn's February 2026 engineering publication (Ramachandran et al.) documents a subsequent four-stage optimization journey that extended SGLang's capabilities from recommendation ranking into search:

- **Batch tokenization and batch send** — Fixing how the system transmitted tokenized batches across the pipeline
- **Scoring-only execution path** — Eliminating decode overhead entirely for pure ranking workloads
- **In-batch prefix caching** — Reusing shared member context computation across candidate items
- **Python runtime optimization** — Garbage collection elimination and multi-process scheduler parallelization

Moving systematically through these stages, the team achieved a **3x throughput improvement** for text-based ranking (750 to 2,200 items per second per GPU) and **2.2x** for mixed-input ranking. The team has stated they are "going deeper into the stack with fine-grained profiling, kernel-level tuning, and further trimming overheads in prefill and attention paths" — confirming this optimization trajectory is ongoing.

The Implication: A System in Constant Flux

By building on vLLM and SGLang — and by investing in custom infrastructure for its primary feed ranker — LinkedIn does not run its AI serving stack on static, internally developed software that engineers update once or twice a year. It is a living, breathing system that benefits from the collective R&D of the entire global AI community alongside LinkedIn-specific specialization.

LinkedIn's own engineers have documented their journey through multiple versions of these frameworks in a matter of months, with each new version bringing significant performance improvements. They are not consumers of this technology alone; they are active contributors, submitting their own performance optimizations back to the open-source projects for everyone to use.

II. **What this means for you**: This is the ultimate reason why chasing short-term "hacks" provides diminishing returns. The very foundation on which the ranking engine runs is changing and improving on an ongoing basis. A loophole or quirk you might discover in the system's behavior at breakfast could be completely gone by lunch — not because a LinkedIn product manager decided to change it, but because the underlying serving frameworks received updates that changed how they schedule and process requests on the GPU.

This constant state of flux and improvement makes it impossible to "game" the system for long. The only durable strategy is to align with the core principles of the ecosystem: providing a clear, complete professional profile that generates strong embeddings at both the retrieval and ranking stages; creating valuable and well-reasoned content; and engaging consistently in ways that build a coherent, legible behavioral signal. The ecosystem will always value these inputs, regardless of how the underlying technology evolves.

References

Ramanujam, S. S., et al. (2025). Large scale retrieval for the LinkedIn Feed using causal language models. *arXiv*. https://arxiv.org/abs/2510.14223

Hertel, L., Srivastava, G., Naqvi, S. A., Kumar, S., Zhang, Y., Ocejo, B., Zelditch, B., Englhardt, A., Cheng, H., Hu, A., Alonso, A., Li, D., Dangi, S., Zhu, C., Zhou, M., Li, W., Huang, T., Borisyuk, F., Parameswaran, G., Tiwana, B., Sankar, S., Lan, Q., Choi, J., & Ghosh, S. (2026). An industrial-scale sequential recommender for LinkedIn Feed ranking. *arXiv*. https://arxiv.org/abs/2602.12354

Danchev, H. (2026, March 12). Engineering the next generation of LinkedIn's Feed. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/feed/engineering-the-next-generation-of-linkedins-feed

360Brew Team. (2025). 360Brew: A decoder-only foundation model for personalized ranking and recommendation. *arXiv*. https://arxiv.org/abs/2501.16450

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Shimizu, S. (2025, December 9). Turbocharging LinkedIn's recommendation systems with SGLang. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/ai/turbocharging-linkedin-recommendation-systems-with-sglang

Ramachandran, S. R., Lan, Q., Nguyen, C., Sheng, J., & Zhu, C. (2026, February 20). Scaling LLM-based ranking systems with SGLang at LinkedIn. *LinkedIn Engineering Blog*.

Zhai, Y., Kumar, S., Ramachandran, S. R., Zhu, C., Nguyen, C., Toddywala, F., Yao, C., Johnson, C., & Lan, Q. (2025, August 26). How we leveraged vLLM to power our GenAI applications at LinkedIn. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/ai/how-we-leveraged-vllm-to-power-our-genai-applications

How LinkedIn's Algorithms Actually Work

This section traces your content's complete journey through LinkedIn's two-stage algorithmic pipeline — from the moment you hit publish to the moment it appears (or doesn't) in a specific viewer's feed. Understanding this pipeline explains why certain profile configurations and content choices consistently outperform others. If you prefer the tactical applications without the full technical walkthrough, the checklists in Sections 5–7 translate these mechanics into direct actions. If you want to understand why those actions work — and when to adapt them — read on.

The Two-Stage Pipeline: Why Both Gates Matter

Your content must pass through two sequential gates. The fundamental architecture of LinkedIn's content distribution system operates as a **two-stage pipeline**, and your success depends on optimizing for BOTH stages:

Stage 1: The Retrieval Gate (Dual-Path Architecture)

From hundreds of millions of potential posts, the retrieval systems must select approximately 2,000 candidates — all within 50 milliseconds. This is the **PRIMARY gating function**. If your content doesn't make it into this initial pool, no one will ever see it, regardless of how well it might rank.

Critical architecture insight: LinkedIn operates a **dual-path retrieval system** with different engines for different content types:

Path A — Connection-Based Content (FishDB)

III. **FishDB**: LinkedIn's high-performance Rust-based generic retrieval engine
IV. Handles content from your network (connections, followed creators, companies)
V. Maintains a **30-day content window** — FishDB excludes content older than 30 days from retrieval
VI. P99 latency: 40ms
VII. Note: FishDB is a generic retrieval engine, not purpose-built exclusively for feed — it also powers other LinkedIn retrieval use cases

Path B — Out-of-Network Content (Causal LLM)

- **Causal LLM Dual Encoder**: Fine-tuned LLaMA-3 (3B parameters)
- Handles "suggested" content from creators you don't follow
- Uses semantic embedding matching rather than network proximity
- Critical for new users and content discovery

Supporting Systems (Both Paths)

- **Cross-Domain GNN**: Understanding your position in LinkedIn's Economic Graph
- **Heuristic systems**: Fast rules for recency, velocity, and recent interactions
- **GPU-RAR cluster**: 72 total H100 GPUs (48 nearline processing + 24 dedicated retrieval)

Stage 2: The Ranking Engine (Generative Recommender / Feed SR)

Only after content passes the retrieval gate does it reach the Generative Recommender (GR) for ranking. This sequential transformer model processes your behavioral history — but **only on the ~2,000 candidates that made it through Stage 1**.

Why This Matters for You

Most advice focuses exclusively on ranking optimization. But if you're not optimizing for retrieval (Causal LLM), you're optimizing for a competition you may never enter.

For new users and those with smaller networks, the retrieval stage is even more critical. LinkedIn's research shows that the Causal LLM retrieval system delivers +1.17% increase in Daily Unique Professional Interactions and +3.29% increase in revenue specifically for these user groups — indicating that high-quality retrieval is the primary driver of their content discovery.

The Multi-System Architecture at a Glance

Stage	System	Parameters	Function	Status
Retrieval (Network)	FishDB	N/A (Rust engine)	Connection-based content, 30-day window	Production
Retrieval (OON)	Causal LLM	3B (LLaMA-3)	Out-of-network "suggested" content	Production
Ranking	Generative Recommender / Feed SR	Sequential Transformer	Ranks candidates that passed retrieval	Production

Profile clarity directly benefits the Causal LLM (which constructs a text prompt from your profile) and the GR ranking stage (which uses LLM-generated profile embeddings). All stages benefit from a clear, coherent professional identity. All matter.

Step 1: Candidate Generation (The Initial Longlist)

This is the very top of the funnel, where personalization begins. The system's first task is to sift through the billions of potential posts in the LinkedIn universe and select a few thousand that might be relevant to you. This is a game of recall, not precision. The goal is not to find the single best post, but to ensure that the best post is somewhere in the initial pool of candidates. If your content doesn't make it into this initial "longlist," no one will ever see it, no matter how good it is.

To accomplish this at incredible speed, the system uses several efficient methods running in parallel. LinkedIn designed these methods to be fast and broad, casting a wide net to pull in a diverse set of potential content. The two primary methods are Cross-Domain Graph Neural Networks and Heuristic-Based Retrieval.

Cross-Domain Graph Neural Networks (GNNs): The Holistic Scout

(Based on LinkedIn's cross-domain GNN research: He et al., 2025, arXiv:2506.12700 — with an important source qualification noted below)

What happens

LinkedIn maintains a colossal, constantly updated map of its entire professional ecosystem, known as the Economic Graph. This isn't just a list of members and companies; it's a complex web of interconnected nodes and edges representing every entity and every interaction: the graph connects you (a node) to your company (a node) with a "works at" edge; it connects you to another member with a "connection" edge; it connects you to a post with a "liked" edge.

A Graph Neural Network (GNN) is a specialized type of AI that LinkedIn designed to learn from this very structure. The GNN can "walk" the graph, learning patterns from the relationships between nodes. The most significant evolution here is that LinkedIn's GNN is now cross-domain.

Previously, LinkedIn might have trained a GNN on Feed data alone to recommend Feed content. The new Cross-Domain GNN takes a holistic approach. It ingests and learns from your activity across the entire platform. It sees the jobs you click on, the notifications you open, the influencers you follow in your email digests, the skills you endorse, and the articles you share. It then uses this complete, 360-degree view of your professional interests to find potential content.

For example, if you've recently started clicking on job postings for "Product Marketing Manager," the GNN learns that you are interested in this topic. It can then walk the graph to find high-quality posts, articles, and discussions about product marketing, even if you've never explicitly engaged with that topic in the feed before. It uses your behavior in one domain (Jobs) to inform its recommendations in another (the Feed).

Source qualification — important context for technical readers: LinkedIn's cross-domain GNN research (He et al., 2025, arXiv:2506.12700) was developed primarily for the **notification system**, not the feed ranking pipeline. The paper describes a large-scale cross-domain GNN deployed for personalized notifications at LinkedIn. Importantly, the paper notes that the member embeddings this system produces are designed for reuse across LinkedIn surfaces — including the Feed. This cross-surface reusability is the basis for the cross-domain behavior described above. However, the GNN paper should not be read as a primary feed architecture source; it is a notification system paper whose outputs have feed implications.

Additionally, note that the GNN notification system uses a **daily refresh cadence** for member embeddings — distinct from the 30-minute refresh cadence used by the Causal LLM retrieval system described later in this section. Different subsystems within LinkedIn's infrastructure operate on different update schedules.

So what?

This means your professional identity on LinkedIn no longer exists in silos. The system builds a single, unified understanding of you based on the totality of your actions. Every click, every follow, every job application refines your "member embedding" — your unique digital fingerprint on the Economic Graph. The system constantly tries to answer the question: "Based on everything this member does on our platform, what are they truly interested in professionally?" The system pulls your content into the longlist when its own "graph neighborhood" — the topics, skills, and people it's connected to — strongly overlaps with a member's holistic interests.

For Priya (Director of Marketing) and Jared (Agency owner): think of this as your LinkedIn reputation system. Everything you do on LinkedIn — not just your posts — contributes to how the system understands who you are and what content is relevant to your world. Your network quality, your engagement patterns, even which companies you follow: these all inform which content gets surfaced to you, and whose content you get surfaced in.

Now what?

Your goal is to create a clear, consistent, and coherent professional identity across the entire platform, not just in your posts.

- **Build a Relevant Network**: Your connections are a primary signal. Connect with professionals in your target industry and with individuals who engage with the kind of content you create. When your connections engage with your content, they signal to the GNN that your post is relevant to that specific "graph neighborhood," increasing the likelihood that the system will show it to their connections (your 2nd and 3rd-degree network).
- **Maintain Your Profile as Your Professional Hub**: The skills listed on your profile, the job titles you've held, and the companies you've worked for are powerful, stable nodes in the graph. The GNN uses this information as an anchor for your identity. If

your profile clearly states your expertise in "B2B SaaS Marketing," the GNN becomes far more likely to identify your content on that topic as relevant.

- **Engage Authentically Beyond the Feed**: Your activity is not just about feed engagement. Clicking on a job ad, following a company, or even watching a LinkedIn Learning video are all signals that feed into the cross-domain system. Engage with the platform in a way that authentically reflects your professional interests and goals. This holistic activity provides the rich data the system needs to understand who you are and, by extension, who your content is for.

FishDB: The Connection-Based Retrieval Engine

What happens

While the GNN and Causal LLM handle semantic matching and network understanding, LinkedIn built a specialized high-performance engine specifically for retrieving content from your direct network: **FishDB**.

FishDB is LinkedIn's **generic** Rust-based retrieval system — "generic" in the sense that LinkedIn designed it to be reusable across LinkedIn's retrieval needs, not purpose-built for the feed alone. For the feed use case, FishDB optimizes for connection-graph queries. When you open your feed, FishDB rapidly traverses your connection graph to find recent content from:

- People you follow
- Your direct connections (1st degree)
- Companies you follow
- Creators you've subscribed to

Critical specification: FishDB maintains a **30-day content window**. FishDB does not index content older than 30 days and will not surface it through this retrieval path. This is a hard architectural constraint, not a ranking preference. The 30-day design boundary is baked into FishDB's in-memory index structure — FishDB does not index content outside this window and therefore cannot retrieve it, regardless of its relevance.

Performance characteristics:

- P99 latency: 40ms (meaning 99% of queries complete in under 40ms)
- LinkedIn designed it for high throughput connection traversal
- Uses a four-component index: **forward index**, **inverted index**, **reference index**, and **attribute stores** (RocksDB-backed key-value stores with bloom filter and LRU cache for sparse data including embeddings and spam classification features)

FishDB's broader role: Because FishDB is a generic retrieval engine (not feed-specific), it also powers other LinkedIn retrieval use cases beyond the feed. The feed deployment

replaced LinkedIn's FollowFeed Java system (which had powered the feed for nearly a decade), achieving 2x efficiency improvement and 50% hardware reduction in the process.

So what?

This 30-day window is one of the most concrete, actionable insights from LinkedIn's architecture. For content from your network:

- Posts older than 30 days face a **hard ceiling on discoverability** through this path
- The "relevance over recency" paradigm operates within this window, not beyond it
- Evergreen content still needs periodic resharing or engagement to remain discoverable

The dual-path implication: Your content's path through retrieval depends on the viewer's relationship to you:

1. Followers and connections → FishDB path (30-day window applies)
2. Non-followers seeing "suggested" content → Causal LLM path (semantic matching, different constraints)

For Priya and Jared: the practical takeaway is that even the best evergreen content ages out of the FishDB retrieval pool after 30 days. This is a system design fact, not an algorithmic judgment about your content's quality. You need a strategy for keeping valuable content in circulation.

Now what?

1. **Maintain consistent posting cadence**: Don't go silent for weeks. The 30-day window means extended breaks create visibility gaps even for followers.
2. **Reshare evergreen content strategically**: If you have cornerstone content that remains valuable, consider resharing or referencing it within the 30-day window to keep it discoverable.
3. **Understand the two audiences**: Content optimized for your existing network (FishDB path) versus content optimized for discovery by new audiences (Causal LLM path) may benefit from different approaches.
4. **Engagement extends visibility**: While the 30-day window is hard, other mechanisms beyond FishDB may surface content that continues receiving engagement.

Heuristics and Similarity Search: The Fast Scouts

While the GNN is incredibly powerful for understanding deep relational patterns, it's also computationally intensive. To supplement it, the system uses several faster methods to fill the longlist with timely, fresh, and highly relevant content.

What happens

This stage combines simpler, rule-based methods (heuristics) with efficient search techniques to quickly find candidates.

- **Heuristic-Based Retrieval**: These common-sense rules execute at massive scale with very low latency. Examples include:
    1. **Timeliness**: The system surfaces very recent posts from a member's direct connections.
    2. **Recent Interaction**: If a member just commented on one of your posts, the system becomes more likely to pull your next post into their longlist.
    3. **Velocity**: The system flags posts that gain unusually high engagement (likes, comments) very quickly and pulls them into more longlists to assess whether they offer broad relevance to additional audiences.
- **Similarity Search (Embedding-Based Retrieval)**: This more sophisticated but still incredibly fast method leverages embeddings. Every piece of content and every member has an "embedding" — a digital fingerprint. The system takes your member embedding and, in a fraction of a second, searches a massive database for posts with the most similar embeddings. "Similar" can mean many things: similar topics, similar style, or content liked by members with similar profiles to yours. This approach allows the system to find topically relevant content even from creators you're not connected to.

So what?

Speed and topical clarity function as crucial factors for getting into the initial longlist. While the GNN looks at your deep, long-term identity, these faster methods focus on the "here and now." A well-timed post on a trending topic, or one that gets a quick burst of initial engagement, can leverage these heuristics to get a significant initial boost in visibility. Similarly, content with a very clear and distinct topical focus generates a "sharper" embedding, making it easier for the similarity search to find and match it with the right audience.

Now what?

Your strategy here should focus on creating clear, timely content and fostering immediate engagement.

1. **Be Clear and Specific**: When you write a post, have a single, clear topic in mind. A post about "The Impact of AI on B2B SaaS Go-to-Market Strategy" generates a much more distinct and matchable embedding than a vague post about "The Future of Business." Avoid muddled or overly broad topics in a single post if you want the similarity search to find you.
2. **Encourage Early Engagement**: Early engagement can help your content pass through velocity-based heuristics, but this is one factor among many. Unlike the old recency-driven system, a post that gains traction more gradually can still succeed if its relevance score is high. Focus on being present to respond to comments when they arrive — this creates richer engagement signals — rather than fixating on a specific time window.
3. **Engage Authentically with Others**: The recent interaction heuristic works as a two-way street. When you thoughtfully engage with content from others in your target audience, you increase the probability that the system will show your next post to them. Authentic engagement does more than build relationships; it sends a direct technical signal to the candidate generation system.
4. **Tap into Trending Topics (When Relevant)**: If there is a significant conversation happening in your industry, creating a timely and insightful post on that topic can leverage the system's ability to identify and boost trending content. Don't force it, but when a topic aligns with your expertise, timeliness can be a powerful amplifier.

By understanding this first crucial step, you can see that getting visibility is not about a single magic bullet. It's about building a strong, coherent professional identity (for the GNN) while also creating clear, timely, and engaging content (for the faster retrieval methods). If you can successfully align your efforts with both of these systems, you will maximize your chances of getting your content into the initial longlist — the gateway to the powerful ranking that follows.

The Retrieval Gate: The Causal LLM Revolution

This is arguably the most important development in LinkedIn's current infrastructure — and the most underappreciated. The Causal LLM retrieval system is the **primary gate** that determines whether your content even enters the competition for visibility among out-of-network users.

What happens

LinkedIn fine-tuned Meta's LLaMA-3 (a 3-billion parameter causal language model) as a "dual encoder" to generate high-quality embeddings for both members and content. This system represents a fundamental architectural shift: it consolidates what previously

existed as a complex patchwork of many separate retrieval systems into a single, unified, semantically-aware engine.

What the old system looked like:

The previous architecture relied on a heterogeneous set of sources — the March 2026 LinkedIn Engineering Blog (Danchev) describes it as "multiple retrieval sources: trending content, collaborative filtering, and embedding-based systems, each maintaining separate infrastructure." Among the sources specifically illustrated in Figure 1 of the Causal LLM paper (arXiv:2510.14223v1) are:

1. Member_EBR (embedding-based retrieval)
2. Global Trending indices
3. Trending in Geo
4. Trending in Industry
5. Cohort EBR

Additionally, collaborative filtering was part of the broader historical retrieval stack, as described in the paper's Related Work section: "LinkedIn's Feed retrieval stack exemplifies this evolution, employing a combination of inverted indices for chronologically ordered activities, trending sources, and collaborative filtering." Each of these required separate engineering teams, separate feature pipelines, and separate maintenance. The Causal LLM consolidates them.

How the new system works:

For retrieval, LinkedIn's system constructs a detailed text prompt from your profile:

1. Your name, headline, and summary
2. Your industry, skills, and location
3. Your job and education history
4. Your certifications and languages
5. **Your recent positive engagement history** (posts you liked, commented on, shared)

The fine-tuned LLaMA-3 model processes this prompt to generate your "member embedding" — a 3,072-dimensional vector (or 512 dimensions as a validated efficient deployment option via Matryoshka learning, with minimal recall loss) that captures your professional identity and current interests.

Similarly, the same model processes every piece of content on LinkedIn to create an "item embedding."

The retrieval process:

When you load your feed:

- The system fetches your member embedding from a key-value store
- A GPU-RAR (GPU Retrieval as Ranking) cluster performs a cosine similarity search across 72 total H100 GPUs (48 nearline processing + 24 retrieval)
- The system retrieves the top candidates — in under 50 milliseconds
- These candidates (and only these candidates) proceed to the Generative Recommender for ranking

A note on candidate counts: The Causal LLM paper (arXiv:2510.14223v1) states "2000 candidates are retrieved" in its abstract (using the paper's own language), referring to the total retrieval pool across all sources combined. The Implementation Details section specifies "top 1,000 candidates to feed to subsequent layers" for the Causal LLM specifically. The combined total from all retrieval sources — FishDB network content plus Causal LLM out-of-network content plus heuristics — reaches approximately 2,000 candidates entering the ranking stage.

Freshness SLAs — a critical distinction:

- **New content**: the system indexes new posts within **1 minute** of creation (near-real-time)
- **New member profiles**: the system generates embeddings within **1 minute**
- **Existing item updates** (interaction signals on existing posts): the system updates embeddings within **30 minutes**
- **Existing member activity** (likes, comments by existing members): the system updates member embeddings within **30 minutes**

The 1-minute SLA for new content means your posts enter the retrieval index almost immediately. The 30-minute cycle applies to ongoing updates, not initial indexing.

Critical insight from the research: LinkedIn found that including ONLY positive engagements (likes, comments, shares) in the history sequence significantly improved retrieval quality compared to including all engagements or no history. The system specifically learns from what you approve of — not what you scroll past or dislike.

So what?

This infrastructure powers the relevance-over-recency shift. The system doesn't just match keywords — it understands what your professional interests actually mean and finds content that genuinely matches those interests.

The cold-start revolution: This system dramatically improves recommendations for new users and those with smaller networks. LinkedIn's A/B testing produced remarkable results:

Metric	Overall Impact	Low-Connection Users
Revenue	+0.8%	**+3.29%**
Daily Unique Professional Interactors	+0.2%	—
Daily Unique Professional Interactions	—	**+1.17%**
Daily Active Users	—	**+0.23%**

The gains for new and low-connection users are **3–4x the overall gains**. This is because:

- The old system relied on network-based signals these users don't have
- The new LLM-based system can match semantic interests even without connection data
- Profile text quality matters most when you have no engagement history

For Priya (small team, building LinkedIn presence) and Jared (agency clients across various maturity stages): this means the old excuse — "I don't have enough followers yet for LinkedIn to work for me" — no longer holds. The system can find your audience semantically even before you've built a large network. But it can only do this if your profile text is clear and specific enough for the LLM to form an accurate representation of who you are.

Now what?

Understanding that your content must pass this gate before the Generative Recommender ever sees it changes everything about optimization.

For Your Profile:

- Write your headline as if an LLM will read it (because one will). "B2B SaaS demand generation specialist" creates a sharper embedding than "marketing guru who helps businesses grow."
- Your About section feeds directly into your member embedding. Clear, specific descriptions of your expertise create sharp, matchable embeddings. Generic buzzwords create fuzzy embeddings that match poorly with potential audiences.
- Use consistent terminology — the same language your target audience uses in their own profiles.

For Your Content:

- The system processes every post to create an item embedding.
- Clear topical focus creates a sharper, more matchable embedding.
- Muddled posts with multiple unrelated topics create fuzzy embeddings that match poorly with any audience.
- Use the specific terminology your target audience would use — if they have "product marketing" in their headlines, use "product marketing" in your content.

For Your Engagement:

- The system includes your positive engagement history in your member prompt.
- Engaging with high-quality content in your niche improves your member embedding.
- Random, unfocused engagement creates noise that degrades retrieval quality.
- The quality and topical coherence of what you engage with directly affects what content the system retrieves for you AND how the system retrieves your own content for others.

For Cold-Start Situations (New Users, New Topics):

- Your profile text is essentially your entire identity in the retrieval system.
- There's no engagement history to correct weak textual positioning.
- Quality of initial profile setup has outsized impact on content discovery.
- Your first 30 days of engagement shape your initial member embedding — be intentional.

Step 1.5: Pre-Ranking (The L1 Layer)

Before candidates reach the full Generative Recommender ranking engine, they may pass through a lighter-weight pre-ranking layer (L1). This stage:

- Reduces candidates further using efficient scoring
- Applies initial relevance filters without full transformer inference
- Ensures that only the most promising candidates receive complete evaluation

This middle layer exists because running a large sequential transformer on thousands of candidates would be computationally prohibitive. The L1 layer narrows the field efficiently.

Note on L1 for Feed: Figure 1 of the Causal LLM paper (arXiv:2510.14223v1) explicitly shows an **L1 Preranking** box in the feed suggested-content architecture, positioned between the retrieval sources and the "Final Feed Ranking Model." This provides direct evidence that an L1 layer exists in the feed pipeline. The MixLM paper (arXiv:2512.07846v1, December 2025) confirms L1 pre-ranking for LinkedIn's Job Search as well, suggesting L1 is a consistent pattern across LinkedIn's ranking surfaces. What matters for content creators is understanding that multiple filtering stages exist between retrieval and final ranking: your content must survive the retrieval gate, a pre-ranking filter, and finally the Generative Recommender's full evaluation before reaching the top of anyone's feed.

Sources: arXiv:2510.14223v1 (Causal LLM retrieval, Oct 2025) · LinkedIn Engineering Blog: FishDB (Nov 2025) · LinkedIn Engineering Blog: Engineering the next generation of LinkedIn's

Feed, Danchev (Mar 12, 2026) · arXiv:2602.12354v1 (Feed SR / Generative Recommender, Feb 2026) · arXiv:2506.12700v1 (Cross-Domain GNN — notification system, He et al., 2025) · arXiv:2512.07846v1 (MixLM / Job Search L1, Dec 2025)

Step 2: The Generative Recommender (GR) Ranking Engine

At this point in the pipeline, the retrieval layer has done its work. The Causal LLM has cast a wide, semantically intelligent net and hauled in roughly 2,000 candidate posts from the vast ocean of LinkedIn content. These candidates have already passed an intelligent gate — they're here because the system believes they're plausibly relevant to you. Now comes the harder question: which handful of these candidates actually belong at the top of your feed, in what order, and why?

This is the job of the Generative Recommender — LinkedIn's production ranking engine, publicly announced in March 2026 and described in the arXiv paper "An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking" (Hertel, Srivastava et al., arXiv:2602.12354v1). The Generative Recommender, or GR, replaced the previous DCNv2-based production ranker and achieved +2.10% time spent in A/B testing — a meaningful improvement at the scale of 1.3 billion professionals.

Understanding how GR works will fundamentally change how you think about your LinkedIn presence.

What It Is: A Sequential Pattern Engine

What happens

GR does not read your profile the way a recruiter reads a resume. It does not process a written description of your professional history and ask itself, "Given everything I know about this person, would they like this post?"

Instead, GR learns who you are by studying what you have *done*.

Specifically, GR processes your last 1,000 interactions on LinkedIn as an ordered, chronological sequence. Not summaries of those interactions. Not text descriptions of them. The actual sequence of posts you saw, what you did with them, what you saw next, what you did with that — all the way back through your recent LinkedIn history.

Think of it less like a reader comprehending your career story as a narrative, and more like a music streaming service that understands your taste by analyzing the precise pattern of everything you've ever played, skipped, replayed, and added to a playlist. Spotify doesn't need you to write an essay about your musical preferences. It figures out that you love introspective indie-folk but skip upbeat pop from the sequence itself. GR operates on the same logic, applied to your professional engagement.

An alternative approach — known internally as the LLM-Ranker — did attempt to work with natural language: it represented candidate posts as text, assembled briefing documents, and asked a large language model to predict whether you'd engage. LinkedIn built and tested that approach. The results were clear: the LLM-Ranker "never achieved superior

online performance over the existing production model," and it struggled specifically with network-based recommendations because "it was difficult to encode the strength of network relationships in a text prompt." (arXiv:2602.12354v1, Section 5.1). LinkedIn set the LLM-Ranker aside. GR is what LinkedIn built instead.

So what?

The engine evaluating your content is not a reader — it is a pattern recognizer. GR doesn't evaluate your post by "reading" it with comprehension the way a human expert would assess its quality. It evaluates your post in the context of what the viewer has done before, building a prediction of what they're likely to do next based on the trajectory of all those prior actions.

This is a more honest and arguably more powerful model. It cannot be fooled by the right keywords or manipulated by writing posts that "sound good to an AI." It is watching behavior, and behavior doesn't lie.

Now what?

Stop optimizing for how your content reads to an AI. Start building a behavioral signal that GR can learn from:

1. **Consistency beats cleverness.** A pattern of engagement on a coherent set of topics creates a strong, clear behavioral signal. Scattered, occasional engagement across random topics creates noise.
2. **Your engagement history is your professional identity, as far as GR is concerned.** What you consistently engage with shapes the behavioral sequence that GR uses to rank everything you see — and, importantly, shapes how it understands the audience for content similar to yours.
3. **For new LinkedIn users, GR supplements your sparse history with profile embeddings (more on this below).** Getting your profile right matters even more when you haven't yet built a behavioral record.

How It Works: The Power of Sequence

What happens

GR is built on a decoder-only sequential transformer — a modern deep learning architecture that, like the language models behind tools such as ChatGPT, is capable of attending to long sequences and finding meaningful patterns within them. The key difference: GR's sequences are not sentences in English. They are sequences of your LinkedIn interactions.

Here is the specific structure, as documented in arXiv:2602.12354v1:

The sequence GR processes:

GR organizes your interaction history chronologically. For each impression — each post you saw — GR creates two elements: a compact representation of the post itself (its content embedding, the author's identity, relevant categorical signals) and a compact representation of what you did with it (clicked, liked, commented, shared, long-dwelled, or skipped). GR represents each post by approximately 2 tokens (item and action embeddings) and interleaves these two elements to create a unified input:

[Post 1, Action 1, Post 2, Action 2, Post 3, Action 3, ... Post 1000, Action 1000]

This interleaved sequence — up to 1,000 post-action pairs — is what the transformer processes. Representing each post as approximately 2 tokens is dramatically more efficient than representing each post as hundreds of words of natural language text.

The causal attention mechanism:

The transformer processes this sequence with *causal attention*. This means each position in the sequence can only attend to previous positions — it can see what came before, but not what came after. This mirrors how you actually experienced that content: each engagement happened in the context of everything you'd done before, not in the context of what you'd do later.

The practical consequence of causal attention is important: **recent interactions naturally receive more contextual weighting, because the model processes them with the benefit of more historical context.** Your engagement from yesterday is seen in light of everything you did last week, last month, and across your full history. Your engagement from a year ago is seen in a much sparser context. The model's architecture means recent behavior carries more signal simply by virtue of how the sequence is built.

The candidate evaluation:

At inference time, LinkedIn appends the candidate posts to be ranked at the end of your historical sequence. GR then scores all candidates in a single forward pass — an efficiency technique called Multi-Item Scoring that avoids redundant recomputation of your history for each candidate. The result is a relevance score for each candidate, predicted for multiple engagement types simultaneously.

So what?

Your most recent behavior is the most influential context for GR's predictions. This is not a policy decision LinkedIn made about recency — it is an architectural consequence of how causal attention works in sequential transformers. LinkedIn designed GR to weight recent behavioral context more heavily, and that design delivers better predictions of near-future engagement.

The corollary for content creators is equally important: GR learns from the *pattern* of behavior across your audience. When your content reliably generates consistent

engagement patterns from a specific type of professional, GR learns to recognize that audience type and route your future content accordingly.

Now what?

For your own feed curation:

- **Engage with strategic purpose.** Every interaction you take on LinkedIn adds to your behavioral sequence. Engaging with posts in your target topic areas — even skimming and briefly dwelling on them — contributes to the sequential pattern GR uses to route content.
- **Cross-session consistency matters more than single-session coherence.** GR's architecture processes your cross-session engagement history — the pattern of topics, authors, and content types you consistently engage with over days and weeks. The model explicitly de-biases within-session co-occurrence to reduce overfitting, so your primary lever is building a consistent behavioral record over time, not engineering any individual browsing session.

For your content strategy:

- **Consistent topic focus builds audience routing intelligence.** If your posts reliably attract a specific professional audience, GR learns this relationship and becomes more efficient at routing your future content to that audience.
- **Your first few dozen interactions on a new topic help prime the model.** Engaging with relevant content before you post on a topic helps populate your recent sequence with contextually aligned behavior.

GR's Profile Layer: The Qwen3 0.6B Embedding

What happens

Sequential patterns are powerful, but they have an obvious limitation: they require history. A new LinkedIn member has no interaction history. Even an existing member who begins exploring a new topic area has sparse signal in that direction. And behavioral sequences, however deep, don't contain explicit professional identity information — GR cannot "see" your job title, your skills, your industry, or your educational background simply by watching what you click.

LinkedIn addressed this with a separate, parallel component: a fine-tuned **Qwen3 0.6B** language model that reads your LinkedIn profile and converts it into a dense vector — a compressed numerical representation of your professional identity.

Late fusion then adds this profile embedding to the ranking model: after the transformer has processed your interaction sequence, the system concatenates the profile embedding to the transformer output as an additional context feature. The profile embedding does not pass through the sequential transformer layers. The causal attention mechanism does not

see it. Late fusion incorporates it into the final representation that feeds the prediction head.

(The arXiv paper is precise about this: "Member profile embeddings are an LLM-based dense representation that captures comprehensive information from LinkedIn member profiles. These embeddings are generated by aggregating member profile information with a Qwen3 0.6 billion parameter fine-tuned model... we integrate it as a late-fused context feature." — arXiv:2602.12354v1, Section 4.5)

The profile embeddings refresh daily, meaning as you update your LinkedIn profile, the embedding GR uses updates within 24 hours.

The offline data is compelling: for members with fewer than 10 historical interactions, adding profile embeddings as a late-fused feature improves GR's Long-Dwell AUC by more than 2%. Your profile matters most precisely when your behavioral history is thinnest.

So what?

Your profile quality affects the ranking stage in two distinct ways — and through two different mechanisms:

- **Retrieval stage (Causal LLM):** Your profile text is encoded by a LLaMA-3 3B dual encoder to generate your member embedding. This embedding determines which content the system retrieves for you in the first place. The retrieval system sees your profile as natural language text and builds a 3,072-dimensional semantic representation of your professional identity.

- **Ranking stage (GR):** A separate Qwen3 0.6B model reads your profile and generates a dense embedding that late fusion incorporates into GR's ranking process. This embedding doesn't determine which content you see — that's the retrieval layer's job — but it helps GR understand who you are when your behavioral history is sparse or ambiguous.

These are two separate LLM-based representations of your profile, serving two different parts of the pipeline. Neglecting your profile means giving both systems degraded inputs.

Now what?

- **New or returning LinkedIn users: prioritize your profile immediately.** The Qwen3 0.6B profile embedding is GR's primary signal for members with sparse behavioral history. A complete, precise profile is the most efficient way to bootstrap relevant content ranking before you've built a behavioral record.
- **Profile quality compounds.** Because profile embeddings feed both retrieval (Causal LLM) and ranking (GR), improvements to your profile have double-layered effects across the entire pipeline.
- **Write your profile for professional specificity, not keyword density.** Both embedding models are semantically sophisticated — they understand professional

concepts, not just surface-level keyword matches. A profile that clearly communicates your domain expertise, specializations, and professional focus generates a more precise embedding than one padded with buzzwords.

- **Keep your profile current.** Qwen3 0.6B profile embeddings refresh daily. If your profile no longer reflects your current focus area, GR's profile-based context features steer the model with outdated professional identity information.

GR's Multi-Task Prediction: Passive vs. Active Engagement

What happens

GR does not make a single prediction about whether you'll "engage" with a post. It simultaneously predicts multiple, distinct types of engagement — and it treats them differently.

LinkedIn's production model (documented in arXiv:2602.12354v1, Section 4.2.4) uses an **MMoE (Multi-gate Mixture-of-Experts)** prediction head with shared DCNv2 experts. The model groups prediction tasks into two categories:

- **Passive tasks:** click, skip, long dwell (dwelling on a post longer than a threshold duration)
- **Active tasks:** like, comment, share

Each group uses its own gating network to combine the shared experts' outputs. This architecture — where passive and active engagement types get specialized routing while still sharing a common representation — outperformed simpler single-tower approaches in ablation tests across both Long Dwell AUC and Contributions AUC (Table 1, arXiv:2602.12354v1).

The GR paper confirms that the objective function weights active and passive engagement signals to produce the final ranking score. Higher-intent active signals (comments, shares) carry more weight than passive signals in this weighting — consistent with LinkedIn's published understanding that different engagement types represent different levels of professional interest.

So what?

GR optimizes for a weighted combination of engagement types, not any single metric. The model is specifically designed to distinguish between "this person paused to read it" and "this person found it valuable enough to share." Both signals matter to the ranking, but the model weights them differently and predicts them through different pathways.

This has a direct implication for content strategy: **comments and shares generate a qualitatively different signal than likes and clicks, because GR's architecture literally routes them through separate prediction pathways.** Your content's ability to generate active engagement — not just passive consumption — carries compounding signal value.

Now what?

- **Design for active engagement, not passive metrics.** A post that reliably generates thoughtful comments sends GR a richer, higher-weighted signal than a post that accumulates passive likes. Ask a genuine question. Invite disagreement. Pose a genuine dilemma your audience faces.
- **Understand that "time spent" is GR's primary online metric.** The A/B test that confirmed GR as the production ranker measured "+2.10% time spent." The system optimizes for content that causes members to stop and engage meaningfully — not content that gets rapid, reflexive reactions.
- **Don't game engagement types.** LinkedIn's MMoE architecture is specifically designed to detect the signal in each engagement type independently. The system knows the difference between a reflexive like and a considered share-with-comment.

The Recency Effect in GR: Sequence Depth vs. Attention

What happens

GR's causal attention mechanism means recent interactions naturally carry more contextual weight. But this also means that interactions deep in your history — 1,000 positions back — have less influence on today's rankings than your most recent behavior.

Researchers call this effect "Lost-in-Distance" — not a limitation of text comprehension, but a principled design choice rooted in how sequential transformers process ordered data. The phenomenon is real, but the mechanism differs from how it operates in text-based LLMs. In text-based LLMs, the challenge is that important information can get lost when many tokens separate it from a relevant part of a natural language prompt. In GR, older interactions appear earlier in the sequence, where the model processes them with less historical context — that earlier position carries less attention weight relative to recent positions.

Practically: your engagement from six months ago has less influence on what GR surfaces for you today than your engagement from last week. And your engagement from last week has less influence than your engagement from this morning.

This is also why LinkedIn's training uses *recency-weighted loss* during model training: the training process weights interactions by exponential decay from their timestamp, with a 60-day half-life. Training the model this way ensures GR's predictions respond more to recent behavioral patterns than to ancient ones.

So what?

The temporal sensitivity of the GR architecture has important implications for both content creators and for anyone looking to influence their own feed:

- **Recent topic consistency matters more than historical depth.** Six months of consistent engagement on a topic does build a strong behavioral signal — but a sudden shift in the last 30 days of your engagement history will influence GR more than that longer background.
- **Your feed reflects your recent professional focus.** If you've been consuming content about a topic, GR begins routing more of that content to you quickly. The effect is not immediate (the model updates daily through incremental training), but it responds within days, not weeks.
- **For content creators: your audience's recent behavior matters.** GR matches your content to audiences whose recent behavioral sequence makes your post a plausible next item. This means your most likely early amplifiers are people who have been actively engaging with related content recently.

Now what?

- **Before publishing important content, align your recent engagement.** Engage with posts on the same topic in the days leading up to your publication. You are not "gaming the algorithm" — you are creating a coherent behavioral signal that accurately reflects your focus area. GR will use that signal to evaluate your post in a context that favors distribution to aligned audiences.
- **Monitor topic drift in your feed.** If your feed starts showing content from areas you're not focused on, that's behavioral sequence drift — your recent engagement history has pulled GR toward different content domains. Deliberate re-engagement with your target topics corrects the signal faster than simply ignoring the drift.
- **Consistency over time still matters.** While recent behavior has more weight, a long, coherent history of engagement in your domain creates a strong baseline signal that prevents GR from being easily confused by occasional off-topic engagement. Build the long record while staying current.

A Note on 360Brew

LinkedIn also developed and published research on **360Brew** — a large language model built on the Mixtral 8x22B architecture with approximately 150 billion total parameters. 360Brew went through a rigorous three-stage training process — Continuous Pre-Training (CPT), Instruction Fine-Tuning (IFT), and Supervised Fine-Tuning (SFT) — on LinkedIn's proprietary data, with the goal of creating a foundational understanding of the professional world.

The LLM-Ranker approach LinkedIn tested for feed ranking was architecturally similar to what the 360Brew research explored: representing posts and member context as natural language text, constructing prompts, and asking an LLM to predict engagement. As described above, LinkedIn evaluated this approach and found it never outperformed the production model in A/B tests for feed ranking.

360Brew represents serious and impressive research. LinkedIn's 360Brew paper describes deployment across "8+ surfaces" — it may well power recommendations in contexts like job matching or people recommendations, where the challenges of encoding network relationship strength in text prompts are less acute. For those applications, an LLM that deeply understands professional concepts, job titles, skill relationships, and career trajectories has genuine advantages.

But for the feed, where the critical signals are behavioral sequences and network relationship strength, the sequential transformer architecture of GR provided the better combination of online metrics and production efficiency. LinkedIn made the pragmatic engineering choice that the data supported.

Any description of the feed ranker as operating through natural language briefing documents, In-Context Learning from text prompts, or "reading" your profile like an expert consultant describes the LLM-Ranker approach — the architecture LinkedIn evaluated and set aside. GR works differently, and understanding the difference matters for how you should think about your content and engagement strategy.

What This Means for You: The Whole Picture

To bring the architecture back to practical reality, here is how the Generative Recommender affects your LinkedIn presence:

What happens (the full ranking pipeline)

When a candidate post reaches GR for ranking, the model has access to three types of information:

- **Your interaction sequence:** The last 1,000 post-action pairs from your LinkedIn history, processed by the sequential transformer with causal attention
- **Your profile embedding:** A dense vector Qwen3 0.6B generates from your LinkedIn profile, added as a late-fused context feature
- **Candidate and context features:** Information about the candidate post itself (content, author, popularity signals, affinity between you and the author) and other contextual signals, also late-fused

GR combines these inputs to produce simultaneous predictions for multiple engagement types (passive: click, long-dwell; active: like, comment, share). The weighted combination of these predictions produces the ranking score.

So what?

Your profile quality and your engagement behavior both matter to this ranking — but through completely separate mechanisms. Improving your profile text improves the Qwen3 0.6B embedding that serves as GR's context feature. Improving the quality and topical

coherence of your engagement history improves the sequential signal at the core of GR's transformer.

These are not redundant: one shapes your professional identity context; the other shapes the behavioral pattern that GR's sequential model learns from. LinkedIn designed the system to use both for good reason — neither alone is sufficient.

Now what?

Integrate these principles into an updated LinkedIn strategy:

On profile quality:

- Write your LinkedIn profile with the specificity and coherence of a precise professional document. Both the Causal LLM (retrieval) and Qwen3 0.6B (ranking) read it to build a numerical representation of who you are professionally. Vague language, keyword stuffing, and scattered focus all produce noisier, less precise embeddings.
- Keep every section current. Profile embeddings refresh daily in GR; retrieval embeddings refresh within 30 minutes for existing members. Your profile should reflect where you want to be recognized, not where you were two years ago.

On engagement behavior:

- Engage authentically and consistently with your target topic areas. Your interaction sequence is GR's primary input. Build a clear, coherent behavioral record.
- Favor active over passive engagement. Comments and shares pass through GR's active task prediction pathway, which the ranking objective function weights more heavily. A thoughtful comment contributes more to your behavioral signal than a quick like.
- Be deliberate in the days before a major publication. Your recent sequence influences the context in which GR evaluates both your content and your feed. Engaging with related content establishes the behavioral pattern GR will use as context.

On content design:

- Create content that your target professional audience would actively engage with, not just passively consume. GR's optimization target weights time spent and contributions (active engagement) more heavily. Content that prompts genuine response outperforms content designed to look impressive.
- Topical coherence across your posting history helps GR learn the audience for your content. An author who consistently posts about a specific domain accumulates a clearer audience routing pattern than one whose content is unfocused.
- The quality of your content's early engagement shapes future distribution. When early readers leave thoughtful comments and shares, GR registers strong active

engagement signals at the beginning of your post's life, which influences how the model evaluates your content for subsequent distribution decisions.

Architecture confirmed by: arXiv:2602.12354v1 (Hertel, Srivastava et al., February 12, 2026) and "Engineering the next generation of LinkedIn's Feed" (Danchev, LinkedIn Engineering Blog, March 12, 2026). All specifications reflect the production GR/Feed SR system as publicly documented.

Step 3: How Your History Becomes a Signal — Behavioral Context in Retrieval and Ranking

At this point in the funnel, the system has gathered roughly 2,000 candidate posts. Now two AI systems need to make sense of *you* — who you are as a professional, what you care about, and how your interests are evolving. Each system answers this question in a fundamentally different way.

This difference is not a technical footnote. It changes what "optimizing for the algorithm" actually means.

How the Causal LLM Retrieval System Understands You

What happens

The Causal LLM uses genuine text processing to build your professional identity. Before any retrieval can happen, the system constructs a detailed text representation of who you are and what you have engaged with — then runs it through the LLaMA-3 3B language model to produce a dense, 3,072-dimensional embedding vector.

This text-based member prompt draws from two sources:

Your profile: name, headline, summary, industry, skills, location, job history, education history, certifications, and languages spoken. Everything you have written in your profile contributes to this representation.

Your positive engagement history: a chronologically ordered sequence of the posts you have previously taken a "professional interaction" on — a long dwell, like, comment, share, or repost. The system uses only positive engagements. Neutral impressions (posts you were shown but scrolled past without action) are deliberately excluded. LinkedIn's research found that this "positives-only" approach produced substantially better embeddings and required significantly less compute than including all impressed posts (Ramanujam et al., 2025).

Once this text representation is assembled, the LLM processes it through its transformer layers and applies **mean pooling** — averaging all token representations equally — to produce the final member embedding. This is an important architectural detail with a practical implication: mean pooling means every part of your profile and every post in your engagement history contributes roughly equally to the final vector. There is no special "front-loading" advantage for the retrieval embedding itself — the model averages across all tokens. The practical advice to put your most important information first in your profile remains sound, but for a different reason: clear, prominent signaling makes your profile easier for any system to interpret, and ensures your expertise is unambiguous throughout the prompt rather than buried in later sections.

Freshness SLAs: The embedding system maintains tight freshness guarantees. New posts receive an embedding within approximately 1 minute of publication. The system reflects

updates to existing posts (including popularity signals) within roughly 30 minutes. Because Causal LLM retrieval embeddings refresh within 30 minutes of an activity update, a profile change quickly propagates into the retrieval system. The pipeline generates first-time embeddings for new member profiles within approximately 1 minute.

So what?

The Causal LLM is a genuine language-comprehension system. It reads your professional narrative and produces a mathematical representation of your identity that the system compares against the embeddings of hundreds of millions of posts. The quality, clarity, and coherence of your written profile directly shapes what content opportunities the system even considers you for. A vague, incomplete, or internally inconsistent profile produces a weak, diffuse embedding — one that places you near the center of a multidimensional space rather than precisely located within relevant professional communities.

Equally important: only your positive engagement history contributes to your member embedding. Actions you take half-heartedly, or content you passively scroll past, do not add noise to the system. The embedding reflects your most intentional professional behaviors.

Now what?

Your profile text is a direct input to the system that decides which content you ever get a chance to see. Invest in it accordingly.

- **Write for coherence, not completeness**: Your headline, summary, and experience sections should all tell a consistent professional story. If you are a product marketing leader who also does data analytics, choose which identity is primary and structure your profile around it. Conflicting signals produce a diffuse embedding; a focused narrative produces a sharp one.
- **Use the language of your field**: The LLM has world knowledge — it understands semantic relationships between concepts. You do not need to stuff keywords like a 2015 SEO strategy. Write naturally about your work in the vocabulary your professional community actually uses. The model will understand the connections. "Electrical engineer who works on grid modernization" and "power systems professional focused on renewable integration" will map to overlapping regions of the embedding space.
- **Engage with what you genuinely care about**: Because only your positive engagements feed into your member embedding, the quality and intentionality of your engagement history matters far more than volume. Ten thoughtful engagements on your core professional topic create a cleaner, more focused embedding than fifty scattered reactions across unrelated subjects.
- **Keep your profile current**: Because Causal LLM retrieval embeddings refresh within 30 minutes of an activity update, a profile change quickly propagates into the

retrieval system. If your professional focus shifts, update your headline and summary to reflect the change. The system will adapt.

How the Generative Recommender Understands Your Behavioral Pattern

What happens

The Generative Recommender (GR) — the production feed ranking system described in LinkedIn's research as "Feed SR" (Feed Sequential Recommender) — takes a fundamentally different approach. GR does not read a natural-language text description of you. It reads the *sequence* of your behavioral history and learns the pattern of your professional engagement over time.

The sequence architecture: GR processes your interaction history as a single, ordered sequence of interleaved post-and-action pairs. Each historical post in your sequence is represented by a compact set of encoded features: actor and content embeddings, lightweight categorical attributes, and semantic content representations. GR encodes each action you took on that post — long dwell, like, comment, share — as an action embedding. The result is a 2T-length input (alternating post and action representations) fed through multiple transformer layers with **causal attention** (Hertel, Srivastava, et al., 2026).

GR keeps the most recent 1,000 impressions from your history, ordered chronologically from oldest to newest. Candidates to be ranked are appended to the end of this sequence and scored in a single forward pass, which allows the system to process all candidates efficiently without re-running the full history for each post.

What GR does NOT do: GR does not construct a natural-language briefing document, ask itself "will this member like this post?", or evaluate content through any form of reading comprehension. The approach that works this way — an LLM-Ranker that encodes member profiles and candidate posts as text prompts — was explicitly tested by LinkedIn's team and rejected. It "never achieved superior online performance over the existing production model" and "struggled with network-based recommendations, because it was difficult to encode the strength of network relationships in a text prompt" (Hertel, Srivastava, et al., 2026). GR's compact token-pair representation is far more efficient, and its explicit handling of popularity and affinity signals via late fusion produces better predictions.

The sequence engineering challenge: The core challenge for GR's behavioral context is not how many words fit in a prompt — it is which 1,000 interactions best represent your current professional interests. By default, GR uses chronological history. This means that if you have been a highly active member for years, interactions from 1,000 engagements ago may be from a professional context quite different from today. The system addresses this through two training techniques: position-weighted loss (which assigns 50% weight to the oldest position and full weight to the most recent) and timestamp-weighted loss (with a 60-day half-life, so interactions from two months ago receive 50% weight during training).

Together, these ensure the model has learned to prioritize recent patterns when making ranking predictions.

Profile embeddings — a separate, late-fused signal: Your profile information feeds into GR ranking through a different channel than the sequence. A fine-tuned Qwen3 0.6B model generates a dense embedding of your complete LinkedIn profile — all the same profile fields used by the retrieval system — and that embedding is incorporated into GR's prediction head through **late fusion**: concatenated to the transformer output after the sequence processing is complete, then processed through the Multi-gate Mixture-of-Experts (MMoE) prediction head alongside other context features. This design means your profile influences ranking without requiring it to be part of the 1,000-interaction sequence. It is particularly valuable for members with short or sparse histories, where profile embeddings provide the primary basis for personalization (+2% Long Dwell AUC improvement for members with fewer than 10 historical actions).

A note on freshness: Qwen3 profile embeddings for GR refresh on a **daily** cadence (Hertel, Srivastava, et al., 2026). This is distinct from the Causal LLM retrieval system, whose member embeddings reflect activity updates within 30 minutes. A profile change becomes fully reflected in GR's ranking within approximately 24 hours — meaningful, but not immediate.

So what?

The "attention span" problem for GR is not about text context length — it is about sequential depth. GR uses causal attention, which means each position in the sequence can only attend to previous positions. Items deep in your history (1,000+ interactions ago) receive progressively less contextual weight than your recent activity, because the model has been trained to prioritize recency in predicting your future behavior. Recent, topically coherent engagement patterns have the strongest influence on what the ranker surfaces for you next.

This phenomenon — call it the Recency Depth Effect — is not a limitation of text comprehension, but a principled design choice that prioritizes your current professional trajectory over stale historical patterns. If your interests have evolved, your recent engagement history will reflect that, and the ranking system will adapt. If your history is scattered across unrelated topics, the sequential pattern is noisy and the predictions less confident.

Your profile is doubly important: it shapes your retrieval eligibility at Stage 1 (Causal LLM) and provides late-fused contextual grounding at Stage 2 (GR). A well-written, current profile helps the system understand you at both stages of the pipeline.

Now what?

You cannot directly specify which of your interactions GR uses, or in what order. But you have complete control over the quality of the behavioral sequence the system learns from.

- **Prioritize high-intent engagement**: GR encodes not just what you engaged with, but what action you took. A "long dwell" signals sustained attention; a comment signals active professional engagement; a share signals that you found something worth amplifying. These distinctions are encoded directly into the action representation tokens that the model reads. Passive scrolling and reflexive likes create a less informative behavioral sequence than deliberate, substantive engagement.
- **Build a recent, topically focused engagement pattern**: Because position-weighted and timestamp-weighted training give more weight to recent interactions, the last few weeks of your engagement history matter more than your complete lifetime record. If you are trying to shift your professional positioning on LinkedIn — to become known for a new area, or to signal a career evolution — the most effective approach is consistent recent engagement on the topics you want the system to associate with you. History changes one interaction at a time; a deliberate recent pattern reshapes the signal quickly.
- **Engage with content that reflects your aspirational professional identity**: The ranking model uses your interaction history to predict what you will find valuable. If your recent engagements accurately represent your professional interests and expertise, the system's predictions will be calibrated to your actual goals. This creates a reinforcing cycle: engage with what you genuinely care about, and the system serves more of what genuinely matters to your professional growth.
- **Remember that your profile anchors the ranking system for sparse histories**: If you are newer to LinkedIn, or if you have been largely inactive, GR relies more heavily on your Qwen3 0.6B profile embedding to inform its predictions. In this case, a complete, well-written profile is your most leverageable asset — it is the primary signal the ranking model has to work with.

Step 4: Finalization, Diversity & Delivery

After GR performs its ranking pass and returns a relevance score for each of the candidate posts, the core "intelligence" work is complete. The system holds a ranked list, ordered from what GR predicts will be most valuable to you down to least. But the process is not yet finished.

If the system delivered the top-scoring posts directly to your screen without further processing, the result might be highly relevant but also monotonous, repetitive, or unbalanced. You might see five consecutive posts from the same prolific person in your network, or a single trending news event dominating your entire feed. A purely relevance-driven ranked list is not necessarily a healthy or engaging one.

This final stage applies a layer of editorial judgment and platform-wide rules to that ranked list. It refines the raw mathematical output of GR to create a balanced, diverse, and safe experience. This involves applying business rules, enforcing feed diversity, and preparing content for delivery to your device.

Applying Final Business Rules: The Platform Guardrails

What happens

Before the feed renders, the system passes GR's ranked list through a rapid series of automated checks. These checks do not re-evaluate relevance; they enforce platform-wide business rules and policies. This governance layer ensures the feed adheres to community standards and delivers a consistently good user experience.

Key filters include:

- **Trust & Safety Moderation**: The most important guardrail. The system checks every piece of content against LinkedIn's professional community policies. Automated systems — and in some cases human reviewers — identify and remove content that violates these policies, including misinformation, hate speech, and spam. Even a post that scores highly for relevance in GR's output will be removed at this stage if Trust & Safety systems flag it.
- **Impression Discounting**: The system maintains a record of what you have recently seen. If you have already seen a particular post in a previous feed session, the system heavily discounts its score or removes it entirely from your next refresh. This prevents the same content from appearing repeatedly.
- **Frequency Capping (Anti-Gaming Rules)**: These rules prevent any single person or topic from dominating your feed. The system applies constraints such as "do not show a member more than X posts from the same author in a single feed session" and "ensure a minimum gap between posts on the same viral topic." Even if individual posts from a prolific creator all score highly in GR's ranking, these rules prevent feed flooding.

- **Block Lists & Mutes**: Personal preference filters. If you have blocked a member, muted them, or unfollowed them, the system removes their content at this stage regardless of its relevance score.

So what?

Raw relevance does not solely determine what you see. LinkedIn actively intervenes to shape the final feed for health, safety, and a good user experience. The platform has made an editorial judgment that a balanced and safe feed is more valuable over time than one that delivers a firehose of the highest-scoring content.

This also means there are hard limits to visibility for any individual creator. No matter how strong your content is, you cannot appear in someone's feed repeatedly in a short window. The system explicitly prevents it. And no amount of technical optimization can override a Trust & Safety flag.

Now what?

Align your strategy with the intent of these guardrails rather than trying to circumvent them. The guardrails exist to protect the feed experience that creates value for everyone on the platform, including you.

- **Post consistently, not repetitively**: Maintain a reliable posting cadence, but avoid publishing so frequently that you trigger frequency caps for your most engaged followers. Spacing out valuable content over time is more effective than bursting multiple posts in a single morning.
- **Vary your content format and angle**: If you post frequently, vary your topics, formats, and perspectives. This not only keeps your content fresh for your audience but reduces the likelihood that anti-gaming rules treat consecutive posts as repetitive. A mix of original analysis, responses to industry news, questions that prompt conversation, and deeper-dive articles signals a healthy content variety.
- **Play the long game**: The business rules are designed to provide a good experience across weeks and months, not just a single session. Building a loyal following who consistently finds your content valuable is more durable than engineering a single viral moment that the system's guardrails may throttle anyway.
- **Always adhere to LinkedIn's Professional Community Policies**: The fastest path to zero visibility is content that violates LinkedIn's rules. Professionalism, accuracy, and authenticity are not just ethical obligations — they are functional prerequisites for platform reach.

Ensuring Feed Diversity: From Manual Rules to Learned Curation

What happens

Beyond the hard-coded business rules, the system also ensures the feed is topically and structurally diverse. The earlier generation of LinkedIn's ranking system accomplished this

primarily through rigid, rule-based re-rankers. A rule might specify, for example, "ensure a minimum gap of two items between any out-of-network posts."

LinkedIn's current architecture can handle diversity in a more intelligent and adaptive way. The system applies diversity-enforcement logic that considers the overall composition of the feed under construction — not just individual post scores in isolation. This reflects a broader industry direction toward evaluating posts as a set rather than one at a time.

Concretely, the system evaluates the top-ranked posts as a group and can ask:

- "Are too many of these posts from the same author?"
- "Are all of these posts about the same trending topic?"
- "Does this set contain a useful mix of content formats — text, video, articles?"

The system then applies rules and learned models to adjust the final composition: down-ranking a post that is too similar to a higher-ranked post, or surfacing a post that adds unique value or a different perspective. Business rules set hard boundaries; learned signals allow the system to personalize diversity preferences over time. The result, for instance, may be that a member who regularly engages with both core-domain and adjacent-topic content sees that variety reflected — rather than a feed saturated with a single perspective.

So what?

What other content is ranking highly for a member at the same moment can affect the visibility of your own post. Even if your post scores strongly in GR's ranking, the broader context of the feed session can influence whether it rises to the top or gets passed over. **Uniqueness and complementary value matter.** If ten other experts in your field have all published posts on the same breaking news story, the feed-level diversity balancing may down-rank your take on that topic for a particular user — even if it is excellent — in favor of a post on a different valuable topic. Conversely, if your post offers a unique angle or covers an underserved niche, the diversity-enforcement logic may actively boost it to add variety to an otherwise monotonous feed.

Now what?

You cannot control what other content is ranking, but you can control the uniqueness and distinctiveness of your own.

1. **Offer a unique angle when engaging with trending topics**: When commenting on a widely-discussed news item, do not simply restate the consensus view. Provide a piece of data others haven't cited, a contrarian argument backed by evidence, or a translation of the story into a specific domain context (e.g., "here's what this means for healthcare compliance professionals specifically"). This makes your post a natural "diversity candidate" — a post the system's diversity-enforcement logic can surface to balance a feed that might otherwise be dominated by identical takes.

2. **Develop a defined niche**: Deep expertise on a specific topic is a structural advantage under feed-level diversity balancing. Your specialized knowledge represents content that the system cannot easily substitute with something else. Being the authoritative voice on a narrow subject makes your posts reliably valuable for diversifying feeds of professionals interested in that area.
3. **Be aware of the competitive landscape**: If your industry is flooding LinkedIn with content on Topic A, that may be precisely the right time to publish your thoughtful piece on Topic B — a related topic that expands rather than echoes the current conversation. Your content may stand out not only to readers, but to the system's diversity-enforcement layer.

Delivery: Formatting for the Final Destination

What happens

In the final milliseconds of the process, the system hands the curated, ranked, and finalized list of posts to the delivery systems. This stage formats content for your specific device — whether a web browser on a large monitor, an iOS app, or an Android device. Specialized rendering processes take the raw content and prepare it for display, ensuring text wraps correctly, images are sized appropriately, and videos are ready to play. The system delivers the formatted feed to your device and renders it on your screen.

So what?

LinkedIn optimizes not only for relevance but for a good consumption experience on every platform. A post that is difficult to read on a mobile device will likely generate lower engagement than one that is immediately readable — and lower engagement feeds back into the system as a weaker signal over time. Format is not cosmetic; it is functional.

Now what?

Always design your content for easy mobile consumption. A significant share of LinkedIn engagement happens on phones, and the feed reaches professionals in quick, intermittent sessions rather than extended reading sessions.

- **Use short paragraphs**: Break up large blocks of text. One to two sentences per paragraph keeps content scannable on a small screen. Dense walls of text signal effort for the reader before a word has been absorbed.
- **Check your visuals**: If you are creating an image or carousel with text, verify that the font is readable at phone scale. Text that is legible on a desktop monitor can be unreadable at 375 pixels wide. Test on an actual device before publishing.
- **Write concise video hooks**: The first three seconds of a video are critical. On mobile, the default is silent autoplay — your opening must communicate its value visually or through on-screen text before a viewer decides to unmute and stay. Design your video's hook for a silent, small-screen first impression.

The Orchestration Layer: How It All Runs in Real Time

One final engineering detail worth understanding: the full pipeline — retrieval, embedding generation, member context loading, GR ranking, business rules, and delivery — runs in real time every time you open your feed, serving more than 1.3 billion members globally (Danchev, 2026).

LinkedIn achieves this through a disaggregated architecture that separates CPU-bound work (feature processing, business logic) from GPU-intensive work (GR's sequential transformer inference). The GR inference system uses **shared context batching**: because all candidates for a given member share the same interaction history, the system computes the history representation once, then appends all candidates and scores them in a single forward pass using a custom attention mask. A custom CUDA kernel (GRMIS — Generative Recommender Multi-Item Scoring; referred to as SRMIS in the academic paper arXiv:2602.12354v1) implements the specific attention pattern required for this multi-item scoring, delivering approximately 2x speedup over standard PyTorch attention implementation. These engineering choices allow the system to run computationally intensive sequential models at production scale without prohibitive latency.

The practical implication for creators: the entire pipeline completes in well under a second. LinkedIn's feed is not a batch process that runs nightly. It is a real-time ranking system that updates its understanding of your interests with each interaction you take and reflects that in subsequent feed refreshes. Your behavior today shapes what you see tonight.

References

Hertel, L., Srivastava, G., Naqvi, S. A., Kumar, S., Zhang, Y., Ocejo, B., Zelditch, B., Englhardt, A., Cheng, H., Hu, A., Alonso, A., Li, D., Dangi, S., Zhu, C., Zhou, M., Li, W., Huang, T., Borisyuk, F., Parameswaran, G., Tiwana, B., Sankar, S., Lan, Q., Choi, J., & Ghosh, S. (2026). An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking. *arXiv*. https://arxiv.org/abs/2602.12354v1

Danchev, H. (2026, March 12). Engineering the next generation of LinkedIn's Feed. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/feed/engineering-the-next-generation-of-linkedins-feed

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B., Somaiya, M., Simon, L., Byrne, D., Ha, S., Zhou, S., Akterskii, A., Liu, Z., Sriram, S., Xiong, C., Pei, Z., Shao, A., Li, A., Xiao, A., Kolb, C., Kistler, T., Moore, Z., & Firooz, H. (2025). Large Scale Retrieval

for the LinkedIn Feed using Causal Language Models. *arXiv*. https://arxiv.org/abs/2510.14223v1

Section 4: Semantic Positioning — Understanding Your Place in Embedding Space

Before applying the checklists in Sections 5–7, one mental model will make all of them coherent: on LinkedIn, you exist as a point in a high-dimensional mathematical space, and every word you write moves that point. This section explains the embedding space model — the conceptual foundation underlying most of what the retrieval and ranking algorithms evaluate — and what it means for how you should approach your professional content and profile.

It sits between the technical walkthrough of Sections 2–3 and the practical checklists ahead because the tactics only make sense once you see the model they're optimizing for.

From Filing Cabinets to Coordinates

In the old feature-based world, your profile existed as a collection of discrete categories: job title, industry, skills. The system put you in boxes. In the new LLM-powered world, you exist as a point in a continuous, high-dimensional embedding space.

When LinkedIn's retrieval system evaluates your profile or content, it doesn't see categories — it sees a vector. A dense mathematical representation in 3,072-dimensional space that captures the semantic meaning of your entire textual presence. Your name, headline, summary, experience, and every post you write all contribute to where this vector lands.

Think of it this way: instead of being in a filing cabinet labeled "Marketing," you exist at a specific coordinate in a vast semantic universe. Nearby are other professionals whose textual profiles share semantic similarity with yours. Far away are those whose professional language and concepts differ significantly.

This is the coordinate-in-concept-space model. LinkedIn applies this model at two distinct stages of its algorithm — not just one.

Two Embedding Systems: Retrieval and Ranking

Here is something most LinkedIn content advice misses entirely: your semantic positioning affects not just whether content reaches your feed at all, but also how it gets

ranked once it's there. LinkedIn deploys two separate LLM-generated profile embeddings across the two stages of feed processing.

The Causal LLM Embedding: Retrieval Stage

LinkedIn's Causal LLM — based on Meta's LLaMA-3 3B architecture — generates a 3,072-dimensional member embedding from your profile text combined with your positive engagement history. This embedding drives the retrieval stage: when the system searches hundreds of millions of posts to find roughly 2,000 candidates for your feed, it asks "which posts are nearest neighbors to this member's embedding vector?" Your semantic positioning determines which content even enters the competition for your feed (Ramanujam et al., 2025, arXiv:2510.14223).

This is the retrieval layer — the system's first pass at determining relevance.

The Qwen3 0.6B Embedding: Ranking Stage

The ranking layer also uses an LLM-generated embedding of your profile — a distinct system operating independently of the retrieval embedding.

LinkedIn's production feed ranker — called the Generative Recommender (GR) or Feed SR — uses a separate, fine-tuned Qwen3 0.6B model to generate a dense representation of your profile. The GR model integrates this embedding as a late-fused context feature into the sequential ranking model, which is particularly valuable for members with shorter interaction histories. Adding profile embeddings improves Long-Dwell AUC by more than +2% for members with fewer than 10 historical actions (Hertel, Srivastava et al., 2026, arXiv:2602.12354).

This is the ranking layer — the system's second pass at determining what rises to the top.

What This Means for You

Your semantic positioning matters at both stages:

- **It determines who discovers your content** — the Causal LLM retrieval embedding decides whether your posts are retrieved as candidates for other members' feeds at all
- **It influences how your content is ranked** — the Qwen3 0.6B profile embedding helps the GR ranking model score candidates, especially for members whose interaction history is sparse

A clear, coherent professional identity in your profile text isn't just good for findability. It feeds directly into both the retrieval and ranking layers that govern how the entire feed works. For Priya building her company's LinkedIn presence or Jared advising clients on semantic strategy, this is the foundational insight: the investment in profile clarity compounds across both stages of the algorithm simultaneously.

How the Retrieval Embedding Is Built: Mean Pooling

Understanding the mechanics helps you make better decisions about your content.

LinkedIn's Causal LLM uses **mean pooling** to generate embeddings. Given a sequence of tokens, the model computes hidden states for each token, then averages all of them:

$$e = (1/L) \times \Sigma H_i \text{ (summed from } i=1 \text{ to } L)$$

This yields a single vector that represents the holistic meaning of the entire text. The Causal LLM paper explicitly tested different pooling strategies and confirmed that pooling all tokens — mean pooling across the full sequence — performs best (Ramanujam et al., 2025, Table 6).

What this means for your content strategy: Every token in your post contributes equally to the item embedding. There is no advantage to front-loading your topic on the assumption that "the system doesn't read the whole post." It reads all of it, and it weights each word equally in the average.

The practical implication is the opposite of a shortcut: write clearly and on-topic throughout your entire post. Filler words, off-topic tangents, and unnecessary verbosity add noise to your average representation. Clarity and precision from beginning to end ensure that more of your tokens contribute meaningful signal rather than diluting it.

This also has implications for your profile. Every section — headline, summary, experience entries, skills — contributes to your member embedding. Consistent terminology across all sections reinforces your semantic positioning by concentrating related concepts in the averaging process.

Embedding Freshness: How Quickly the System Updates

LinkedIn's embedding system operates on two distinct timelines (Ramanujam et al., 2025):

What Changed	Update SLA
New item published	Within **1 minute** of creation
New member profile created	Within **1 minute**
Existing item receives new engagement	Within **30 minutes**
Existing member's activity updated	Within **30 minutes**

The distinction matters for how you think about the system's responsiveness:

- A brand-new post gets indexed in the retrieval system almost immediately after you publish it
- Your member embedding — shaped by your engagement history — refreshes within 30 minutes of activity

For Priya managing content calendars: new posts enter the system fast. For Jared advising clients: engagement activity shapes their semantic position within the hour. The system is not static; it is continuously responsive to new text and new behavior.

Note: LinkedIn refreshes the Qwen3 0.6B profile embeddings used by the GR ranking model daily, on a longer cycle than the real-time retrieval system.

The Compounding Effect at Scale

At LinkedIn's scale of over one billion members, small positioning differences compound dramatically.

Our research team tested 406 pairs of identical professional content with only the author's name changed, measuring positioning differences in the base LLaMA-3 model (Penn & Robbert, 2025). We found measurable differences in embedding positioning — approximately 0.6 percentage-point deviation in cosine similarity, with a large statistical effect size (Cohen's $d = -0.93$, $p < 0.0001$).

Important caveat: We tested the base LLaMA-3 model, not LinkedIn's production system. LinkedIn's Causal LLM underwent three extensive post-training stages — Continuous Pre-Training on trillions of LinkedIn-specific tokens, Instruction Fine-Tuning on proprietary datasets, and Supervised Fine-Tuning on millions of labeled engagement examples. Each stage substantially reshapes the model's representational patterns. No one outside LinkedIn knows whether this extensive fine-tuning process preserves, amplifies, or mitigates the bias patterns we observed in the base model. Treat this research as indicative of a potential concern to monitor, not as a direct measurement of LinkedIn's production behavior.

What the research does confirm is the sensitivity of the system: the mathematics of embedding retrieval makes microscopic differences in cosine similarity determinative. At LinkedIn's scale:

- Retrieval systems return the top-K nearest neighbors to a member's query embedding
- A fraction-of-a-percentage-point difference positions you marginally closer or farther from target audiences
- These differences determine whether you appear in the 2,000-candidate retrieval pool — or disappear entirely

- The system performs these calculations for every search, every recommendation, every feed generation

---

Cold-Start Vulnerability: Why Day One Matters Most

This is particularly important for new users and new accounts. LinkedIn's Causal LLM research shows that newer members and those with fewer connections benefit most from the LLM-based retrieval system, with a +1.17% increase in Daily Unique Professional Interactions for the low-connection cohort (Ramanujam et al., 2025). However, new users are also most vulnerable to embedding-layer positioning because the system has no behavioral data to provide context signals.

For a new user, your profile text defines essentially your entire identity in the retrieval system. You have no engagement history to provide corrective signal. The quality and clarity of your textual presence matters enormously from day one — before you've posted anything, before you've engaged with anything, your profile alone places you at a specific coordinate in semantic space.

For Jared's agency clients onboarding new LinkedIn company pages, or for Priya's company expanding into a new LinkedIn vertical: the first text you commit to the system sets your initial semantic coordinates. Make them intentional.

The New Optimization Question

In the feature-based era, the question was: *"What keywords should I use?"*

In the LLM era, the question becomes: *"What semantic neighborhood do I want to occupy — and am I sending that signal clearly enough for two separate embedding models to confirm it?"*

This is a fundamentally different framing. You're not trying to match keywords; you're trying to position yourself in concept-space near your ideal audience and the topics you want to be known for. Every section of your profile, every post you write, nudges your embedding vectors — plural now, since both the retrieval embedding and the ranking embedding read your profile — in some direction.

For Marcus evaluating this from an analytics perspective: the mathematical framing is precise. You are optimizing a vector's position in two separate high-dimensional spaces simultaneously. The optimization function is: write text whose semantic content places your profile embedding near the embeddings of the audience you want to reach and the topics you want to represent.

For Priya and Jared working through practical strategy: the application is equally precise. Build a clear, consistent topical identity across your profile and your content. Each post

that coherently extends your established topic cluster reinforces your semantic position. Each off-topic post adds noise to your average.

The Embedding Coherence Principle

Our research on embedding coherence (Penn & Robbert, 2025) found that topical consistency compounds over time. When a member's profile and post history share overlapping semantic content, the mean-pooled embeddings concentrate signal in consistent directions in embedding space — making the member's semantic position both clearer and stronger.

The practical application:

- **Profile and posts should share semantic territory.** If your profile positions you as a B2B SaaS demand generation expert but your posts are primarily about personal productivity, your member embedding receives mixed signals
- **Consistent terminology reinforces positioning.** Using the same conceptual vocabulary across profile sections and posts helps the averaging process concentrate signal rather than scatter it
- **Topic coherence is more important than posting frequency.** A lower volume of topically consistent posts generally produces a clearer semantic position than a higher volume of scattered content

---

The Actionable Summary

As you work through the checklists that follow, keep this mental model in mind: you're not filling out a form. You're authoring the document that determines your coordinates in two separate high-dimensional semantic spaces — one that governs whether your content is retrieved, and one that influences how it is ranked once retrieved.

Every word votes for where you want to exist in those spaces. Choose words that position you closer to your ideal audience, closer to the topics you want your name to evoke, and closer to the professionals you want audiences to discover alongside you.

Write with precision. Write with clarity. Write with intention — and write it throughout your entire post, because every word contributes equally to the average.

References

Hertel, L., Srivastava, G., Naqvi, S. A., Kumar, S., Zhang, Y., Ocejo, B., et al. (2026). An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking. *arXiv*. https://arxiv.org/abs/2602.12354

Penn, C. S., & Robbert, K. (2025). *Gender bias in LLaMA-3 embeddings: Implications for LinkedIn-style retrieval systems* [Research report]. Trust Insights. https://doi.org/10.5281/zenodo.17982122

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B., et al. (2025). Large Scale Retrieval for the LinkedIn Feed using Causal Language Models. *arXiv*. https://arxiv.org/abs/2510.14223

LinkedIn Profile Checklist for Marketers & Creators

LinkedIn's algorithm has fundamentally shifted from recency-based to relevance-based content prioritization — a change that makes your profile the foundation of everything that follows. We've moved from a world of numerical signals to a world of semantic understanding, from a feature factory to a two-stage AI pipeline. These checklists translate that shift into action.

We have completely revised these checklists to align with this new paradigm. They are your practical, step-by-step guides to providing the highest-quality raw materials for both of LinkedIn's AI systems. The new guiding principle: **Communicate your value with clarity, because your profile now serves two distinct AI systems — and both matter for your content's reach.**

New Guiding Principle: Your profile serves as more than a biographical record. It provides the foundational text and signals that power *both stages* of LinkedIn's feed pipeline. Your profile feeds the Causal LLM that determines whether your content is even considered for a viewer's feed, *and* it feeds the Generative Recommender (GR) that ranks content against competitors for viewers who reach the ranking stage. Profile clarity and specificity is doubly important under the current architecture: it shapes both who discovers your content and how that content ranks once discovered.

How Your Profile Affects Both Stages of the Pipeline

Understanding where your profile enters the system helps you optimize it deliberately, not just generally.

Stage 1 — Retrieval (Causal LLM): When LinkedIn's retrieval system evaluates whether to include your content in a viewer's candidate set, it uses a 3,072-dimensional "member embedding" that encodes your professional identity. A fine-tuned LLaMA-3 3B model constructs that embedding from your profile text *and* your positive engagement history. Every field you fill in — headline, About section, job descriptions, skills — contributes to this embedding. A clearer, more topically consistent profile creates a more accurate embedding, which improves the system's ability to match you with the right viewers. *(Ramanujam et al., 2025, arXiv:2510.14223)*

Stage 2 — Ranking (Generative Recommender / GR): After the retrieval stage selects candidates, the Generative Recommender processes your last 1,000+ interactions as a sequential history — interleaved post and action pairs — to predict which content a viewer is most likely to engage with. GR also uses your profile to generate a separate dense embedding via a fine-tuned **Qwen3 0.6B** model. GR integrates this profile embedding as a late-fused context feature, adding profile-based context to the sequential ranking decision. This is especially significant for members with short or sparse interaction

histories, where the profile embedding provides the system with signal it otherwise wouldn't have. *(Hertel, Srivastava et al., 2026, arXiv:2602.12354; Danchev, 2026, LinkedIn Engineering Blog)*

The practical implication: Profile optimization is not just about search or discovery. It directly affects how the *ranking* system understands you as a viewer and as a creator. A vague, keyword-stuffed profile creates noisy embeddings at both stages. A clear, specific, professionally consistent profile creates sharp embeddings that help both systems work in your favor.

Key Concepts *(if you haven't read Sections 2–4 yet)*

Member Embedding: LinkedIn converts your profile text and positive engagement history into a numerical vector that encodes your professional identity across thousands of dimensions. The algorithm uses this to match you with relevant content, jobs, and connections. More precise, topically consistent profiles produce more useful embeddings. *(Ramanujam et al., 2025, arXiv:2510.14223)*

The Retrieval Gate: Before any content can surface in a viewer's feed, LinkedIn's dual-path retrieval system must first select you as a candidate. For viewers in your network, FishDB retrieves content from connections within its 30-day window. For viewers outside your network, the Causal LLM retrieval system uses member embeddings to determine relevance. If neither path selects you, the ranking system never evaluates your content. *(Ramanujam et al., 2025, arXiv:2510.14223)*

The Generative Recommender (GR): LinkedIn's production feed ranking model. GR is a sequential transformer that processes 1,000+ historical member interactions as a causal sequence — interleaved post and action pairs — to predict which retrieved candidates a member is most likely to engage with. It does *not* work by reading natural-language prompts about members; it learns from the pattern of interaction sequences. Your profile enters GR as a dense embedding generated by Qwen3 0.6B, late-fused as context after the transformer layers. *(Hertel, Srivastava et al., 2026, arXiv:2602.12354; Danchev, 2026, LinkedIn Engineering Blog)*

Embedding Freshness: For the retrieval system, the pipeline captures new member profiles in member query embeddings within **1 minute** of creation. For existing members, activities and engagement updates refresh member embeddings within **30 minutes**. For the GR ranking system, the Qwen3 0.6B pipeline refreshes profile embeddings **daily**. *(Ramanujam et al., 2025, arXiv:2510.14223; Hertel, Srivastava et al., 2026, arXiv:2602.12354)*

Positive Engagement History: The record of a viewer's positive engagements — posts they liked, commented on, shared, or dwelled on — which the Causal LLM uses to construct member embeddings and which GR processes as its primary sequential input. LinkedIn's research shows that using only positive interactions (removing negatives) significantly

improves both systems. Your profile helps determine whose positive engagement history overlaps semantically with your content. *(Ramanujam et al., 2025, arXiv:2510.14223; Danchev, 2026, LinkedIn Engineering Blog)*

---

1. Profile Photo & Background Photo

Why it Matters in the Current System

Important clarification: Both AI systems — the Causal LLM and the GR ranking model — process only textual and structured input. Your photos do not enter either AI system's processing pipeline. The Causal LLM paper explicitly confirms it generates embeddings from "only textual input" *(Ramanujam et al., 2025, arXiv:2510.14223)*, and the GR model's Qwen3 0.6B profile embedding is similarly text-based.

However, these visual elements serve as crucial trust and engagement signals for the humans who ultimately interact with your content. GR's primary job is to predict human behavior — specifically whether a member will click, like, comment, share, or dwell on a post. Members trust and engage more readily with profiles featuring professional, high-quality photos. That positive human engagement then becomes the sequential interaction history that GR processes and the positive engagement data that shapes retrieval embeddings. Strong photos generate better human signals, which feed both systems indirectly through the engagement they produce.

What to do

Use a clear, professional headshot and a relevant, high-quality background photo.

How to do it

- **Profile Photo:**
    - Use a high-resolution, well-lit photo where your face is clearly visible.
    - Dress professionally, consistent with your industry and role.
    - Ensure the background is simple and not distracting.
    - Use a real photo. LinkedIn's systems increasingly detect AI-generated or fake images, flagging them as negative trust signals.
- **Background Photo:**
    - Use a high-quality image (1584 x 396 pixels is ideal).
    - Reflect your personal brand, company, industry, or a key professional achievement.
    - If you use text, ensure it's legible on both desktop and mobile devices without being cut off.

---

2. Headline

Why it Matters in the Current System

Your headline is the single most important line of text on your profile for both AI systems.

For the **Causal LLM retrieval system**, your headline is among the first pieces of profile text encoded into your 3,072-dimensional member embedding. The Causal LLM creates your embedding via mean pooling across all tokens in your profile prompt — meaning every word contributes equally to your professional identity signal. Your headline ensures your most important professional terms appear prominently in the profile text the system reads, giving it clear, unambiguous signal from your very first field. A headline like "B2B SaaS Content Strategist | AI in Marketing" immediately establishes the semantic territory your embedding inhabits.

For the **GR ranking system**, your headline is part of the profile text that the Qwen3 0.6B model encodes into the dense profile embedding used as a late-fused context feature. Clear, specific headlines produce more useful profile embeddings, which improve GR's ranking performance especially for members with sparse interaction histories.

Both systems benefit from the same properties: precision, topical clarity, and terminology that matches how your target audience would describe themselves or search for your expertise.

What to do

Craft a concise, keyword-rich headline (up to 220 characters) that clearly states who you are, what you do, and the value you bring.

How to do it

- **Lead with Your Most Important Terms:** Place your 2–3 most important keywords or titles at the very beginning. Leading with your most important terms helps human readers immediately understand your expertise — and clear, natural professional language gives both AI systems the rich context they need throughout.
- **State Your Value Proposition:** Briefly explain the problem you solve or the value you create. Example: "Helping enterprise tech companies build their content engine." This gives both the retrieval LLM and the GR profile encoder rich, conceptual context about your professional purpose.
- **Use the Language of Your Audience:** Think about the terms your ideal connections or clients would use. Use that language in your headline. The retrieval system connects you to viewers with semantically similar professional identities — your headline's terminology shapes which neighborhoods you occupy in embedding space.
- **Keep it Updated:** If your professional focus or key skills shift, update your headline promptly. The retrieval pipeline reflects profile changes in embeddings within 30

minutes for existing members, and the Qwen3 pipeline refreshes the GR profile embedding daily.

---

3. About (Summary) Section

Why it Matters in the Current System

If your headline is the title, your About section is the executive summary of your professional identity — and the largest block of narrative text that both AI systems have to work with.

For **retrieval**, the Causal LLM reads your entire About section as part of the member prompt it uses to generate your embedding. The richer and more specific the content, the more accurately the embedding captures your expertise. For **GR ranking**, the Qwen3 0.6B model encodes all profile text into a dense profile embedding — a well-written About section contributes directly to that encoding's quality and specificity.

A strong About section delivers a rich, conceptual understanding that goes beyond basic keywords, enabling both models to make more accurate associations between your profile and content in your domain.

What to do

Write a compelling, detailed summary that tells your professional story, weaving in your key skills, achievements, and goals naturally.

How to do it

- **Start with a Strong Opening Paragraph:** Your first paragraph should summarize your core expertise and value proposition. A strong opening paragraph helps human readers quickly grasp who you are — and it establishes the thematic focus that carries through the entire section, giving both AI encoders a coherent narrative to work with.
- **Tell a Story with Keywords:** Weave skills into the narrative of your accomplishments rather than listing them. Instead of "Skills: SEO," write "I led the SEO strategy that resulted in a 300% increase in organic traffic for our flagship product." Context and results make embeddings more semantically precise.
- **Quantify Your Achievements:** Numbers add concrete, verifiable data points that signal impact and credibility. "Managed a team of 10" and "grew revenue by $5M" are more semantically meaningful than vague claims about leadership and growth.
- **Mention Key "Entities":** Naming notable companies you've worked with, technologies you've used, or significant projects you've led helps the retrieval system link your profile to relevant nodes in LinkedIn's Economic Graph, strengthening embedding connections.

---

4. Experience Section

Why it Matters in the Current System

The Experience section provides the evidence that backs up your headline and About section claims. Both AI systems parse each job description as text. For retrieval, the Causal LLM builds a chronological narrative of your career trajectory from these entries. For GR ranking, the Qwen3 0.6B profile encoder includes your full work history when constructing the dense profile embedding.

Detailed, achievement-oriented experience entries create more specific and accurate embeddings at both stages — and make your profile more credible to the humans who discover you through those systems.

What to do

Detail each role with achievement-oriented descriptions, using industry-standard language and keywords.

How to do it

- **Link to Official Company Pages:** Always link your role to the correct, official LinkedIn Company Page. This creates a clean, unambiguous link in the Economic Graph.
- **Use Precise Titles and Dates:** Use your exact job title and accurate employment dates. This helps both AI systems build a clear timeline of your career progression.
- **Focus on Achievements, Not Responsibilities Alone:** Use bullet points to describe your accomplishments. Instead of "Responsible for social media," write "Grew social media following by 50,000 and increased engagement by 25% in one year." Use the STAR method (Situation, Task, Action, Result) to frame accomplishments — this provides structured, high-information-density text that both AI systems can process more effectively.
- **Embed Relevant Skills in Each Role:** Naturally weave the specific skills and keywords relevant to each job into its description. This shows both AI systems when and in what context you applied your expertise — providing temporal depth that pure skill listings lack.

---

5. Skills Section (Endorsements & Skill Badges)

Why it Matters in the Current System

The Skills section provides structured, verifiable data points that complement your profile's narrative text. Both the Causal LLM and the Qwen3 0.6B profile encoder process your skills list as part of the text they use to understand your professional identity.

Endorsements from skilled professionals and Skill Badges from LinkedIn assessments add third-party validation signals. While published research does not explicitly document the precise weighting of endorsements within the AI systems, endorsements from recognized experts in the same skill area likely contribute positively to professional credibility signals across the platform.

What to do

Curate a comprehensive list of your most relevant skills, seek endorsements for them, and complete LinkedIn Skill Assessments where possible.

How to do it

- **Pin Your Top 3 Skills:** Place your most critical, relevant skills at the top so they are immediately visible to both human viewers and AI processing.
- **Use Standardized Skill Terms:** As you type, LinkedIn will suggest standardized skills. Use them. This maps your profile cleanly to canonical "Skills" nodes in the Economic Graph, which strengthens the semantic signals available to the retrieval system.
- **Seek Strategic Endorsements:** Ask connections who have direct knowledge of your work to endorse your key skills. Endorsements from recognized experts in the same skill area likely contribute to professional credibility signals.
- **Earn Skill Badges:** Passing a LinkedIn Skill Assessment adds a "verified" credential to your profile. This is a strong credibility signal to both human viewers and the AI systems processing your profile.

---

6. Recommendations

Why it Matters in the Current System

Recommendations serve as qualitative, third-party testimonials in your professional profile. LinkedIn's published research confirms that the retrieval system's Causal LLM encodes profile fields — including your headline, About section, skills, job history, education, and certifications — into the member embedding, and the GR ranking system's Qwen3 model encodes them into the profile embedding. Published sources have not explicitly confirmed whether recommendation text enters these AI encodings.

Research confirms this: recommendations build credibility with human viewers, and that human credibility drives the positive engagement signals — likes, comments, shares, dwell time — that both AI systems learn from. A recommendation that says "She led the migration of our analytics stack to dbt, cutting reporting latency by 40%" generates more trust with the humans who click through to your profile than generic praise, which translates into stronger engagement signals in the data both systems train on.

The recommender's identity also strengthens your connection to them in the Economic Graph, which has independent value for network-based signals in the retrieval pipeline.

What to do

Request and give thoughtful, specific recommendations that highlight key skills and impactful achievements.

How to do it

- **Guide Your Recommenders:** When requesting a recommendation, politely suggest the specific project or skills you'd like them to highlight. A recommendation that says "She led the migration of our analytics stack to dbt, cutting reporting latency by 40%" builds far more credibility with human viewers — and credibility drives the engagement signals both AI systems learn from — than "She's a great team player."
- **Give Detailed, Valuable Recommendations:** When recommending others, be specific. Mention the context of your work together, the skills they demonstrated, and the impact of their contribution. This reflects positively on you as a thoughtful professional and strengthens your connection to them in the Economic Graph.

---

7. Education, Honors & Awards, Certifications, etc

Why it Matters in the Current System

These sections provide additional structured entities and keywords that enrich your profile's context for both AI systems. A certification from a recognized body (like Google, HubSpot, or PMI) or an award from a respected industry organization adds verifiable credibility. The Causal LLM and Qwen3 0.6B profile encoder recognize these entities and incorporate them into your embedding alongside your other profile signals.

Certifications and awards from well-known organizations also create specific entity links in LinkedIn's Economic Graph, connecting your profile to recognized credentials in a way that vague self-descriptions cannot.

What to do

Thoroughly complete all relevant sections with accurate, specific, and official information.

How to do it

- **Be Comprehensive:** List your relevant degrees, certifications, publications, patents, and awards. Each item adds signal.
- **Use Official Names:** Use the exact official names for institutions, certifications ("Project Management Professional (PMP)"), and publications. Link to the issuing organization where possible. Both AI systems recognize official entity names; vague or abbreviated versions may not trigger the same associations.

- **Use Description Fields:** If a description field is available, add context and relevant keywords. Explain what the project was about or what you learned. These descriptions contribute directly to the text that both AI systems process.

---

8. LLM-Optimized Writing Principles

Why it Matters in the Current System

Two distinct AI models process your profile text: the LLaMA-3 3B Causal LLM (which reads your full profile prompt to generate a 3,072-dimensional retrieval embedding) and the Qwen3 0.6B model (which encodes your profile into a dense vector for GR ranking). Every word contributes to how both systems understand your professional identity.

Clear, precise language helps both models build accurate embeddings. Ambiguous or cluttered text creates noise that can push your embeddings in unintended directions — the retrieval system may match you with the wrong audience, and the GR profile vector may provide less useful context for ranking.

What to do

Write your entire profile with both human readers AND the AI systems in mind. Prioritize clarity, precision, and semantic consistency.

How to do it

- **Use Clear, Parseable Language:** Avoid run-on sentences, overly complex structures, or stream-of-consciousness writing. The Causal LLM processes your profile as a sequential text prompt; the Qwen3 0.6B model encodes it as a dense vector. Clear structure helps both build an accurate understanding of your professional identity.
- **Maintain Consistent Terminology:** If you're an "AI consultant" in your headline, use that same term in your About section and Experience descriptions. Don't switch between "AI consultant," "artificial intelligence advisor," and "machine learning expert" unless you genuinely want to occupy all those semantic neighborhoods. Consistency reinforces your positioning in both retrieval and ranking embedding spaces.
- **Prioritize Precision Over Length:** In embedding models, a concise, specific description carries more per-token semantic weight than a vague, expansive one. Every word should earn its place. Ask yourself: "Does this word help the AI understand what I offer and who I serve?"
- **Expand Acronyms and Jargon:** Write "Search Engine Optimization (SEO)" at least once before using "SEO" alone. Both AI systems understand the full form, but explicit expansion removes ambiguity and strengthens the semantic signal.

- **Avoid Ambiguous Terms:** Words like "strategy," "solutions," or "helping businesses grow" are semantically vague. Be specific. Instead of "marketing solutions," write "demand generation campaigns for B2B SaaS companies." Instead of "helping businesses grow," write "increased pipeline revenue by 40% through account-based marketing." Specificity creates sharper embedding positioning in both systems.
- **Write for Cold-Start:** Imagine a reader — or an AI — with zero context about you. Does your profile make sense stand-alone? New connections who haven't engaged with your content yet see you entirely through this textual lens. And for new members, the 1-minute indexing SLA for new profiles means the system is ready to use your profile almost immediately — so write it well from the start.

---

9. Optimizing for Retrieval: The Causal LLM Perspective

Why it Matters in the Current System

Before GR can rank your content, it must first pass through the Causal LLM retrieval gate. This system creates a "member embedding" — a 3,072-dimensional mathematical representation of your profile and engagement history — that determines which content candidates LinkedIn even considers for a viewer's feed.

The same logic works in reverse: when you post content, the retrieval system uses embeddings to determine which members should see it. Your profile serves as the primary source for your member embedding, making it the foundation of all discovery. And because GR also uses your profile (via Qwen3 0.6B) as a context feature in ranking, profile optimization affects both stages.

What to do

Optimize your profile for retrieval discoverability, not ranking quality alone. Your goal: create a clear, strong embedding that accurately represents your expertise and connects you to the right audiences — at both the retrieval and ranking stages.

How to do it

- **Understand the Two-Stage Embedding Foundation:**
    - **Retrieval stage**: The Causal LLM reads your profile text and engagement history to create a 3,072-dimensional embedding. This vector positions you near other professionals with similar expertise and interests. The clearer and more consistent your profile, the more accurate your positioning in retrieval embedding space.
    - **Ranking stage**: The Qwen3 0.6B model encodes your profile into a dense embedding that GR uses as a late-fused context feature. This matters most for viewers and creators with sparse interaction histories, where GR has less

behavioral data to work from. A clear profile compensates for limited engagement data.

- **Single-Topic Density Over Breadth:** For retrieval, concentrated expertise signals work better than diluted generalist profiles. A profile clearly focused on "B2B SaaS marketing" will have a stronger embedding position than one vaguely covering "marketing, sales, and business development." If you have multiple areas of expertise, prioritize the one most central to your goals. The same principle applies for GR's profile embedding — topical clarity produces a more useful profile vector.

- **Strategic Entity Mentions:** The retrieval system learns from associations. Mentioning specific companies, technologies, and industry terms creates stronger embedding connections. "Led growth marketing at Salesforce using HubSpot and Marketo" creates more retrievable signals than "Led growth at a major tech company using marketing automation."

- **Cold-Start Optimization — For New and Returning Members:**
    - LinkedIn's research confirms that the Causal LLM delivers a **+3.29% revenue lift** for users with fewer connections and less engagement history — the group for whom suggested content plays the most vital role. *(Ramanujam et al., 2025, arXiv:2510.14223)*
    - For **new members**: the retrieval pipeline generates your embeddings within **1 minute** of profile creation. Your profile immediately begins generating discovery signals. Write it carefully from day one — the system is ready before you expect it.
    - For **existing members**: activity and engagement updates refresh retrieval embeddings within **30 minutes**. Profile updates to the Qwen3 0.6B GR embedding refresh daily.
    - If you're rebuilding your LinkedIn presence or pivoting your professional focus, treat profile quality as an urgent priority. The faster the system indexes accurate signals about you, the sooner content discovery improves.
- **Engagement-Profile Alignment:** Your member embedding combines both profile text AND engagement history. If your profile says "AI consultant" but your engagement history is full of unrelated content, your embedding becomes incoherent. The retrieval system may struggle to position you correctly, reducing discoverability to your intended audience. GR's sequential model also relies on engagement history as its primary signal — incoherent engagement patterns reduce ranking effectiveness regardless of profile quality. Both systems work best when your stated expertise and actual engagement patterns point in the same direction.

By meticulously optimizing these sections, you author more than a form — you create the foundational document that LinkedIn's two-stage AI pipeline uses to understand you at both the retrieval and ranking stages. You build the professional identity that the Causal LLM uses to match your content with the right viewers, and that the Generative Recommender uses as context when deciding how to rank your content among everything else competing for those viewers' attention.

The profile is where both AI systems meet. Clarity, specificity, and topical consistency are not just good writing — they are the inputs that determine how both systems perform on your behalf.

LinkedIn Content Pre-Launch Checklist for Creators

Spring 2026 Edition

How the System Evaluates Your Content

Before you post, it helps to understand what you are actually optimizing for — and what you are not.

LinkedIn's feed uses a two-stage pipeline. Your content must pass retrieval before the ranking stage ever considers it. These two stages evaluate your content in fundamentally different ways.

Retrieval (Causal LLM): The retrieval system generates a vector embedding of your post by processing the full text of your content through a fine-tuned LLaMA-3 language model and averaging the resulting token representations across every word in your post (mean pooling). The retrieval system then compares this embedding against member embeddings — which LinkedIn builds from each member's profile text and positive engagement history — to identify the roughly 2,000 most relevant candidates from hundreds of millions of posts. *(Ramanujam et al., 2025, arXiv:2510.14223)*

Ranking (Generative Recommender / GR): The GR model does not read your post as a text document. It is a sequential transformer that processes each member's most recent 1,000 feed impressions — posts shown to that member, interleaved with the actions they took on each — as a causal sequence. It learns, from patterns in that behavioral history, which content types and topics are valuable for members with similar behavioral profiles. Your content's text generates the retrieval embedding; the engagement it earns then becomes part of the behavioral patterns the ranking model learns from. *(Hertel, Srivastava et al., 2026, arXiv:2602.12354)*

This distinction matters for how you think about content quality. Clear, focused writing helps you in two ways — by creating a precise, matchable retrieval embedding, and by producing content that earns authentic engagement, which then generates the patterns the ranking model learns from.

Key Concepts *(if you haven't read Sections 2–4 yet)*

The Two-Stage Pipeline: LinkedIn's feed uses two AI systems sequentially. The retrieval stage operates through dual paths: FishDB retrieves content from your connections within a hard 30-day window, while the Causal LLM selects out-of-network content based on embedding similarity — together producing approximately 2,000 candidates. Then the Generative Recommender (GR) scores and ranks those candidates individually for each

viewer. Your content must pass retrieval before ranking begins. A post that fails to match the right member embeddings at the retrieval stage generates zero engagement, regardless of its quality. *(Ramanujam et al., 2025, arXiv:2510.14223; Li et al., 2025, FishDB blog)*

Embedding Coherence: Every post you publish contributes to LinkedIn's ongoing model of the topics you cover. A post that diverges significantly from your established topic area introduces noise into your embedding representation, making the retrieval system less confident about when to surface your content. Consistent topical focus produces cleaner embeddings and more reliable retrieval. *(Ramanujam et al., 2025, arXiv:2510.14223)*

The 30-Day FishDB Window: LinkedIn's connection-based feed retrieval system maintains a hard 30-day data window. The connection feed cannot retrieve posts older than 30 days, regardless of relevance score. The relevance-first ranking shift significantly improves how content surfaces *within* that window but does not extend the window itself. For out-of-network suggested content, LinkedIn has not publicly documented the exact retention window in the embedding-based index. *(Li et al., 2025, FishDB blog)*

Past Interaction Data: The record of a viewer's positive engagements — posts they liked, commented on, or shared — shapes their member embedding in the retrieval system and forms the interaction sequence the GR ranking model processes. When someone comments on your post and you reply, that exchange becomes part of their engagement history, strengthening the signal that your content type is valuable for future retrieval and ranking decisions. *(Ramanujam et al., 2025, arXiv:2510.14223)*

Dwell Time: LinkedIn captures the time a viewer spends reading your post as a training signal. Long dwell forms part of LinkedIn's Professional Interaction definition, which trains the retrieval model — meaning the system optimizes over time to surface content that holds attention. A strong opening that stops the scroll directly influences this signal. *(Ramanujam et al., 2025, arXiv:2510.14223)*

I. Before You Post: Content Strategy & Creation

This phase is about ensuring your content creates a sharp, accurate embedding and gives your audience genuine reasons to engage with it.

1. Topic Selection & Conceptual Alignment

Why It Matters

The retrieval system works through semantic similarity. It generates an embedding of your post and compares it against member embeddings built from profile text and engagement history. A post on a topic that precisely matches your audience's professional identity — and your own established expertise — creates a strong embedding match.

The GR ranking model learns from behavioral patterns: which content types and topics generate authentic engagement from members with particular profile characteristics. A post that earns genuine engagement from the right audience strengthens the signal that similar members should see similar content.

Both mechanisms reward the same thing: clear, genuine alignment between your expertise and your audience's actual professional interests.

What to Do

Strategically choose topics that sit at the intersection of your deep expertise and your audience's real professional challenges.

How to Do It

- **Identify audience pain points:** What are the key challenges, questions, and goals of your target audience? Frame your topics around providing solutions, insights, or new perspectives on these specific challenges. Priya needs posts she can act on with a lean team; Jared needs frameworks he can build into client deliverables; Denise needs credible content that holds up in an educational context.
- **Find your niche intersection:** The most effective content lives where three elements meet: your deep expertise, your audience's needs, and a perspective they can't easily find elsewhere. Instead of broadly discussing "AI in Marketing," consider "How Mid-Sized B2B SaaS Companies Can Use AI to Automate Competitive Analysis." Specificity produces a cleaner embedding and a more targeted audience match.
- **Align with your profile:** The topics you post about should directly reflect the expertise stated in your headline and About section. Consistency between your profile and your content creates a coherent embedding identity that the retrieval system can reliably match to the right audiences.

---

2. Content Format Selection

Why It Matters

Different formats generate different types of engagement signals. The GR ranking model learns from engagement patterns — which means the type of engagement your format generates shapes how the system treats similar content for similar audiences.

Video excels at generating long dwell time. Polls drive rapid, lower-friction responses. Long-form articles become cornerstone evidence of topical authority. Choosing the format that best fits your message also determines what kind of engagement signal you send.

What to Do

Choose a format that suits your message and that your target audience actually engages with. Experiment consistently; the system learns your format-audience fit over time.

How to Do It

- **Text Posts:** Ideal for focused insights, sharp perspectives, and discussion questions. The full text of your post is the primary input for the retrieval embedding — well-structured, topically focused text posts are among the most effective formats for precise retrieval matching.
- **Articles/Newsletters:** Best for establishing deep topical authority. Long-form content provides a rich source of semantic information for the retrieval model and gives readers genuine reasons to dwell.
- **Images/Carousels:** Excellent for making complex information digestible. Provide a strong, topic-specific introductory paragraph — this text is the primary context the retrieval model processes, and clear descriptive text in your caption creates a sharper item embedding than image content alone.
- **Native Video:** Great for building personal connection and capturing attention. Add captions. The system can process your video transcript, so what you say contributes to your content embedding.
- **Polls:** Effective for generating quick, broad engagement. A successful poll generates engagement signals that can raise a post's profile in the ranking model's learning, helping the content gain initial visibility.

---

3. Crafting High-Quality, Engaging Content

Why It Matters

Content quality affects your visibility through two distinct mechanisms that both point toward the same conclusion.

For retrieval: The Causal LLM generates your post's embedding by mean pooling — averaging the token representations of every word in your post equally. This means a post with topically coherent language throughout — not just in the first sentence, but across the entire piece — produces a cleaner, more accurate embedding than a post that meanders or covers unrelated ground. The system does not weight your opening sentences more heavily than the rest; every word contributes. The path to a sharp item embedding is topical focus and clarity from start to finish.

For ranking: The GR model learns from engagement patterns, not from reading your text. High-quality content earns authentic engagement from the right audience — comments from people with relevant profiles, replies that extend the conversation, shares that carry your post into new networks. These engagement patterns become the training signal that teaches the ranking model which members value your content type. Writing that genuinely serves your audience's professional needs produces the engagement patterns that make the ranking model work in your favor.

Write for your reader. When you genuinely serve your reader's professional needs with clear, well-structured content, you generate both the precise retrieval embedding and the authentic engagement patterns that the two-stage system rewards.

What to Do

Create content that is valuable, well-structured, and encourages meaningful interaction. Write for an intelligent professional reader, and the AI systems will follow.

How to Do It

- **Hook attention immediately:** Your first sentence is the most important for human readers. It must clearly state the value proposition and create a reason to keep reading. A scroll-past is a negative signal; stopping the scroll earns dwell time.
- **Maintain topical focus throughout:** Because every word in your post contributes equally to the item embedding (mean pooling across all tokens), topical clarity throughout the entire post produces a sharper embedding than a post that opens strongly and then wanders. A post that stays on-topic from the first sentence to the last generates a more coherent embedding and a better retrieval match.
- **Structure your argument:** Use formatting — bolding, bullet points, short paragraphs — to make your main points easy to follow for both human readers and the retrieval model's text processing.
- **Provide genuine value first:** Your primary goal is to educate, inform, or inspire. Authentic, valuable content tends to generate higher-intent engagement signals (comments, shares) rather than low-effort reactions alone.
- **Encourage discussion:** End with an open-ended question. When a member comments and you reply, you create a conversation thread that becomes part of their engagement history, strengthening the connection between your content type and their profile.
- **Proofread meticulously:** A post with errors signals low quality to human readers, generates weaker engagement, and may produce a noisier item embedding. Professionalism in your prose matters.

II. As You Post: Optimizing for Discovery & Initial Engagement

This phase focuses on packaging your well-crafted content correctly — making it easy for the retrieval systems to identify its topic and connect it to the right audience.

A Note on LinkedIn's Multi-System Architecture: LinkedIn uses separate systems for different stages of content distribution. The Causal LLM handles candidate retrieval for out-of-network content by generating item and member embeddings and comparing them via cosine similarity. FishDB handles connection-based feed retrieval with a 30-day window. The Generative Recommender (GR) then handles ranking, processing each member's sequential interaction history to score and order the retrieved candidates. The

guidance in this section primarily addresses retrieval optimization — getting your post into the candidate pool in the first place.

---

4. Writing Compelling Copy & Headlines

Why It Matters

The text of your post is the literal input to the retrieval embedding system. Clear, engaging copy with relevant topical language creates a precise item embedding that matches accurately against the right member embeddings. A strong opening also earns dwell time, which feeds back into the system as a positive training signal.

What to Do

Craft clear, concise, and compelling text that maintains topical focus throughout and encourages viewers to engage further.

How to Do It

- **Strong opening:** Make the first one or two sentences captivating. They should convey the core value and create curiosity. For connection-based feed content, this is also the first thing human readers see before deciding whether to expand the post.
- **Stay on-topic throughout:** Because the Causal LLM generates item embeddings by averaging all tokens equally, a post that introduces off-topic content in the middle or end produces a less coherent embedding than one that maintains focus from start to finish. If you need to cover multiple angles, consider whether two focused posts would serve you better than one sprawling one.
- **Incorporate concepts naturally:** Weave in the 1–3 primary concepts your audience associates with your topic. Use the terminology your audience actually uses in their own profiles and conversations — this improves the precision of the retrieval match. Do not keyword-stuff; write naturally about the core ideas.
- **Clear call to action (CTA):** What do you want readers to do? A direct question like "What's your experience with this?" or "Share your approach in the comments" explicitly frames the post as a conversation starter.

---

5. Strategic Use of Hashtags

Why It Matters

Hashtags appear in your post text and likely contribute to topic identification, helping the retrieval systems understand the primary topic of your post and connect it to broader interest categories. While the retrieval model can infer topics from your text, hashtags may provide an additional signal during the retrieval phase.

(Note: LinkedIn has not publicly documented hashtag-specific feature extraction in its engineering publications, so we base this guidance on reasonable inference from the system architecture.)

What to Do

Use a small number of highly relevant hashtags that mix broad and niche topics.

How to Do It

1. **Use 3–5 relevant hashtags:** This range provides useful topic signal without diluting the coherence of your post text. Too many hashtags can look spammy and introduce noise into your item embedding.
2. **Mix broad and niche:** Use one or two broad hashtags (e.g., #marketing, #leadership) for wider discovery and two or three niche hashtags (e.g., #productledgrowth, #b2bsaas) to attract a more targeted, high-intent audience whose profiles align with your content.
3. **Avoid irrelevant hashtags:** A mismatch between your content text and your hashtags introduces incoherence into your item embedding. The retrieval model processes your entire post text including hashtags — irrelevant hashtags degrade your embedding quality.

---

6. Tagging Relevant People & Companies (When Appropriate)

Why It Matters

Tagging creates a direct connection between your post and the person or company you tag within LinkedIn's Economic Graph. This can strengthen the retrieval signal by connecting your post to an additional set of network edges. It also triggers a notification, encouraging initial engagement from the tagged party.

What to Do

Tag individuals or companies only when they are genuinely relevant to the content.

How to Do It

- **Relevance is key:** Tag people you are referencing, quoting, or collaborating with. Tag companies you are analyzing or celebrating. Do not tag influencers purely for visibility — both human readers and the system treat this behavior as spam.
- **Notify and engage:** When a tagged person finds your content valuable and engages with it, their interaction enters the engagement history that shapes both retrieval embeddings and the GR ranking model's input sequence for their connections.

III. After You Post: Fostering Engagement & Learning

This phase is about capitalizing on the initial visibility your post receives and generating the strongest possible engagement signals for both the retrieval and ranking systems.

7. Engaging with Comments Promptly & Thoughtfully

Why It Matters

Comments are among the most powerful engagement signals in the system. When someone comments on your post, that action enters their engagement history — which both updates their member embedding in the retrieval system and becomes part of the sequential impression-and-action data that the GR ranking model processes.

When you reply, you extend the conversation. The exchange becomes part of the commenter's positive engagement record, strengthening the signal that your content type is valuable for members with their profile characteristics. Substantive replies also increase dwell time for other readers who follow the thread.

What to Do

Monitor your posts and respond to comments in a timely, substantive manner.

How to Do It

- **Acknowledge all comments:** Even brief acknowledgments signal that your post is active, which can increase dwell time for other readers.
- **Answer questions:** Detailed answers demonstrate expertise and give readers additional reasons to engage.
- **Ask follow-up questions:** Keep the conversation going. Your replies are part of the content — they contribute to the engagement pattern that the ranking model learns from.
- **Foster respectful debate:** Differing professional opinions generate substantive comment threads. A healthy debate produces the kind of high-intent engagement signals that reflect strongest in both retrieval and ranking.

---

8. The Evergreen Content Advantage

LinkedIn's relevance-first retrieval architecture has a profound implication for content strategy: your best content can continue working for you weeks after publication.

Why It Matters

The current architecture rewards quality over timing. When a post earns strong engagement relative to its impressions, the retrieval system can continue surfacing it to relevant members throughout the 30-day FishDB window. A high-quality post can

resurface in feeds days or even weeks later when the retrieval system determines it matches a member's current interests.

What This Changes

- **Quality beats timing:** Initial engagement still matters for generating early signals, but creators no longer need to engineer posts around "optimal" posting times. A genuinely valuable post published at an inconvenient time can still find its audience within the retrieval window.
- **Evergreen content has staying power:** Posts with lasting professional value — frameworks, how-tos, industry analysis — can continue generating engagement throughout their retrieval window. Time-sensitive content ("breaking news") can achieve high relevance scores when matched with members actively tracking that topic, but it has a narrower window of peak relevance compared to evergreen content that compounds value over weeks.
- **Your content library matters:** All your posts within the active retrieval window are potential candidates for resurfacing. This rewards creators who build a consistent library of high-quality content rather than chasing volume.

Important architectural constraint: LinkedIn's connection-based feed retrieval system (FishDB) maintains a hard 30-day data window. The connection feed cannot retrieve posts older than 30 days, regardless of relevance score. The relevance-first ranking approach significantly improves how content surfaces *within* that 30-day window but does not extend the window itself. For suggested content from outside your network, LinkedIn has not publicly documented the exact retention window in the embedding-based index. *(Li et al., 2025, FishDB blog)*

How to Leverage This

- Create content with lasting professional value — insights that will be relevant next week and next month, not just today.
- Don't obsess over posting time. Focus on posting when you can commit to engaging with comments in the first hour, regardless of clock time.
- Aim for 2–3 high-quality posts per week rather than daily low-effort content. The system rewards sustained quality over volume.

---

9. Embedding Coherence — Your Content as Part of Your Identity

Why It Matters

Every post you publish does not just compete for engagement — it contributes to the embedding representation of who you are on LinkedIn. The retrieval system builds your member embedding from both your profile text and your recent content and engagement history. If your profile presents you as an AI marketing strategist but your posts cover cooking, travel, and random observations, you create embedding incoherence. The

system's representation of your topical identity becomes blurred, and your discoverability for the audience you actually want to reach declines.

This same coherence principle applies to the GR ranking model, which uses member profile embeddings (generated from your profile text by a separate language model) as one of its input features. Consistent alignment between your profile positioning and your posting topics reinforces your professional identity at both the retrieval stage and the ranking stage.

What This Changes

- **Topic drift has real costs:** Each off-topic post nudges your embedding position away from your stated expertise area. Over time, this can dilute discoverability with the audience you most want to reach.
- **Content reinforces profile:** Your posts and your profile are not separate signals — they are additive evidence of the same professional identity. Posts that reinforce your profile's positioning strengthen your retrieval match for the right audience.
- **Clarity for both audiences:** Writing that is clear and conceptually coherent serves your human readers and creates the kind of precise item embedding the retrieval system needs to match your content to the right people.

How to Leverage This

- Develop 3–5 core content pillars that align directly with your profile positioning. The majority of your content should reinforce these pillars.
- When you post outside your core topics, do so intentionally. Personal content has value for human connection, but understand the tradeoff with embedding coherence.
- Periodically audit your recent posts: do they collectively reinforce who you claim to be? Would someone reading them understand your professional expertise?
- Use consistent terminology between your profile and your posts. If your headline uses "digital transformation," use that phrase in your content too — vocabulary consistency creates cleaner embedding alignment.

---

10. Optimizing for Retrieval Discovery

Why It Matters

Before the GR ranking model ever scores your content, the Causal LLM retrieval system must select it as one of approximately 2,000 candidates worth considering — from a pool of hundreds of millions of posts. *(Ramanujam et al., 2025, arXiv:2510.14223)* Your content needs to pass this retrieval gate, which works through cosine similarity between your post's item embedding and potential viewers' member embeddings.

Understanding how the Causal LLM generates item embeddings is critical here. It uses mean pooling: it averages the token representations across every word in your post

equally. No special weight falls on the opening sentences at the embedding level — every word contributes to the final embedding vector. This means the path to a precise, matchable item embedding is topical focus and clarity throughout the entire post, not just front-loading your topic in the first sentence.

(The practical advice to front-load your topic remains valuable — for your human readers, who decide in the first two seconds whether to keep reading. But the reason front-loading helps with the AI isn't that the system "stops reading" after the first paragraph. It's that a post with a clear, topically focused opening is more likely to maintain that topical coherence throughout — which produces a sharper embedding. A post that opens strong and then rambles gains nothing from that strong opening at the embedding level.)

What This Changes

- **Item embeddings are holistic:** Your post generates a single embedding vector from the mean of all its token representations. A clear, focused post throughout creates a cleaner embedding that matches more precisely with relevant member embeddings than a post that introduces off-topic material anywhere in the text.
- **Topical vocabulary signals matching:** The specific terminology your target audience uses in their profiles and engagement history influences which member embeddings your post matches. If your audience discusses "product-led growth," using that phrase — rather than a generic alternative like "growth strategies" — improves retrieval precision.
- **Single-focus posts outperform scattered topics:** A post covering one topic deeply generates a sharper embedding than a post touching several loosely related topics. If you have multiple insights, consider multiple posts rather than a single sprawling one.

How to Leverage This

- **Maintain topical clarity throughout:** Make your subject unmistakably clear from the first sentence to the last. Because every word contributes equally to your item embedding, topical drift anywhere in your post degrades the embedding's precision.
- **Use domain-specific vocabulary:** The words you use determine which semantic space your content occupies. Technical terminology, industry-specific phrases, and role-specific language all influence which member embeddings your content matches against.
- **Single-focus posts outperform scattered topics:** One topic, covered well, produces a sharper embedding and a better retrieval match than a post that tries to cover multiple themes.
- **Align with your member embedding:** The retrieval system more frequently surfaces your content to audiences with similar professional profiles to your own. A post about enterprise AI implementation from someone whose profile focuses on enterprise AI retrieves more accurately to that audience than the same post from someone with a consumer marketing profile.

Quick-Reference Summary

What you're optimizing	How the system uses it	Practical action
Post text (all of it)	Mean pooling generates item embedding — every word contributes equally	Write topically focused content throughout, not just in the opening
Profile text + engagement history	Builds your member embedding for retrieval matching	Keep profile aligned with posting topics; engage with topically relevant content
Engagement signals (comments, shares, dwell)	GR ranking model learns from engagement patterns of members with similar profiles	Write content that earns genuine professional engagement, not just reactions
Posting consistency (3–5 pillars)	Reinforces coherent embedding identity	Stay on-topic most of the time; off-topic posts dilute retrieval precision
Content age (30-day FishDB window)	Hard limit for connection-based retrieval	Create evergreen content; quality matters more than posting time

---

By following this checklist, you are systematically creating content aligned with how LinkedIn's two-stage feed pipeline actually works: a retrieval system that matches your post's semantic content to the right audiences through embedding similarity, and a ranking system that learns from the engagement patterns your content generates. Your profile, your posts, and your engagement behavior together form the signals that determine your discoverability — and all of them reward the same underlying practice: clear, focused, genuinely valuable professional content.

Citations:

- *Ramanujam et al. (2025). "Large Scale Retrieval for the LinkedIn Feed using Causal Language Models." arXiv:2510.14223v1.*
- *Li et al. (2025). "FishDB: a generic retrieval engine for scaling LinkedIn's feed." LinkedIn Engineering Blog, November 2025.*
- *Hertel, Srivastava et al. (2026). "An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking." arXiv:2602.12354v1.*
- *Danchev (2026). "Engineering the next generation of LinkedIn's Feed." LinkedIn Engineering Blog, March 2026.*

- *Ramanujam et al. (2025). "Large Scale Retrieval for the LinkedIn Feed using Causal Language Models." arXiv:2510.14223v1.*

LinkedIn Engagement Checklist for Marketers and Creators

Spring 2026 Edition

New Guiding Principle: Your Activity Shapes Two Separate Systems

Every engagement you make — a like, a comment, a share — enters two distinct systems simultaneously. Understanding how each system works changes how you think about strategic engagement.

For retrieval (which content you ever get the chance to see): Your positive engagements build the textual prompt that the Causal LLM uses to generate your member embedding. More topically coherent engagement produces a sharper embedding. A sharper embedding means better retrieval matches — which means more relevant content enters your feed in the first place, and your own content gets retrieved for more relevant audiences. Your embedding updates within 30 minutes of new activity. *(Ramanujam et al., 2025, arXiv:2510.14223v1)*

For ranking (which of those retrieved candidates rises to the top of your feed): the Generative Recommender (GR) processes your entire engagement history as a chronological sequence of more than 1,000 interactions, ordered from oldest to most recent. This is not a reading-comprehension exercise — GR doesn't read your comments as text. It reads the fact that you engaged, and it uses the pattern of what you engage with, how often, and how recently to learn your professional preferences. The most recent interactions sit at the end of the sequence and receive the highest attention weight — the paper confirms the final (most recent) position receives full weight, while earlier positions receive progressively less. Your engagement history is the primary behavioral signal teaching GR what type of content you value. *(Hertel, Srivastava et al., 2026, arXiv:2602.12354v1)*

Both systems reward the same underlying behavior: consistent, topically focused, high-quality engagement. The mechanisms differ; the strategy doesn't.

Key Concepts *(if you haven't read Sections 2–4 yet)*

Member Embedding (Causal LLM): A 3,072-dimensional mathematical representation of your professional identity, generated by LinkedIn's LLaMA-3 3B dual-encoder retrieval system. LinkedIn builds your embedding from your profile text AND your positive engagement history. It determines which content from outside your network gets retrieved for your feed, and reciprocally, which audiences can discover your content.

Embeddings update within 30 minutes of new engagement activity. *(Ramanujam et al., 2025, arXiv:2510.14223v1)*

Generative Recommender (GR) / Feed SR: LinkedIn's production feed ranking model as of 2026. A sequential transformer that processes your last 1,000+ interactions as a causal sequence of interleaved post+action pairs, ordered chronologically from oldest to most recent. GR learns engagement patterns — not comment text. It uses causal attention, which means the most recent interactions (at the end of the sequence) receive the highest weight — specifically, the final (most recent) position receives full weight while earlier positions receive progressively less via position-weighted loss. Your engagement history is not a prompt to be written; it is a behavioral record to be shaped. *(Hertel, Srivastava et al., 2026, arXiv:2602.12354v1)*

Positive-Only Engagement History: The Causal LLM retrieval system builds your member embedding using only positive engagement signals — posts you liked, commented on, or shared. LinkedIn's research found that including only positive engagement in the activity history sequence performed best. This means your positive engagements define who you are to the retrieval system. Deliberate, strategic positive engagement has outsized impact. *(Ramanujam et al., 2025, arXiv:2510.14223v1)*

Professional Interaction (PI) Signals: LinkedIn tracks specific engagement types — Long Dwell, Reacts, Comments, and Reposts — as the core signals the platform uses to train and optimize its retrieval and ranking models. Consistent engagement from the same people over time builds a stronger matching signal between your content and that professional audience. Early engagement in the first hours after posting is particularly important for triggering the system's amplification logic.

GR Sequence Attention: GR orders your interaction history chronologically from oldest to most recent. The most recent interactions sit at the end of the sequence and receive the highest attention weight — the paper confirms the final (most recent) position receives full weight, while the first (oldest) position receives approximately 50% weight via position-weighted loss. This creates a functional recency effect: recent engagement in a topic sends a stronger signal than older engagement in the same topic. Consistent engagement within a professional topic across time creates a recurring pattern that GR treats as evidence of sustained professional identity, not just momentary interest.

Dwell Time: LinkedIn captures the time a viewer spends reading your post as a training signal that improves the ranking model's future behavior. Long Dwell is part of LinkedIn's Professional Interaction definition the platform uses for training the retrieval model — meaning LinkedIn optimizes the system over time to surface content that holds attention. Creating content that earns genuine reading time produces meaningful long-term signal.

I. Quick Daily Engagements (5–15 minutes per day)

Consistency matters, but relevance matters more. These are small, daily actions that keep your engagement history current and topically aligned with your goals.

1. Reacting Strategically to Relevant Feed Content

Why It Matters

Each reaction you make creates a data point in two systems. For the Causal LLM, the system logs your reactions in your member prompt, contributing to your embedding position. For the GR ranking system, each reaction enters your interaction sequence — and because GR orders your history from oldest to most recent, with the most recent interactions receiving the highest weight, recent reactions carry greater influence than older ones on how GR scores today's content. Consistent reactions within your professional domain create a recognizable engagement pattern that GR interprets as evidence of sustained professional interest.

A reaction directly signals "this content is relevant to me." Reacting to content from your target audience or on your core topics reinforces your position in a topically coherent embedding neighborhood — which means the system retrieves you to similar audiences and retrieves similar content for you.

For Marcus (VP Analytics): Every off-topic reaction is a literal noise injection into both your embedding and your GR sequence. What's the signal-to-noise ratio of your last 20 reactions?

For Dr. Raymond (CDO, Healthcare): Healthcare data leaders who engage consistently with clinical informatics content build embeddings and GR sequences that reflect their professional domain — not their personal interests. Strategic engagement is a form of professional boundary-setting.

What to do

Quickly scan your feed and thoughtfully react to 3–5 posts that are highly relevant to your expertise, industry, or target audience.

How to do it

- **Prioritize relevance over volume.** Focus on reacting to posts from key connections, industry leaders, and on topics central to your brand. A single reaction on a highly relevant post sends a better signal than 20 reactions on random content. This holds for both systems: retrieval (sharper embedding) and ranking (cleaner GR sequence).
- **Use diverse reactions for nuance.** Don't only "Like" everything. Using "Insightful" on a data-driven post or "Celebrate" on a colleague's promotion provides a richer signal. Diverse reactions supply more detailed semantic information.
- **Avoid indiscriminate reacting.** Mass-liking dozens of posts in a few minutes dilutes both your embedding clarity and your GR sequence coherence. Be deliberate.

2. Brief, Insightful Comments on 1–2 Key Posts

Why It Matters

A comment is one of the highest-intent engagement signals available. Two things happen when you comment:

For the Causal LLM (retrieval): Your comment creates a high-weight entry in your positive engagement history — the record the system uses to update your member embedding. Because you wrote a comment, the interaction registers as particularly high-intent.

For the GR ranking system: GR records the fact that you commented, and the pattern of what you comment on — which topics, which people, which content formats — trains its understanding of your professional priorities. GR does not read your comment text. It reads the behavioral signal: "this member comments on posts about clinical data governance, health informatics, and AI ethics in healthcare." The pattern is the signal.

The practical implication: a topically relevant comment on a relevant post contributes to both systems simultaneously. It sharpens your embedding and adds a high-weight, topically coherent interaction to your GR sequence.

For Jared (Agency owner): The explanation you give clients is this: "When you comment on ten posts about demand generation in a week, the algorithm doesn't read your ten comments. It learns that you're someone who actively participates in demand generation discussions. That's who it thinks you are. That's who it surfaces you to."

For Denise (VP Education): Engagement that reflects your professional mission — education leadership, learning design, association management — builds the identity signal that attracts the right professional audience to your own content.

What to do

Identify 1–2 highly relevant posts in your feed and add a brief, thoughtful comment that contributes to the discussion.

How to do it

- **Add value, don't merely agree.** Instead of writing "Great post!", expand on a point, ask a clarifying question, or share a brief, related experience. For the embedding, this adds a richer positive interaction to your prompt history. For GR, this confirms topical relevance through a high-intent signal.
- **Use relevant professional vocabulary naturally.** When you comment on a post, the Causal LLM includes the text of that post — not your comment text — in your member prompt. Engaging with posts that use your professional vocabulary reinforces your topical positioning in embedding space. The retrieval system reads

the posts you engage with, not your comment text. GR reads neither — it records the behavioral signal that you commented.

- **Prioritize relevance over recency.** Both systems benefit more from topically focused engagement than from being first. Don't race to comment on the newest content at the expense of commenting on the most relevant content.
- **Keep it professional and constructive.** Your comments are a permanent part of your professional record, readable by both humans and the retrieval system.

II. Focused Daily/Regular Engagements (15–30 minutes per day or several times a week)

These activities require more effort but create stronger, more durable signals in both the retrieval and ranking systems.

3. Participating Actively in 1–2 Relevant LinkedIn Groups

Why It Matters

Group activity sends a powerful signal of deep topical interest. Your interactions within a group — the posts you share, the questions you answer, the discussions you enter — provide a concentrated stream of topically aligned engagement. This concentration benefits both systems: for retrieval, it creates dense topical clustering in your embedding; for GR, it creates recurring topical patterns in your interaction sequence that register as strong professional domain signals.

For GR specifically: consistent engagement in a professional group creates a clearly recognizable pattern. A member who comments in "Clinical Data & Informatics Leaders" weekly signals sustained, credible professional interest — not a passing curiosity. GR uses this pattern as evidence of genuine expertise and professional identity in that domain.

What to do

Identify and actively participate in 1–2 LinkedIn Groups that are highly relevant to your industry, expertise, or target audience.

How to do it

- **Share valuable content.** Post relevant articles, insights, or questions within the group. This establishes you as a contributor.
- **Engage with others' posts.** Like, comment, and answer questions in group discussions. This creates a rich trail of high-intent, topically focused engagement signals in both your embedding history and your GR sequence.
- **Choose active, well-moderated groups.** The quality of the conversation matters. A well-moderated group provides higher-quality engagement context.

---

4. Sending Personalized Connection Requests

Why It Matters

Expanding your relevant network strengthens your position in LinkedIn's Economic Graph and directly shapes both retrieval and ranking. An accepted connection request adds a new source of high-relevance content to your feed — meaning your future engagement opportunities become more topically coherent. The people you connect with become a primary source of content you'll engage with, which in turn shapes your GR sequence and your member embedding.

For GR: the content your connections publish is a primary source of ranking candidates. Connecting with people whose content you genuinely want to engage with creates a virtuous cycle — better candidates, more relevant engagement, stronger GR sequence.

What to do

Send a few targeted, personalized connection requests each week to individuals relevant to your professional goals.

How to do it

- **Always add a personal note.** Explain why you want to connect. Reference a shared interest, a recent post they wrote, or a mutual connection. This dramatically increases the acceptance rate.
- **Focus on mutual value.** Think about what value the connection brings to them as well.
- **Connect with people who engage with your content.** If someone consistently reacts to or comments on your posts, they have already demonstrated an interest in your expertise and are ideal connection candidates. Their future engagement will be high-quality signal.

III. More Involved Weekly/Bi-Weekly Engagements (30–60+ minutes per session)

These high-effort activities create cornerstone assets that generate sustained engagement history and durable embedding signal.

5. Writing and Publishing LinkedIn Articles or Newsletters

Why It Matters

A long-form article or newsletter creates a permanent, high-quality content asset that generates ongoing engagement from your audience. That engagement feeds both systems:

- **For retrieval:** Reader reactions and comments on your article become positive engagement signals in their own member prompts, linking their embeddings to your topical domain.
- **For GR:** Your article creates a sustained engagement source. Members who regularly engage with your articles build a pattern in their GR sequences that signals ongoing interest in your content — meaning GR ranks your future posts higher for them.

A successful newsletter also attracts subscribers — a very strong signal of audience validation that creates recurring, predictable engagement.

What to do

If you have in-depth insights to share, consider publishing LinkedIn Articles or starting a Newsletter on a topic relevant to your expertise and target audience.

How to do it

- **Choose a niche focus.** Consistency is key. A newsletter that consistently delivers value on a specific topic builds a loyal audience and creates a coherent body of work for the retrieval system to match against.
- **Provide substantial value.** Articles should offer deep insights, comprehensive guides, or unique perspectives.
- **Optimize for readability.** Use headings, subheadings, bullet points, and images to break up the text.
- **Engage with comments.** Foster a discussion on your published pieces. The conversation in the comments extends the article's engagement lifespan — and each new comment is a new high-intent interaction entering the commenter's GR sequence.

---

6. Reviewing and Endorsing Skills for Connections

Why It Matters

Endorsing a skill creates a structured data signal that reinforces your position in LinkedIn's Economic Graph. While this action primarily benefits the person you endorse, the activity also signals your own professional domain and engagement within your community. It tells the system: "I am a professional in this domain, and I am qualified to validate the skills of others." This provides a subtle but valuable form of demonstrating professional standing in your expertise area.

What to do

Periodically review connection profiles and endorse skills for which you can genuinely vouch.

How to do it

- **Be authentic.** Only endorse skills you know the person possesses.
- **Focus on key skills.** Prioritize endorsing the most relevant and important skills for your connections.

IV. Professional Interaction (PI) Signals: What LinkedIn Actually Measures

LinkedIn's systems track specific engagement types that constitute "Professional Interactions" (PI). Understanding these helps you focus on signals that matter.

What Counts as a Professional Interaction

LinkedIn's ranking and retrieval systems specifically track:

- **Long Dwell**: Extended time spent reading content (not just scrolling past)
- **React**: Likes, Celebrates, Insightful, and other reactions
- **Comment**: Any text you add to a discussion
- **Repost/Share**: Amplifying content to your network

The Positive-Only History Insight

Critical finding from LinkedIn's research: The Causal LLM retrieval system uses only positive interactions in your member embedding. When LinkedIn tested including negative engagement signals (dismissals, "I don't want to see this") in member prompts, performance degraded. LinkedIn's research found that using only positive engagement in the activity history sequence performed best.

This means:

- Your positive engagements define who you are to the retrieval system
- The system learns from what you approve of, not what you reject
- Deliberate, strategic positive engagement has outsized impact on your member embedding

For the GR ranking system: GR processes your full interaction history as a sequence, but the retrieval system — which determines which content ever reaches GR for ranking — runs on positive-only signals. Getting the retrieval right is foundational.

Dwell Time: The Hidden Signal

Unlike reactions and comments, dwell time is a passive signal you generate just by reading. LinkedIn tracks how long you spend on content. Long dwell signals genuine interest even without explicit engagement.

Implication for creators: Long Dwell is part of LinkedIn's Professional Interaction definition the platform uses for training the retrieval model — meaning LinkedIn optimizes the system to surface content that generates long dwell, among other actions. Creating content that holds attention long enough to generate an explicit reaction or comment produces the strongest, most unambiguous signals for both systems.

V. Retrieval-Aware Engagement Strategy

Your engagement history serves not only for ranking — it functions as a critical component of your member embedding that determines what content gets retrieved for you and whether your content gets retrieved for others.

7. Understanding Your Engagement as Embedding Input

Why It Matters

The Causal LLM retrieval system creates your member embedding from two sources: your profile text AND your positive engagement history (likes, comments, shares). Every engagement you make contributes to your embedding position in semantic space. If you consistently engage with AI content, your embedding moves closer to the AI neighborhood. If you engage with content outside your professional domain, your embedding drifts away from your intended topical territory.

What to do

Treat your engagement activity as a deliberate positioning strategy, not merely a social activity. Your engagements function as votes that shape where you exist in embedding space.

How to do it

- **Curate for positioning.** Before engaging, ask yourself: "Does this engagement move my embedding toward or away from my professional goals?" A reaction on off-topic content actively repositions you.
- **Quality over quantity.** The retrieval system analyzes engagement semantically. Ten thoughtful engagements on highly relevant content position you better than 100 random reactions.
- **Strategic topic selection.** Focus the majority of your engagement on content within your core professional domain. This creates embedding concentration rather than diffusion.
- **Leverage comments for semantic weight.** When you comment on a post, the Causal LLM pulls that post's text into your member prompt — not your comment text, but the topically rich, professional-domain content of the post itself. Comments also generate a higher-intent signal in your GR sequence than reactions alone. This dual benefit makes commenting the highest-ROI engagement action: it

contributes richer topical context to your embedding and registers as stronger behavioral signal for ranking.

VI. Ranking-Aware Engagement Strategy

Your engagement history also serves as the primary input to the GR ranking model. Understanding how GR reads your engagement changes how you think about consistency and recency.

8. Understanding Your Engagement as GR Sequence Input

Why It Matters

The Generative Recommender processes your last 1,000+ interactions as a sequential behavioral record. It does not read your comments as text. It learns from patterns: what topics do you engage with? How often? How recently?

GR orders your interaction history chronologically from oldest to most recent. The final (most recent) position in the sequence receives full weight; the first (oldest) position receives approximately 50% weight. This creates a functional recency effect: if you've engaged heavily with AI leadership content in the past two weeks, those interactions sit near the end of your sequence and carry more influence over how GR ranks today's feed than older engagement from months ago.

The practical insight: your engagement history is a living behavioral profile that GR continuously updates. Every interaction you make either reinforces or dilutes the pattern of professional interest that GR has learned about you.

For Marcus (VP Analytics): Think of your GR sequence as a rolling behavioral regression. The model fits a curve to your engagement patterns. The more consistently you engage within your professional domain, the tighter the fit — and the more accurately GR can predict (and surface) what's relevant to you.

For Priya (Director of Marketing): The daily routine doesn't need to be complex. Fifteen minutes of focused, relevant engagement every day creates a more coherent GR sequence than one hour of scattered engagement once a week.

What to do

Engage consistently within your professional niche. Think of each engagement as a data point teaching GR what kind of professional you are — not what you want to say, but what you consistently return to.

How to do it

- **Consistency over volume.** Regular engagement within your professional domain creates a clear, learnable pattern in your GR sequence. Sporadic bursts followed by silence create noise.
- **Recency matters for ranking.** Because GR places recent interactions at the end of your sequence — where they receive the highest position weight — engaging with topically relevant content shortly before publishing your own posts means those interactions occupy the highest-weight positions in your sequence. GR will evaluate your post in a context where your most recent behavioral signals strongly indicate professional interest in that topic.
- **Pattern coherence.** GR learns from patterns, not individual data points. A single comment on a healthcare data governance post doesn't change GR's understanding of you. Three months of consistent engagement with healthcare data governance content does.
- **Engagement type hierarchy for GR.** High-intent actions (comments, reposts) register as stronger behavioral signals than passive actions (reactions alone). GR learns more from what you invest time in than from what you click quickly.

---

9. The "Warm-Up" Tactic: Engage Before You Post

One of the most tactically valuable engagement practices is timing your own engagement activity before publishing a post.

Why It Works

Engaging with topically relevant content before you publish accomplishes two things simultaneously:

For retrieval (Causal LLM): Your member embedding updates within 30 minutes of new engagement activity. If you engage with five highly relevant posts before you publish, your embedding refreshes to reflect strong, recent topical engagement. When LinkedIn indexes your post and the system evaluates which audiences to retrieve it for, your member embedding will reflect your most current topical positioning.

For ranking (GR): LinkedIn adds your recent engagements to the end of your GR interaction sequence — the highest-weight position. When GR evaluates candidates for your connections' feeds — including your post as a candidate for people who engage with similar content — your recent behavioral signals carry the most influence over the ranking outcome. Members who engage with your post add that interaction to the end of their own GR sequences, where it receives maximum weight in future ranking decisions.

The practical result is the same as the old "warm-up the algorithm" advice, but for accurate technical reasons: the system evaluates your new post in a context where your most recent behavior signals strong interest in the relevant topic.

How to do it

- **20–30 minutes before posting:** Engage thoughtfully with 3–5 posts in your professional domain.
- **Comment at least once.** A comment creates a higher-intent signal than a reaction alone, adding richer topical context to both your embedding prompt and your GR sequence.
- **Stay on topic.** The warm-up effect depends on topical coherence. Engaging with off-topic content before posting dilutes, rather than strengthens, the signal.

VII. Cold-Start Engagement Strategy

New users and those rebuilding their presence face a specific challenge: both systems have limited history to work with.

10. Building a Strong Engagement History Quickly

Why It Matters

LinkedIn's research shows the Causal LLM delivers +3.29% revenue lift for users with fewer connections and less engagement history — the group for whom suggested content plays the most vital role. The data also shows a +1.17% increase in Daily Unique Professional Interactions for users in this low-connection cohort. These are members whose experience the system most dramatically shapes by whether it has enough signal to understand them. Building a coherent engagement history quickly accelerates discoverability. *(Ramanujam et al., 2025, arXiv:2510.14223v1)*

For Dr. Raymond (CDO, Healthcare): In highly specialized fields, cold-start clarity matters even more. A new LinkedIn presence in clinical data analytics that starts with 30 days of highly focused engagement in health informatics groups and with leading clinical data voices establishes embedding positioning that general professionals take years of scattered engagement to approximate.

What to do

If you're new to LinkedIn or rebuilding, prioritize strategic engagement in your first 30 days to rapidly build a coherent member embedding and a clear GR sequence pattern.

How to do it

- **Week 1 focus**: Establish your profile completely, then engage heavily with 5–10 top voices in your specific professional niche. Your initial engagements set the foundation for your embedding and the opening of your GR sequence.
- **Avoid topic scatter.** Resist the temptation to engage broadly. Concentrated engagement in your core area builds a stronger, more directionally clear embedding and a more coherent GR pattern faster than distributed engagement across many topics.

- **Comment over react.** Your early comments provide rich text signals that help the Causal LLM understand you when other data is sparse. The GR system also registers comments as higher-intent behavioral signals than reactions. Make early comments count.
- **Connect strategically.** New connections become sources of content that influences your future engagement history. Connect with people whose content you genuinely want to engage with — they shape your input environment for both systems.
- **Monitor your feed.** As you engage, watch how your feed evolves. If you're seeing more relevant content, your embedding is positioning correctly. If you're seeing random content, your engagement may be too scattered.

Engagement Priorities at a Glance

Activity	Frequency	Retrieval Impact	Ranking (GR) Impact	Effort
Targeted reactions on relevant posts	Daily (3–5 posts)	Medium — adds to positive history	Medium — adds to sequence	Low
Insightful comments on relevant posts	Daily (1–2 posts)	High — rich text enters prompt	High — high-intent sequence signal	Low–Medium
Active group participation	2–3x week	High — concentrated topical data	High — consistent pattern signal	Medium
Strategic connection requests	Weekly	Medium — shapes content environment	Medium — shapes candidate pool	Low
Articles and newsletters	Bi-weekly to monthly	High — generates sustained engagement	High — creates recurring engagement source	High
Skill endorsements	Monthly	Low–Medium — Graph signal	Low	Low
Warm-up engagement before posting	Before each post	High — refreshes 30-min embedding	High — adds high-weight recency to sequence	Low

---

The Strategic Summary

By consistently applying these engagement strategies, you actively and deliberately curate the behavioral record that defines you to both of LinkedIn's core feed systems.

To the **Causal LLM retrieval system**, your engagement history is literal input: it enters your member prompt and shapes your embedding position in semantic space. Topically coherent positive engagement produces an embedding that retrieves you to relevant audiences and retrieves relevant content for you.

To the **Generative Recommender ranking system**, your engagement history is a behavioral sequence: GR processes the pattern of what you engage with, how often, and how recently — and uses that pattern to determine which content rises to the top of your feed. GR doesn't read what you wrote. It learns what you consistently return to.

This approach centers not on "being active" for the sake of it, but on strategic, relevant, and valuable interactions that provide the clearest possible behavioral signal for both systems to understand your professional identity and amplify your voice.

Sources: Ramanujam et al. (2025), "Large Scale Retrieval for the LinkedIn Feed using Causal Language Models," arXiv:2510.14223v1. Hertel, Srivastava et al. (2026), "An Industrial-Scale Sequential Recommender for LinkedIn Feed Ranking," arXiv:2602.12354v1.

A Note on Embedding-Layer Realities

Before we move into the technical specifications, we need to acknowledge something that comes with any LLM-based system: the models encode patterns from their training data, including patterns we might not intend or desire.

What Research Has Shown

Our research examined LinkedIn's Causal LLM retrieval system using the same base model (LLaMA-3) and methodology LinkedIn published. We tested 406 identical professional posts with only the author's name changed and found systematic differences in embedding positioning — approximately 0.6 percentage-point deviation in cosine similarity between paired posts (0.994 vs. expected 1.0), with a large statistical effect size (Cohen's $d = -0.93$, $p < 0.0001$).

Important caveat: We tested the base LLaMA-3 model, not LinkedIn's production system. LinkedIn's Causal LLM underwent three extensive post-training stages: Continuous Pre-Training on trillions of LinkedIn-specific tokens, Instruction Fine-Tuning on proprietary datasets, and Supervised Fine-Tuning on millions of labeled engagement examples. Each stage substantially reshapes the model's representational patterns. We do not yet know whether this extensive fine-tuning process preserves, amplifies, or mitigates the bias patterns we found in the base model. Treat our findings as indicative of a potential concern to monitor, not as a direct measurement of LinkedIn's production behavior.

Our research demonstrates that embedding systems encode semantic information from every text field, including author names. While we focused specifically on gender-coded names, the broader implication is that LLM-based systems process all textual input in ways that may not be immediately obvious.

What This Means for You

Focus on what you control. You cannot change how the model processes your name, but you have complete control over the quality, clarity, and consistency of everything else — your headline, summary, experience descriptions, and content. These represent your highest-leverage optimization points.

Quality compounds. The ~0.6% deviation our research found occurs per-evaluation. At LinkedIn's scale, these small systematic differences aggregate across millions of queries. Your small improvements in text quality aggregate across every ranking decision involving your profile or content.

Semantic positioning now affects both stages. The relevance system that determines what you see in your feed operates through two sequential stages, and your textual identity shapes both:

- **Stage 1 — Retrieval (Causal LLM)**: The fine-tuned LLaMA-3 3B model generates 3,072-dimensional embeddings of your member profile using mean pooling across all tokens. These embeddings determine which content you are ever considered for — the system never ranks content that fails to reach retrieval.
- **Stage 2 — Ranking (Generative Recommender / Feed SR)**: A separate Qwen3 0.6B model, fine-tuned specifically for LinkedIn, generates dense profile embeddings that late-fuse into the Generative Recommender's ranking predictions. Your profile text contributes to both stages independently, through architecturally distinct models.

This means the concern our research raises is not limited to retrieval. If embedding patterns encode systematic differences from textual inputs, those patterns appear at both the retrieval selection stage and the ranking scoring stage, through two different LLM architectures applied to your profile.

New users experience the greatest impact. Without engagement history to provide behavioral signals, new users define themselves almost entirely through their textual presence. This creates both a vulnerability (less data to correct for noise) and an opportunity (your text matters enormously from day one). For the Generative Recommender specifically, the benefit of high-quality profile embeddings is greatest for members with sparse histories: the Qwen3 0.6B profile embeddings deliver more than +2% AUC gains for members with fewer than 10 historical actions (Hertel et al., 2026).

The system continues to evolve. LinkedIn iterates on their systems continuously. Adjustments to relevance-over-recency prioritization demonstrate their responsiveness to user feedback. As these systems mature, the industry actively researches debiasing techniques and fairness improvements.

The Bottom Line

LLM-based systems deliver power but carry imperfections. Understanding their mechanics — including their limitations — helps you make informed decisions about how you present yourself professionally. The strategies in this guide focus on what you can optimize: clear communication, semantic consistency, and high-quality engagement that helps any system (human or AI) understand the value you bring.

Technical Specifications

For reference, LinkedIn's embedding systems operate with the following specifications:

System	Dimension	Pooling	Use Case
Feed Retrieval (Causal LLM — LLaMA-3 3B)	3,072	Mean pooling	Member/content matching for retrieval (out-of-network)

System	Dimension	Pooling	Use Case
Feed Ranking profile feature (Qwen3 0.6B)	Dense vector (dimension not published)	Aggregated profile embedding	Late-fused context feature for Generative Recommender ranking
Job Matching (MixLM)	H (not stated in paper)	Not stated in source paper	Job recommendation ranking
Compressed deployment option	512	Varies	Reduced-cost deployment via Matryoshka learning

Matryoshka Representation Learning can reduce the Causal LLM's 3,072-dimensional vectors to 512 dimensions with minimal recall loss — this is a validated experimental option, though published sources do not confirm the production deployment dimension (Ramanujam et al., 2025). The MixLM paper (arXiv:2512.07846) does not state the explicit hidden dimension H for job matching embeddings.

Key insight: Different LinkedIn products use different embedding configurations. The feed retrieval system uses mean pooling across all tokens for the Causal LLM embeddings. The feed ranking system applies a separate LLM (Qwen3 0.6B) to profile data specifically, late-fusing that embedding as an additional feature after the main transformer processes interaction history. MixLM for job matching uses a mix-interaction approach where the system encodes items as embedding tokens rather than full text sequences; the published paper does not describe the specific pooling method. Both retrieval and ranking respond to the quality and semantic coherence of your profile text — but through entirely different model architectures.

References

Penn, C. S., & Robbert, K. (2025). *Gender bias in LLaMA-3 embeddings: Implications for LinkedIn-style retrieval systems* [Research report]. Trust Insights. https://doi.org/10.5281/zenodo.17982122

Hertel, L., Srivastava, G., et al. (2026, February 12). An industrial-scale sequential recommender for LinkedIn feed ranking. *arXiv*. arXiv:2602.12354v1.

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B., et al. (2025). Large scale retrieval for the LinkedIn feed using causal language models. *arXiv*. arXiv:2510.14223.

LinkedIn Newsfeed Technologies

This section provides a granular, technical outline of the LinkedIn newsfeed generation architecture, which we synthesized from publicly available research papers and engineering blogs. It details the specific systems, models, and data flows that drive the end-to-end process, from offline model training to real-time content delivery.

> **Architecture status note**: Specifications below reflect values at the time of each cited paper's publication. LinkedIn upgrades its serving infrastructure continuously. These figures are point-in-time measurements, not current production specifications.

I. Offline Ecosystem: AI Asset Generation & Training

The offline ecosystem handles all large-scale data processing and model training. It operates on a cadence of hours to months, producing the versioned AI models and embeddings that the online serving systems consume.

A. Pipeline Orchestration & Execution Environment

- 1.1. **Orchestration Platform (OpenConnect)**: LinkedIn built this platform on Flyte for defining, executing, and managing all AI/ML workflows. It replaces the legacy ProML ecosystem.
- 1.2. **Dependency Management**: The platform utilizes Docker containers and resource manifests to decouple component dependencies, enabling rapid iteration and eliminating full-workflow rebuilds for minor changes.
- 1.3. **Compute Environment**: Multi-region, multi-cluster Kubernetes setup with a global scheduler for intelligent routing based on data locality and resource availability (CPU/GPU).
- 1.4. **Distributed Training Frameworks**: LinkedIn primarily uses PyTorch Fully Sharded Data Parallel (FSDP) for large model training and Horovod for certain distributed tasks.
- 1.5. **Resilience**: The platform employs active checkpointing and automated job retries via Flyte to handle node maintenance and infrastructure disruptions, reducing training failures by a reported 90% (LinkedIn Engineering, 2024).

B. The Production Feed Ranker: Generative Recommender (GR) / Feed SR

> **Production status**: Confirmed as the primary member experience on LinkedIn's Feed as of February–March 2026. Replaces the previous DCNv2-based ranker. A/B test: +2.10% time spent over the prior production model (Hertel et al., arXiv:2602.12354v1). Publicly announced as "Generative Recommender (GR)" by

LinkedIn (Danchev, LinkedIn Engineering Blog, March 12, 2026). The academic paper uses the name "Feed SR" (Feed Sequential Recommender).

The Generative Recommender is a transformer-based sequential ranking model. It is architecturally distinct from both the Causal LLM retrieval system and from 360Brew. It does not use natural-language text prompts.

B.1 Architecture

- **Model class**: Decoder-only transformer (sequential recommender, not a text-generation LLM)
- **Attention**: Causal attention mask — each position attends only to preceding positions, modeling the temporal flow of how a member experienced content
- **Normalization**: Pre-LayerNorm
- **Positional encoding**: Rotary Position Embeddings (RoPE, Yang et al.) — delivers +0.20% Long Dwell AUC over learned absolute position embeddings in offline evaluation (Hertel et al., 2026)
- **Input sequence**: 1,000+ historical member interactions represented as interleaved post+action token pairs — **2 tokens per item** (one item embedding token, one action embedding token), compared to the rejected LLM-Ranker approach which represented each post as full natural-language text
- **After transformer blocks**: The model discards outputs corresponding to action positions; the pipeline concatenates remaining outputs with additional context features before the prediction head

B.2 Context Features (Late Fusion)

Late fusion adds numeric signals after the transformer, avoiding any increase in the transformer's dimensionality:

- Count features (engagement counts)
- Affinity features (member-to-content and member-to-author relationship signals)
- Device type

These features are critical for encoding network relationship strength and popularity signals that text prompts cannot represent — the primary reason LinkedIn evaluated and rejected the LLM-Ranker approach (Hertel et al., 2026; see Section B.6 below).

B.3 Member Profile Embeddings (Qwen3 0.6B)

- **Model**: A **Qwen3 0.6B** parameter model, fine-tuned specifically for LinkedIn
- **Function**: Generates a dense embedding of member profile information (headline, experience, skills, education, and related structured profile data)
- **Integration**: Late-fused as an additional context feature after the transformer blocks — not concatenated to the input sequence
- **Refresh cadence**: Daily

- **Value**: Adding profile embeddings improves Long-Dwell AUC by more than **+2% for members with fewer than 10 historical actions** — the model compensates for sparse histories using profile identity (Hertel et al., 2026)
- **Practical implication**: Your LinkedIn profile text contributes to ranking independently of your engagement history, through a different LLM than the Causal LLM retrieval system

B.4 Prediction Head: Multi-gate Mixture-of-Experts (MMoE) with DCNv2 Experts

- **Architecture**: Multi-gate Mixture-of-Experts (MMoE) head with shared DCNv2 (Deep & Cross Network v2) cross-network experts
- **Separate gating networks**: One gate for passive tasks (click, skip, long-dwell); a separate gate for active tasks (like, comment, share). This separation allows the model to weight experts differently depending on the prediction target
- **Multi-task output**: Predicts probabilities for multiple user actions simultaneously
- **Training loss**: Binary cross-entropy with recency weighting (more recent interactions given higher weight in the loss function)

B.5 Custom Serving Infrastructure

Feed SR/GR does **not** use SGLang. It runs on a purpose-built custom serving stack:

- **Architecture**: Disaggregated CPU/GPU architecture with two primary components:
  - **Inference Driver** (CPU-based): Handles feature fetching, feature tracking, and CPU-bound transformations (not GPU-friendly operations). LinkedIn optimized member history parsing from 450ms to 2ms (225x speedup); the pipeline reduced sparse-to-dense tensor conversion from 254ms to 5ms (50x speedup)
  - **PyTorch Inference Server** (GPU-based): Python-based service optimized for GPU execution; exposes a high-performance gRPC interface wrapping Apache Arrow buffers inside protobuf messages, enabling zero-copy conversion of large data payloads to PyTorch tensors
- **Custom attention kernel**: **SRMIS (Sequential Recommender Multi-Item Scoring)** — a specialized CUDA kernel extending Flash Attention to support Feed SR's multi-item scoring pattern. Unlike standard approaches requiring explicit mask tensors, SRMIS accepts two scalar parameters (`context_length` and `candidate_length`) and implements attention masking directly within the Flash Attention computation. This eliminates mask materialization overhead and enables online (streaming) softmax
- **Multi-item scoring**: The SRMIS kernel scores all candidates concurrently while preventing cross-candidate leakage. For typical workloads with approximately 500 candidates and a history length of 1,000, this delivers an **80x speedup** on the transformer forward pass compared to non-batched approaches
- **Operational flexibility**: CPU and GPU components scale independently based on their respective bottlenecks

Note on naming: The task instructions and some documentation refer to this kernel as "GRMIS" (Generative Recommender Multi-Item Scoring). The arXiv:2602.12354v1 source paper uses "SRMIS" (Sequential Recommender Multi-Item Scoring). Both names refer to the same custom Flash Attention kernel for Feed SR/GR multi-item scoring.

B.6 The LLM-Ranker Was Evaluated and Rejected

Before Feed SR, LinkedIn built and tested an **LLM-Ranker** in which all features of a candidate post were represented as text and passed into an LLM as part of a prompt — structurally equivalent to the approach described in the 360Brew paper. LinkedIn fine-tuned the model to output "Yes" or "No" for each candidate.

LinkedIn rejected the LLM-Ranker for feed production for three reasons documented in arXiv:2602.12354v1, Section 5.1:

- **Feature encoding**: Encoding numeric features (engagement counts, affinity scores) as text with sufficient precision proved difficult
- **Sequence length and cost**: The text representation of each post was much longer than Feed SR's 2-token compact representation, making long-history processing prohibitively expensive
- **Network recommendations**: "The LLM-Ranker never achieved superior online performance over the existing production model. Although it performed well on posts from out-of-network recommendations, the model struggled with network-based recommendations, because it was difficult to encode the strength of network relationships in a text prompt." (Hertel et al., 2026, Section 5.1)

Feed SR's late-fusion architecture explicitly solves the third problem by incorporating affinity features outside the transformer — something a text-prompt ranker cannot do naturally.

Sources: Hertel, L., Srivastava, G., et al. (2026, February 12). An industrial-scale sequential recommender for LinkedIn feed ranking. arXiv:2602.12354v1. Danchev, H. (2026, March 12). Engineering the next generation of LinkedIn's Feed. LinkedIn Engineering Blog.

C. 360Brew Foundation Model

Production status for the feed: 360Brew's approach to feed ranking (the LLM-Ranker paradigm) was explicitly evaluated and rejected — see Section B.6. The 360Brew paper's own label of "pre-production model" was accurate for feed ranking. However, LinkedIn states 360Brew is applied across "8+ surfaces"; it may be active for job recommendations, people recommendations, notification ranking, and other surfaces. The paper (arXiv:2501.16450v4) was recalled from arXiv for IP licensing reasons; this appears to have been due to concerns about proprietary information disclosure, not evidence of feed production deployment.

- **2.1. Base Model**: LinkedIn builds 360Brew on the **Mixtral 8x22B** pre-trained Mixture-of-Experts (MoE) architecture. Total parameters: approximately **150B**. This is a decoder-only MoE Transformer.
- **2.2. MoE Routing**: Mixtral 8x22B uses **8 experts per layer with top-2 routing** — 2 experts are active per token. Note: A separate LinkedIn Foundation Model described in arXiv:2502.14305v2 uses 16 experts with 4 active per token, but this is a different model — see Section D.
- **2.3. Training Stage 1 — Continuous Pre-Training (CPT)**: LinkedIn further pre-trains the base model on trillions of tokens of verbalized, first-party data (member profiles, interactions, Economic Graph data) to imbue it with domain-specific knowledge.
- **2.4. Training Stage 2 — Instruction Fine-Tuning (IFT)**: The training pipeline fine-tunes the model on a blend of open-source and proprietary instruction datasets using preference alignment algorithms like DPO (Direct Preference Optimization) to enhance instruction-following and zero-shot reasoning capabilities.
- **2.5. Training Stage 3 — Supervised Fine-Tuning (SFT)**: The pipeline fine-tunes the model on millions of labeled examples in a Multi-Turn Chat (MTC) format to learn specific ranking and recommendation tasks. The loss function combines prompt loss and masked MTC loss with weighted terms.
- **2.6. Final Artifact**: A frozen, versioned 360Brew Foundation Model with approximately 150B parameters, applicable across multiple LinkedIn surfaces.

D. SLM Deployment Strategy (Production Details)

LinkedIn's SLM compression research (Behdin et al., 2025, arXiv:2502.14305v2) covers a Foundation Model for RecSys that is architecturally distinct from 360Brew V1.0. This FM uses a Mixture-of-Experts architecture **motivated by the Mixtral family but initialized from Llama 3.1 8B Instruct**, with **16 experts in total (4 active per token)**. This is the model for which LinkedIn published compression and serving details.

> **Note on expert counts**: The 360Brew paper (arXiv:2501.16450v4) describes a model built on Mixtral 8x22B, which has 8 experts per layer with top-2 routing. The SLM compression paper (arXiv:2502.14305v2) describes a different Foundation Model initialized from Llama 3.1 8B with 16 experts (4 active). These are two distinct models. The 16-expert/4-active configuration does NOT apply to 360Brew V1.0.

- **Model Size Ladder**: LinkedIn runs multiple model sizes in production for different latency requirements:
    - Full Foundation Model: ~150B parameters (per 360Brew paper) / MoE FM described in compression paper: ~16 × 8B architecture
    - Compressed variants: 8B → 6.4B → 3B → 2.4B → 2.1B parameters

    - LinkedIn achieves **20x+ model size reduction** through the full compression pipeline (from the ~150B foundation model to 2.1B compressed variants); the structured pruning stage alone compresses an 8B distilled model to 2.1B (~3.8x reduction)
- **Compression Workflow**:
    - Step 1: The pipeline distills knowledge from the foundation model
    - Step 2: The pipeline applies structured pruning via the OSSCAR algorithm (gradual pruning preferred over one-shot for quality retention)
    - Step 3: The team fine-tunes the compressed model to recover performance
    - Result: Structured head pruning delivers 40% attention latency improvement and 28%+ prefill speedup
- **Hardware Specifications** (per serving node, from arXiv:2502.14305v2 as of publication date):
    - 256 CPU cores
    - 2TB host memory
    - 8 NVIDIA H100 GPUs
    - SGLang v0.4.1 as serving engine (at time of publication; upgraded subsequently — see Section D.2)
    - FlashInfer as attention backend
    - FP8 quantization for weights and activations
    - RadixAttention for prefix caching ("hot prefill" for shared member context)
- **Performance Benchmarks** (from arXiv:2502.14305v2, Table 8):
    - P99 Time-to-First-Token (16k context, tp=4 GPUs, m=1, k=1):
        - Full FM: 1,039ms
        - 3B model: 209ms
        - 2.1B model: 184ms
    - Throughput: 14k–115k tokens/sec depending on model size and batching configuration
- **Training Efficiency Optimizations**:
    - Liger Kernel: 20% training time reduction, 60% memory reduction
    - ZeRO++: 2.4x speedup over vanilla ZeRO

Strategic implication: LinkedIn does not run "one big model" for all use cases. The compression ladder allows latency-sensitive applications to use smaller, compressed models while reserving the full foundation model for higher-stakes decisions or batch processing.

D.2. SGLang for LLM-Based Search Ranking (Ramachandran et al., 2026)

Confirmed scope: SGLang serves **AI Job Search** and **AI People Search** ranking. SGLang does NOT power feed ranking — the production Feed SR/GR model uses its own custom PyTorch inference server with the SRMIS kernel (see Section B.5).

> The Scaling LLM blog conclusion states explicitly: "At LinkedIn, these advancements power AI Job Search and AI People Search to deliver state-of-the-art LLM ranking to millions of members."

Products: AI Job Search (cross-encoder SLM ranking) and AI People Search ranking

Architecture: Prefill-only ranking — no decoding; long shared prefix; high concurrency; strict P99 latency SLAs. This represents a different workload class from generative LLMs.

4-Stage Optimization Framework:

- **Stage 1 — Batch Tokenization**: Sequential tokenization consumed hundreds of milliseconds as a CPU bottleneck before GPU involvement. LinkedIn implemented in-request batch tokenization (PR #5141) and Async Dynamic Batch Tokenizer (PR #9382). "Batch send" (PR #9436) transmitted the entire tokenized batch as a single ZMQ message, preserving batch boundaries through the scheduler; this change alone reduced average latency from 70.39ms to 41.12ms per request — a **41.5% reduction** for 300-token/50-item batches.

- **Stage 2 — Scoring-Only Fast Path**: A dedicated scoring API (PR #6460) skips the decode/sampling loop and extracts only final-token logits. LinkedIn tightened CPU-GPU synchronization (PR #8840, #9748) by skipping per-token log-probability extraction and replacing fragmented GPU→CPU memory copies with a single vectorized gather. Combined result: **13.7x P99 improvement** (6,220ms to 454ms on a 0.6B model at 100 QPS), **25% throughput increase**.

- **Stage 3 — In-Batch Prefix Caching**: The engine reuses the query-prefix KV cache within a single forward pass. The prefix KV is computed once using the first prompt in the batch; remaining items reuse that KV directly via attention merging (combining prefix attention and suffix attention using log-sum-exp). This delivers throughput comparable to Multi-Item Scoring: **~2,200 items/s/GPU vs. MIS ~2,100 items/s/GPU** on a pruned 0.4B model. The key difference from MIS: in-batch prefix caching operates at a higher abstraction level in the execution stack without specialized kernel-level masking, preserving standard batched inputs and requiring no concatenation.

- **Stage 4 — Python Runtime Optimization**: Three sub-optimizations:

    - **GC Freeze** (PR #9241): `gc.freeze()` prevents Python's generational garbage collector from scanning long-lived objects, eliminating 100–300ms periodic stalls under sustained load
    - **Multi-process gRPC**: Multiple gRPC servicer processes handle network I/O and request preprocessing while dedicated SGLang engine processes execute inference; requests pass between processes via ZMQ. This decouples request handling from the GIL bottleneck

- **Multi-process scheduler parallelization**: Multiple SGLang scheduler processes per GPU (e.g., 2 workers), with GPU memory partitioned across schedulers. Together with multi-process gRPC, this delivered an **additional ~40% throughput increase** beyond what a single Python process could sustain

Production Results (H100 GPUs, P99 ≤ 500ms SLA):

Workload	Model	Query / Item Tokens	Baseline	Optimized	Improvement
Text-based ranking	375M decoder-only ranker	50 / 150, batch 50	750 items/s/GPU	2,200 items/s/GPU	~3x
Mixed-input ranking	0.6B decoder-only ranker	60 / 1 embedding + 1 special token, batch 50	10,000 items/s/GPU	22,000 items/s/GPU	~2.2x

Relationship to Turbocharging Paper (Shimizu et al., 2025): The December 2025 Turbocharging paper covered recommendation systems (feed personalization via LLM-for-ranking components): Multi-Item Scoring (69% latency reduction), FlashAttention 3, FP8 kernels, Knock-Knock. The February 2026 Scaling paper covers search ranking (AI Job Search, AI People Search): 4-stage optimization framework above. Both use SGLang. The February 2026 paper explicitly references Turbocharging's MIS (PR #10979) as a parallel approach to its in-batch prefix caching. Both papers share co-authors (Qing Lan, Sundara Raman Ramachandran), confirming a continuous optimization effort by an overlapping team.

Stated next steps: LinkedIn engineers committed to "going deeper into the stack with fine-grained profiling, kernel-level tuning, and further trimming overheads in prefill and attention paths" and published a public SGLang Prefill-Only Roadmap for community contributions.

Source: Ramachandran, S. R., Lan, Q., Nguyen, C., Sheng, J., & Zhu, C. (2026, February 20). Scaling LLM-based ranking systems with SGLang at LinkedIn. LinkedIn Engineering Blog.

E. MixLM: Full-Traffic Job Search Ranker

Production status: Confirmed as the full-traffic deployed job search ranker as of November 2025. A/B test result: **+0.47% Daily Active Users (DAU)** in online tests. Served via SGLang.

MixLM (arXiv:2512.07846v1, Li et al., 2025) is LinkedIn's production semantic job search ranker, enabling the first full-traffic LLM-ranking-based job search deployment.

The core problem MixLM solves: Traditional LLM cross-encoder rankers require full text of both query and candidate in each forward pass. For job search, this creates long context prefill-heavy workloads that exceed production latency budgets, limiting LLM ranking to low-traffic experiments.

Text-Embedding Mix-Interaction approach:

- Encodes all items in the catalog into a small number of embedding tokens (compact representations), which MixLM stores in a nearline cache
- Online inference uses these cached embedding tokens rather than full item text, dramatically reducing context length
- Input: approximately 900 tokens of job/member text condensed to 2 tokens per item (1 embedding token + 1 special token)
- **Compression factor**: approximately 450x — 900 tokens → 2 tokens per item
- **Throughput**: 22,000 items/second/GPU — a **10x speedup** over a summarized-text LLM baseline (2,200 items/sec/GPU) and approximately 75x over a full-text LLM baseline (290 items/sec/GPU)

Architecture:

- Uses 0.6B parameter encoder and ranker models
- Pooling method: not specified in the source paper (arXiv:2512.07846v1)
- Enables standard batched inference via SGLang without specialized masking

Deployment path: Prior text-only LLM-based job search operated only at limited-scale traffic. MixLM's efficiency gains enabled full-traffic deployment at production scale, resulting in the +0.47% DAU improvement.

Source: Li, G., He, R., Jing, S., Behdin, K., et al. (2025, November 25). MixLM: High-throughput and effective LLM ranking via text-embedding mix-interaction. arXiv:2512.07846v1.

F. Ancillary Model Training & Asset Generation

- **3.1. Candidate Generation Model (Cross-Domain GNN)**: LinkedIn trains a Graph Neural Network on a unified, heterogeneous graph that consolidates data from multiple domains. Published research on this architecture (arXiv:2506.12700) focuses specifically on LinkedIn's notification system, with member embeddings noted as reusable across surfaces including feed. The member embedding refresh cadence for the GNN system is daily — distinct from the sub-30-minute SLAs of the Causal LLM retrieval system.
- **3.2. Efficient Model Generation (SLMs)**:

    - 3.2.1. **Knowledge Distillation**: LinkedIn uses a large foundation model (teacher) to train a smaller, more efficient model (student), often by minimizing the KL divergence between their output logits.
    - 3.2.2. **Structured Pruning**: The compression pipeline applies the OSSCAR algorithm for one-shot or gradual pruning of MLP layers and attention heads, creating a smaller model (e.g., from 8B to 6.4B parameters) that the team then fine-tunes via distillation to recover performance.
    - 3.2.3. **Final Artifact**: A set of versioned Small Language Models (SLMs) for deployment in latency-sensitive or cost-constrained scenarios.

II. Real-Time Data Infrastructure

This layer captures and serves the most recent member activity, which the online systems require for embedding freshness and interaction-sequence currency.

A. Event Streaming & Ingestion

- 1.1. **Event Bus**: The system publishes all client-side interactions (impressions, clicks, dwells, comments, shares) as events to a Kafka stream.

B. Real-Time Data Storage & Serving

- 2.1. **In-Memory Datastore**: Real-time processing systems consume the Kafka stream and write the recent interaction data into low-latency, key-value stores like Pinot or Venice.
- 2.2. **Function**: This datastore serves as the source of truth for member interaction history used during online inference — including the interaction sequences that the Generative Recommender processes and the member prompt history that LLM-based retrieval uses.

C. Feed Retrieval Infrastructure (FishDB)

- **3.1. Architecture**: FishDB is a **Rust-based generic retrieval engine** that replaces the legacy FollowFeed system. "Generic" is precise: FishDB is not purpose-built for feed retrieval; it is designed to serve multiple recommender system use cases (Li et al., 2025).
- **3.2. Efficiency**: FishDB delivers 2x efficiency improvement with 50% hardware reduction compared to the previous system (Li et al., 2025).
- **3.3. Data Model**: FishDB implements a four-component index:
    - **Forward Index**: Direct lookup of content by ID for filtering and scoring.
    - **Inverted Index**: Actor timelines for efficient feed construction; implemented as an in-memory hashmap using Rust's dashmap concurrent hashmap library.

    - **Reference Index**: Graph-based filtering allowing documents to reference other documents, enabling relationship-based content discovery.
    - **Attribute Stores**: RocksDB-backed key-value stores with row-level bloom filter and LRU cache so that most single key lookups avoid hitting RocksDB entirely. Used for sparse data including document embeddings as ranking features and spam classification features as sparse filtering attributes.
- **3.4. Content Window**: FishDB maintains a **30-day content window** — the system retrieves feed data only from the past 30 days, which fits entirely in memory across the 48 partitions. This is an architectural design constraint, not a ranking preference: FishDB does not index content older than 30 days, so the pipeline cannot retrieve it through this path.
- **3.5. Performance**: P99 latency of **40ms** for connection-based content retrieval (99% of queries complete in under 40ms while supporting twice the QPS on a single host).
- **3.6. Scale**: 16 replicas × 48 partitions per replica.
- **3.7. Scope**: FishDB handles content from the user's direct network (connections, followed creators, followed companies, subscribers) — the "Path A" content that represents your defined network.

C.1. Freshness SLAs (Causal LLM Retrieval System)

Distinct SLAs for new vs. existing content (Ramanujam et al., 2025):

Event	SLA
New post created	Within **1 minute** of posting
New member profile created	Within **1 minute**
Interaction updates (likes, comments) on existing items	Within **30 minutes**
Existing member activity → member embedding refresh	Within **30 minutes**

D. LLM-Based Embedding Retrieval (Causal LLM)

- **4.1. Architecture**: LinkedIn fine-tunes **LLaMA-3 (3B parameters)** as a dual encoder for generating member and item embeddings. A single shared LLM encodes both members and items into a shared embedding space.
- **4.2. Pooling**: Mean pooling across all tokens in the sequence — the representation averages information from every token equally.
- **4.3. Matryoshka Learning**: Matryoshka Representation Learning (MRL) enables nested, size-adaptive representations by optimizing multiple sub-representations simultaneously. The full dimension is 3,072; reducing to 512 dimensions shows minimal recall loss (Recall@10: 0.4225 vs. 0.4242 for full 3,072). The 512-dimension

option is a validated experimental configuration, not a confirmed production deployment dimension.

- **4.4. Training**: LinkedIn trained on 5 million member-item pairs from public engagement in the LinkedIn Feed, using 8 H100 GPUs per training run with per-GPU batch size of 4.
- **4.5. Positive Engagement Only**: LinkedIn found that using only positive engagement events (likes, comments, shares, long dwells) — and removing negative engagements from the history sequence — significantly improved retrieval quality. All subsequent model versions use positive-only history.
- **4.6. A/B Results** (Ramanujam et al., 2025):
  - +0.8% revenue overall
  - +3.29% revenue for low-connection users
  - +0.2% Daily Unique Professional Interactors (overall)
  - +1.17% Daily Unique Professional Interactions (low-connection cohort)
  - +0.23% Daily Active Unique Users

III. Online Serving Funnel (Real-Time Inference)

LinkedIn executes this end-to-end, sub-second process for every feed request.

A. L0: Candidate Generation

- **1.1. Network Content (FishDB — Path A)**: FishDB retrieves posts from actors the member directly follows or is connected to (connections, followed creators, companies, subscribers). P99 latency: 40ms. Content window: last 30 days only.
- **1.2. Out-of-Network Content (Causal LLM — Path B)**: The Causal LLM dual encoder generates a member embedding and performs cosine similarity kNN search on the GPU-RAR cluster (72 H100 GPUs) against pre-computed item embeddings. Retrieves top candidates (the Causal LLM paper specifies "top 1,000 candidates to feed to subsequent layers of the ranking stack" in its Implementation Details; the abstract states 2,000 candidates from the total retrieval pool across all sources). Latency: sub-50ms.
- **1.3. Heuristic-Based Retrieval**: Fast, rule-based systems pulling in candidates based on timeliness signals and engagement velocity.
- **1.4. Aggregation**: The system collects, de-duplicates, and passes candidates from all sources to the ranking stage. Output is a longlist of candidate item IDs.

B. L2: Ranking — The Generative Recommender (GR)

Note: The primary feed ranking model (Feed SR / Generative Recommender) runs on its own custom PyTorch inference server with the SRMIS kernel, not SGLang. See Section B.5 above for full serving infrastructure details.

- **2.1. Member History Assembly**: The inference driver fetches the member's interaction history (1,000+ historical posts and actions as interleaved pairs). The offline pipeline generates member history features and stores them as compact Arrow columnar buffers in key-value stores for efficient access. LinkedIn optimized member history parsing from 450ms to 2ms through vectorized NumPy strided array processing.
- **2.2. Profile Embedding Retrieval**: The inference driver retrieves the Qwen3 0.6B-generated member profile embedding (refreshed daily) from the embedding store for late fusion.
- **2.3. GR Transformer Processing**: The decoder-only transformer processes the interleaved post+action sequence with causal attention. The transformer blocks output contextual representations for each position in the sequence.
- **2.4. SRMIS Multi-Item Scoring**: The SRMIS kernel appends all candidates to the end of the interaction sequence and scores them simultaneously via the custom Flash Attention kernel. This delivers an 80x speedup on the transformer forward pass for typical workloads (500 candidates, history length 1,000).
- **2.5. Late Fusion and MMoE Head**: The pipeline concatenates context features (count features, affinity features, device type) and the Qwen3 0.6B profile embedding with transformer outputs. The MMoE prediction head applies separate gating for passive (click, skip, long-dwell) vs. active (like, comment, share) prediction tasks.
- **2.6. Output**: A list of candidate items, each with relevance scores across multiple predicted action types.

C. Online Serving Engine — Other Components

- **C.1. SGLang (AI Job Search and AI People Search ranking)**:
    - Products: AI Job Search cross-encoder SLM ranking and AI People Search ranking — confirmed as the scope in the February 2026 Scaling LLM blog
    - Multi-Item Scoring (MIS): Concatenates multiple candidates with the member prompt, scoring them in a single pass. Latency reduction: **69%** vs. single-item scoring with prefix caching (Shimizu et al., 2025)
    - FlashAttention 3: LinkedIn contributed FA3 as the default attention backend in SGLang
    - FP8 Quantization: The per-token FP8 kernel (sgl_per_token) delivers **9% additional latency improvement** (Shimizu et al., 2025)
    - Knock-Knock Latency Hiding: Preemptively runs the LLM on member context while retrieval executes, reducing overall latency by **~38%** (520ms → 200ms) by overlapping prefill with item retrieval (Shimizu et al., 2025)
    - MixLM for job search: 22,000 items/second/GPU throughput via SGLang (see Section E)
- **C.2. vLLM (GenAI Applications)**:

    - Powers 50+ GenAI use cases including LinkedIn Hiring Assistant (Zhai et al., 2025)
    - AI Job Search query understanding: vLLM powers the query understanding component that converts free-form search queries into structured interpretations and facet suggestions — this is the generative text production workload in AI Job Search, distinct from the SGLang-powered cross-encoder ranking layer
    - Core technology: PagedAttention for efficient GPU memory management and high-concurrency request batching
    - Optimization: Tensor Parallelism to distribute models across multiple GPUs

D. Finalization, Delivery & Feedback

- **3.1. Business Rule Filters**: A final post-processing step applying non-relevance-based rules:
    - Trust & Safety: Content moderation filters
    - Impression Discounting: Down-ranks or removes content the member has already seen
    - Frequency Capping: Rules to prevent author or topic saturation in the feed
- **3.2. Diversity Modeling (Setwise Ranking)**: An optional layer that re-ranks the top-N candidates by evaluating them as a collective "set" to optimize for session-level diversity and coherence.
- **3.3. Delivery**: The system sends the final, ordered list of item IDs to a Render Service, which formats the content for the specific client (Web, iOS, Android).
- **3.4. Feedback Loop**: The system logs all final impressions and subsequent user interactions, streaming them back via Kafka to both the real-time and offline data systems, closing the loop for the next cycle of model training and embedding updates.

IV. GPU-RAR (GPU Retrieval as Ranking)

LinkedIn's Causal LLM retrieval system bypasses traditional database lookups:

- **GPU Cluster**: 72 total H100 GPUs — 48 nearline processing (item and member embedding inference, including backfill for new experimental models) + 24 dedicated retrieval (indexing item embeddings and GPU-RAR kNN retrieval with attribute-based matching)(Ramanujam et al., 2025)
- **Architecture**: Item embeddings stored directly in GPU memory; the 24-GPU retrieval cluster handles the index and kNN search
- **Retrieval**: Top candidates retrieved via cosine similarity search on GPU
- **Latency**: Sub-50ms from a corpus of hundreds of millions of items

This architecture blurs the traditional boundary between retrieval infrastructure and ranking infrastructure — both happen on GPU compute, with semantic matching and ranking happening in the same hardware environment.

V. Embedding Specifications Across Systems

Different LinkedIn systems use different embedding configurations optimized for their specific tasks:

System	Model	Embedding Dimension	Pooling Method	Use Case
Feed Retrieval (Causal LLM)	LLaMA-3 3B fine-tuned	3,072 (full) / 512 (Matryoshka option)	Mean pooling	Member/content semantic matching for out-of-network retrieval
Feed Ranking profile feature (GR)	Qwen3 0.6B fine-tuned	Dense vector (dimension not published)	Aggregated profile embedding	Late-fused context for GR prediction head
Job Matching ranking (MixLM)	0.6B encoder/ranker	H (not stated in paper)	Not stated in source paper	Job recommendation cross-encoder ranking
SLM compressed deployment	Various (2.1B–8B range)	Task-dependent	Task-dependent	Latency-sensitive serving across surfaces

The MixLM paper (arXiv:2512.07846) does not state the explicit hidden dimension H. The 512-dimensional Matryoshka option for the Causal LLM is a validated experimental configuration with minimal recall loss, not a confirmed production deployment dimension.

Matryoshka Representation Learning: Flexible Deployment

LinkedIn uses Matryoshka Representation Learning (MRL) for the Causal LLM retrieval system (Ramanujam et al., 2025):

- **Full dimension**: 3,072-dimensional vectors for maximum accuracy

- **Reduced dimension**: 512-dimensional vectors with minimal recall loss (Recall@10: 0.4225 vs. 0.4242 full)
- **No retraining required**: The same model produces both by optimizing multiple sub-representations simultaneously during training

---

MixLM Compression (Job Search)

MixLM achieves extreme compression for efficient ranking (Li et al., 2025):

- **Input**: ~900 tokens of job/member text
- **Output**: 2 tokens (1 embedding token + 1 special token) per item
- **Compression factor**: ~450x
- **Throughput**: 22,000 items/second/GPU via SGLang

References

Behdin, K., Song, Q., Kaushal, A., Ma, X., et al. (2025). Scaling down, serving fast: Compressing and deploying efficient LLMs for recommendation systems. *arXiv*. arXiv:2502.14305v2.

Danchev, H. (2026, March 12). Engineering the next generation of LinkedIn's Feed. *LinkedIn Engineering Blog*.

Hertel, L., Srivastava, G., Naqvi, S. A., Kumar, S., Zhang, Y., Ocejo, B., et al. (2026, February 12). An industrial-scale sequential recommender for LinkedIn feed ranking. *arXiv*. arXiv:2602.12354v1.

Li, G., He, R., Jing, S., Behdin, K., Wang, Y., Ramachandran, S. R., et al. (2025, November 25). MixLM: High-throughput and effective LLM ranking via text-embedding mix-interaction. *arXiv*. arXiv:2512.07846v1.

Li, K., Raveendran, N., Gupta, P., Ki, E., Singh, S., & Krishnamurthy, V. (2025, November 17). FishDB: A generic retrieval engine for scaling LinkedIn's feed. *LinkedIn Engineering Blog*.

Ramachandran, S. R., Lan, Q., Nguyen, C., Sheng, J., & Zhu, C. (2026, February 20). Scaling LLM-based ranking systems with SGLang at LinkedIn. *LinkedIn Engineering Blog*.

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B., et al. (2025). Large scale retrieval for the LinkedIn feed using causal language models. *arXiv*. arXiv:2510.14223v1.

Shimizu, S., Lan, Q., Dharamsi, T., Ramachandran, S. R., De, A., Wang, Y., et al. (2025, December 9). Turbocharging LinkedIn's recommendation systems with SGLang. *LinkedIn Engineering Blog*.

Zhai, Y., et al. (2025, August 26). How we leveraged vLLM to power our GenAI applications at LinkedIn. *LinkedIn Engineering Blog*.

Zhang, H., et al. (2025). 360Brew: A decoder-only foundation model for personalized ranking and recommendation. *arXiv*. arXiv:2501.16450v4.

Methodology and Disclosures

About Us

We are Trust Insights, a management consulting firm that helps organizations transform data into meaningful business outcomes. We specialize in analytics, data science, machine learning, and artificial intelligence implementations that deliver practical, measurable results. Our services range from training and education to fully managed AI deployments. Contact us to discuss your data and insights needs.

- Learn more about us: https://www.trustinsights.ai
- Learn more about our AI services: https://www.trustinsights.ai/aiservices

How This Guide Was Researched

Source Categories

This guide synthesizes two categories of primary sources:

Original Sources (Pre-2025 LinkedIn Engineering Era)

These 13 publications cover LinkedIn's foundational systems — feed ranking history, multi-task learning, graph neural networks, dwell time, near real-time personalization, and the data infrastructure layers (Venice, FollowFeed/FishDB lineage) that continue to underpin the current platform. These sources provide essential context for understanding how LinkedIn's systems evolved into their current form. Where these sources describe approaches that LinkedIn has since superseded (such as DCNv2-based ranking), we note the transition explicitly.

Current Sources (2025–2026 LinkedIn Engineering Publications)

These 20 publications are the primary evidentiary basis for the guide's technical claims. They include peer-reviewed papers published in KDD, RecSys, CIKM, and AAAI proceedings; arXiv preprints; and LinkedIn Engineering Blog posts. This category encompasses two 2026 sources: arXiv:2602.12354v1 (Hertel, Srivastava et al., February 12, 2026) and the LinkedIn Engineering Blog post "Engineering the next generation of LinkedIn's Feed" (Danchev, March 12, 2026). These two sources are critical: they establish Feed SR / Generative Recommender as the actual production feed ranker, replacing the DCNv2-based model that preceded it.

Total primary sources consulted: 31 publications across both categories.

Source Hierarchy and Conflict Resolution

When sources conflict, we apply the following priority order:

- **Most recent LinkedIn official publication** (Engineering Blog posts and peer-reviewed papers dated 2026)
- **Peer-reviewed academic papers** (KDD, RecSys, CIKM proceedings)
- **arXiv preprints** from LinkedIn researchers
- **Older Engineering Blog posts** (for historical context and foundational system descriptions)

Where a more recent source explicitly supersedes an older claim, the guide reflects the current state and notes the change.

Production Status Claims: What Is and Is Not Confirmed

This guide distinguishes between systems that LinkedIn has confirmed in production and systems that remain at the research stage.

Confirmed production systems as of Q1 2026:

- **Causal LLM retrieval** (arXiv:2510.14223v1): LLaMA-3 3B dual encoder generating embeddings for suggested content retrieval. Multiple sources including the March 2026 Engineering Blog confirm this deployment.
- **FishDB** (Li et al., November 2025 blog): Rust-based retrieval engine for connection-based content. Confirmed as replacement for FollowFeed.
- **Feed SR / Generative Recommender (GR)** (arXiv:2602.12354v1; Danchev, March 2026): Sequential transformer ranking model. Two independent 2026 sources confirm this as "currently the primary member experience on LinkedIn's Feed."
- **SGLang for AI Job Search and AI People Search** (Ramachandran et al., February 2026): Confirmed for these two surfaces explicitly; the blog states this scope directly.
- **vLLM for 50+ GenAI applications** (Zhai et al., August 2025): Confirmed for Hiring Assistant and query understanding components.

Pre-production or uncertain production status:

- **360Brew** (arXiv:2501.16450v4, recalled late 2025): The 360Brew decoder-only foundation model describes itself as "pre-production" in its own paper. arXiv:2602.12354v1 (Section 5.1) documents that LinkedIn evaluated a text-prompt LLM ranker architecturally equivalent to 360Brew for feed ranking and found it did not achieve superior online performance. This guide does not claim 360Brew is the primary feed ranker. The guide describes 360Brew as LinkedIn's LLM foundation model research, actively deployed across other non-feed surfaces per its own disclosure ("8+ surfaces"), but not confirmed as the primary feed ranker.

Note on the arXiv recall: LinkedIn withdrew arXiv:2501.16450 in late 2025 due to licensing concerns — the submitter did not have rights to publish proprietary information. This recall confirms the paper described real LinkedIn systems, but it does not by itself establish which surface or production role those systems occupy. The guide treats the recall as confirming the paper's description of the 360Brew model architecture while remaining agnostic about its production scope beyond what other sources confirm.

- **User-prompt-driven feed recommendation**: arXiv:2510.14223v1 (October 2025) describes this as "prototyping." We have not confirmed production deployment as of this guide's writing.

Synthesis Methodology

We used Google's Gemini 2.5 Pro model and Anthropic's Claude to synthesize approximately 400,000 words of source material across all primary sources. We trace all technical claims to specific passages in official LinkedIn publications, peer-reviewed research, or verified news sources. We label claims involving future directions as stated goals from LinkedIn's own publications, not as confirmed production deployments.

The guide also incorporates an independent Trust Insights research study (Penn & Robbert, 2025) examining gender bias in LLaMA-3 embeddings using LinkedIn's published Causal LLM methodology. This study tested 406 paired posts and found systematic bias based on author name (Cohen's $d = -0.93$, $p < 0.0001$). This finding is relevant to practitioners relying on the retrieval system and is included in the guide's bias and fairness discussion.

Limitations and Uncertainty Disclosures

This guide is not endorsed by LinkedIn. It is an independent analysis of LinkedIn's public research publications. LinkedIn has not reviewed, approved, or validated the interpretations presented here.

LinkedIn's systems are continuously evolving. The technical specifications in this guide reflect the state of LinkedIn's systems as documented in published sources through March 2026. LinkedIn deploys updates continuously; some details may change between the guide's writing and your reading.

Architectural inferences are labeled as such. Where source documents do not specify a detail but the logic of the architecture implies it, we note this explicitly. We do not present inferences as confirmed specifications.

Source coverage is not complete. LinkedIn's engineering organization publishes selectively. Many production systems, features, and optimizations are never publicly

documented. This guide describes what LinkedIn has chosen to disclose. Undisclosed systems and decisions may differ from what the published record implies.

The 360Brew paper was recalled. LinkedIn withdrew the primary source for the 360Brew model description (arXiv:2501.16450v4) from arXiv in late 2025 due to licensing concerns. The technical information in the paper remains the most comprehensive public documentation of LinkedIn's foundation model approach, and the recall itself confirms the paper described real LinkedIn systems. However, readers should be aware that this source exists in an unusual legal context, and the guide treats it as historical documentation of a research system rather than confirmed production deployment for the feed.

A/B test results represent LinkedIn's conditions. LinkedIn researchers measured the performance metrics cited from LinkedIn's published A/B tests (e.g., +2.10% time spent for Feed SR, +0.8% revenue for Causal LLM) under LinkedIn's member population, content distribution, and infrastructure conditions as of those tests' execution dates. These figures describe system quality improvements; they do not constitute marketing claims or guarantees of outcomes for individual users or content creators.

Disclosures

Trust Insights is a commercially independent firm. We have no financial relationship with LinkedIn, Meta, Google, Anthropic, or any other company referenced in this guide, except as ordinary customers of their services. This guide was not commissioned, sponsored, or funded by LinkedIn.

We use AI tools in our work. The synthesis of this guide used Google's Gemini 2.5 Pro and Anthropic's Claude. We used these tools because the source volume (400,000+ words of technical documentation) exceeds what a small team can synthesize manually at the accuracy level this subject demands. All AI-assisted synthesis was reviewed against primary sources by human editors.

This guide may be updated. The LinkedIn algorithm is a living system. We intend to update this guide as significant new publications become available. The edition designation (Q1 2026) reflects the primary sources current as of March 2026.

Consolidated References

This section contains all references cited throughout The Unofficial LinkedIn Algorithm Guide, Q1 2026 Edition, organized by category and alphabetically within each category in APA format.

Academic Papers & Technical Research

Behdin, K., Fatahibaarzi, A., Song, Q., Dai, Y., Gupta, A., Wang, Z., et al. (2025). Scaling down, serving fast: Compressing and deploying efficient LLMs for recommendation systems. *arXiv*. https://arxiv.org/abs/2502.14305

Behdin, K., Song, Q., Vasudevan, S., et al. (2025). Scaling up large language models serving systems for semantic job search. *arXiv*. https://arxiv.org/abs/2510.22101

Borisyuk, F., Hertel, L., Parameswaran, G., Srivastava, G., Ramanujam, S., Ocejo, B., Du, P., Akterskii, A., Daftary, N., Tang, S., Sun, D., Xiao, C., Nathani, D., Kothari, M., Dai, Y., & Gupta, A. (2025). From features to transformers: Redefining ranking for scalable impact. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. https://arxiv.org/abs/2502.03417

Borisyuk, F., He, S., Ouyang, Y., Ramezani, M., Du, P., Hou, X., Jiang, C., Pasumarthy, N., Bannur, P., Tiwana, B., Liu, P., Dangi, S., Sun, D., Pei, Z., Shi, X., Zhu, S., Shen, Q., Lee, K.-H., Stein, D., Li, B., Wei, H., Ghoting, A., & Ghosh, S. (2024). LiGNN: Graph neural networks at LinkedIn. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery. https://doi.org/10.1145/3637528.3671566

Borisyuk, F., Song, Q., Zhou, M., et al. (2024). LiNR: Model based neural retrieval on GPUs at LinkedIn. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '24)*. https://arxiv.org/abs/2407.13218

Borisyuk, F., Zhou, M., Song, Q., Zhu, S., Tiwana, B., Parameswaran, G., Dangi, S., Hertel, L., Xiao, Q. C., Hou, X., Ouyang, Y., Gupta, A., Singh, S., Liu, D., Cheng, H., Le, L., Hung, J., Keerthi, S., Wang, R., Zhang, F., Kothari, M., Zhu, C., Sun, D., Dai, Y., Luan, X., Zhu, S., Wang, Z., Daftary, N., Shen, Q., Jiang, C., Wei, H., Varshney, M., Ghoting, A., & Ghosh, S. (2024). LiRank: Industrial large scale ranking models at LinkedIn. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery. https://doi.org/10.1145/3637528.3671561

Firooz, H., Sanjabi, M., Jiang, W., & Zhai, X. (2025). Lost-in-distance: Impact of contextual proximity on LLM performance in graph tasks. *arXiv*. https://arxiv.org/abs/2410.01985

He, S., Choi, J., Li, T., Ding, Z., Du, P., Bannur, P., Liang, F., Borisyuk, F., Jaikumar, P., Xue, X., & Gupta, V. (2025). Large scalable cross-domain graph neural networks for personalized notification at LinkedIn. *arXiv*. https://arxiv.org/abs/2506.12700

Hertel, L., Srivastava, G., Naqvi, S. A., Kumar, S., Zhang, Y., Ocejo, B., Zelditch, B., Englhardt, A., Cheng, H., Hu, A., Alonso, A., Li, D., Dangi, S., Zhu, C., Zhou, M., Li, W., Huang, T., Borisyuk, F., Parameswaran, G., Tiwana, B., Sankar, S., Lan, Q., Choi, J., & Ghosh, S. (2026, February 12). An industrial-scale sequential recommender for LinkedIn feed ranking. *arXiv*. https://arxiv.org/abs/2602.12354

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of experts. *arXiv*. https://arxiv.org/abs/2401.04088

Juan, Y., Shen, J., Zhang, S., et al. (2025). Scaling retrieval for web-scale recommenders. In *Proceedings of the 19th ACM Conference on Recommender Systems (RecSys '25)*. Association for Computing Machinery. https://doi.org/10.1145/3705328.3748116

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*. https://arxiv.org/abs/2309.06180

LinkedIn Engineering. (2025). MixLM: High-throughput and effective LLM ranking via text-embedding mix-interaction. *arXiv*. https://arxiv.org/abs/2512.07846

Liu, P., Shen, J., Shen, Q., et al. (2025). Powering job search at scale: LLM-enhanced query understanding in job matching systems. *arXiv*. https://arxiv.org/abs/2509.09690

Penn, C. S., & Robbert, K. (2025). Gender bias in LLaMA-3 embeddings: Implications for LinkedIn-style retrieval systems [Research report]. Trust Insights. https://doi.org/10.5281/zenodo.17982122

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B., et al. (2025). Large scale retrieval for the LinkedIn feed using causal language models. *arXiv*. https://arxiv.org/abs/2510.14223

Sanjabi, M., Firooz, H., & 360Brew Team. (2025). 360Brew: A decoder-only foundation model for personalized ranking and recommendation. *arXiv* [Recalled from arXiv in late 2025 due to licensing concerns; technical content remains the most comprehensive public documentation of LinkedIn's 360Brew foundation model approach]. https://arxiv.org/abs/2501.16450

LinkedIn Engineering Blog Posts

Ackerman, I., & Kataria, S. (2021, August 19). Homepage feed multi-task learning using TensorFlow. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2021/homepage-feed-multi-task-learning-using-tensorflow

Borisyuk, F., Xu, J., Ma, X., Vasudevan, S., Wu, M., Zheng, R., Le, B., Zhang, S., Metkar, S., Gupta, R., Shen, Q. K., Hooshmand, A., Racca, D. N., Katarya, V., Behdin, K., Lapchuk, I., Lu, X., Zhang, L., Mohanasundaram, G., Bottaro, J. P., … Ramachandran, S. R. (2026, January 21). Reimagining LinkedIn's search tech stack. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/search/reimagining-linkedins-search-tech-stack

Dangi, S., Jia, J., Somaiya, M., & Xuan, Y. (2020, October 29). Understanding dwell time to improve LinkedIn feed ranking. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2020/understanding-feed-dwell-time

Danchev, H. (2026, March 12). Engineering the next generation of LinkedIn's feed. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/feed/engineering-the-next-generation-of-linkedins-feed

Ghike, S., & Gupta, S. (2016, March 3). FollowFeed: LinkedIn's feed made faster and smarter. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2016/03/followfeed-linkedin-s-feed-made-faster-and-smarter

Gupta, R., Ovsankin, S., Li, Q., Lee, S., Le, B., & Khanal, S. (2022, April 26). Near real-time features for near real-time personalization. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2022/near-real-time-features-for-near-real-time-personalization

GV, F. (2022, September 28). Open sourcing Venice: LinkedIn's derived data platform. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2022/open-sourcing-venice-linkedin-s-derived-data-platform

Jurka, T., Ghosh, S., & Davies, P. (2018, March 15). A look behind the AI that powers LinkedIn's feed: Sifting through billions of conversations to create personalized news feeds for hundreds of millions of members. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2018/03/a-look-behind-the-ai-that-powerslinkedins-feed-sifting-through

Li, K., Raveendran, N., Gupta, P., Ki, E., Singh, S., & Krishnamurthy, V. (2025, November 17). FishDB: A generic retrieval engine for scaling LinkedIn's feed. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2025/fishdb-generic-retrieval-engine

Lyu, L., Zhang, C., Shang, Y., Jha, S., Jain, H., Ahmad, U., & the OpenConnect Team. (2024). OpenConnect: LinkedIn's next-generation AI pipeline ecosystem. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2024/openconnect

Mohamed, A., & Li, Z. (2019, June 27). Community-focused feed optimization. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2019/06/community-focused-feed-optimization

Ouyang, Y., Gupta, V., Basu, K., Diciccio, C., Gavin, B., & Guo, L. (2020, August 27). Using Bayesian optimization for balancing metrics in recommendation systems. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/recommendations/using-bayesian-optimization-for-balancing-metrics-in-recommendat

Ramachandran, S. R., Lan, Q., Nguyen, C., Sheng, J., & Zhu, C. (2026, February 20). Scaling LLM-based ranking systems with SGLang at LinkedIn. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/ai/scaling-llm-based-ranking-systems-with-sglang-at-linkedin

Shimizu, S., Lan, Q., Dharamsi, T., et al. (2025, December 9). Turbocharging LinkedIn's recommendation systems with SGLang. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/ai/turbocharging-linkedins-recommendation-systems-with-sglang

Zhai, Y., Kumar, S., Ramachandran, S. R., Zhu, C., Nguyen, C., Toddywala, F., Yao, C., Johnson, C., & Lan, Q. (2025, August 26). How we leveraged vLLM to power our GenAI applications at LinkedIn. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/ai/how-we-leveraged-vllm-to-power-our-genai-applications

Zhang, F., Kothari, M., & Tiwana, B. (2024, August 7). Leveraging dwell time to improve member experiences on the LinkedIn feed. *LinkedIn Engineering Blog*. https://www.linkedin.com/blog/engineering/feed/leveraging-dwell-time-to-improve-member-experiences-on-the-linkedin-feed

Zhu, J. (S.), Ghoting, A., Tiwana, B., & Varshney, M. (2023, May 2). Enhancing homepage feed relevance by harnessing the power of large corpus sparse ID embeddings. *LinkedIn Engineering Blog*. https://engineering.linkedin.com/blog/2023/enhancing-homepage-feed-relevance-by-harnessing-the-power-of-lar