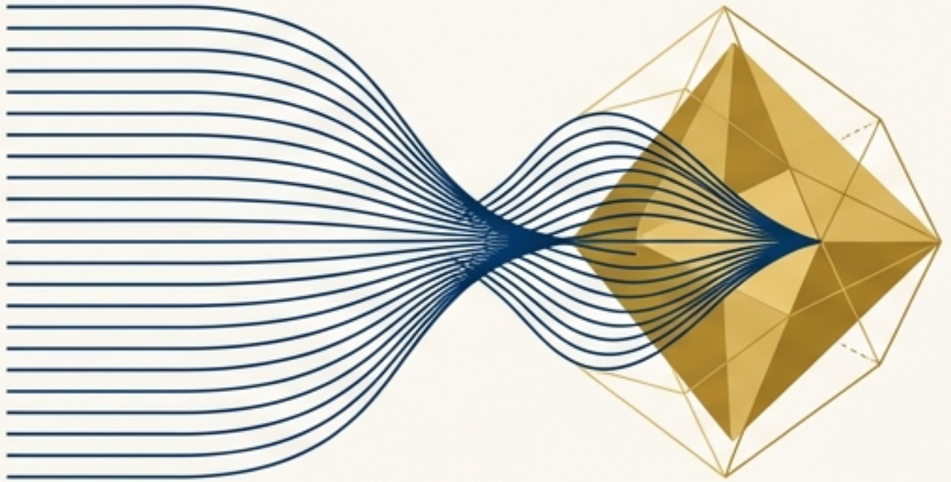
LinkedIn AI

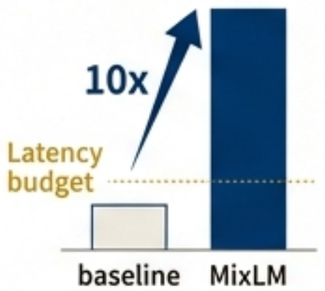
MixLM: テキストと埋め込み表現の統合による LLMランキングスループット10倍向上

大規模な検索・推薦システムにおいて、LLMの持つ高度な意味理解能力と、
本番環境で求められる高い効率性を両立させるための新しいフレームワーク



MixLMがもたらしたブレークスルー

10倍



スループット向上

同一のレイテンシバジェット（P99 < 500ms）内で、既存の要約テキストベースライン（2,200 items /sec）に対し、MixLMは22,000 items/secを達成。

+0.47%

DAU（デイリーアクティブユーザー）増加

効率化によりLLMランキングの全面展開が可能となり、オンラインA/BテストでLinkedInの主要ビジネス指標に有意な向上をもたらした。

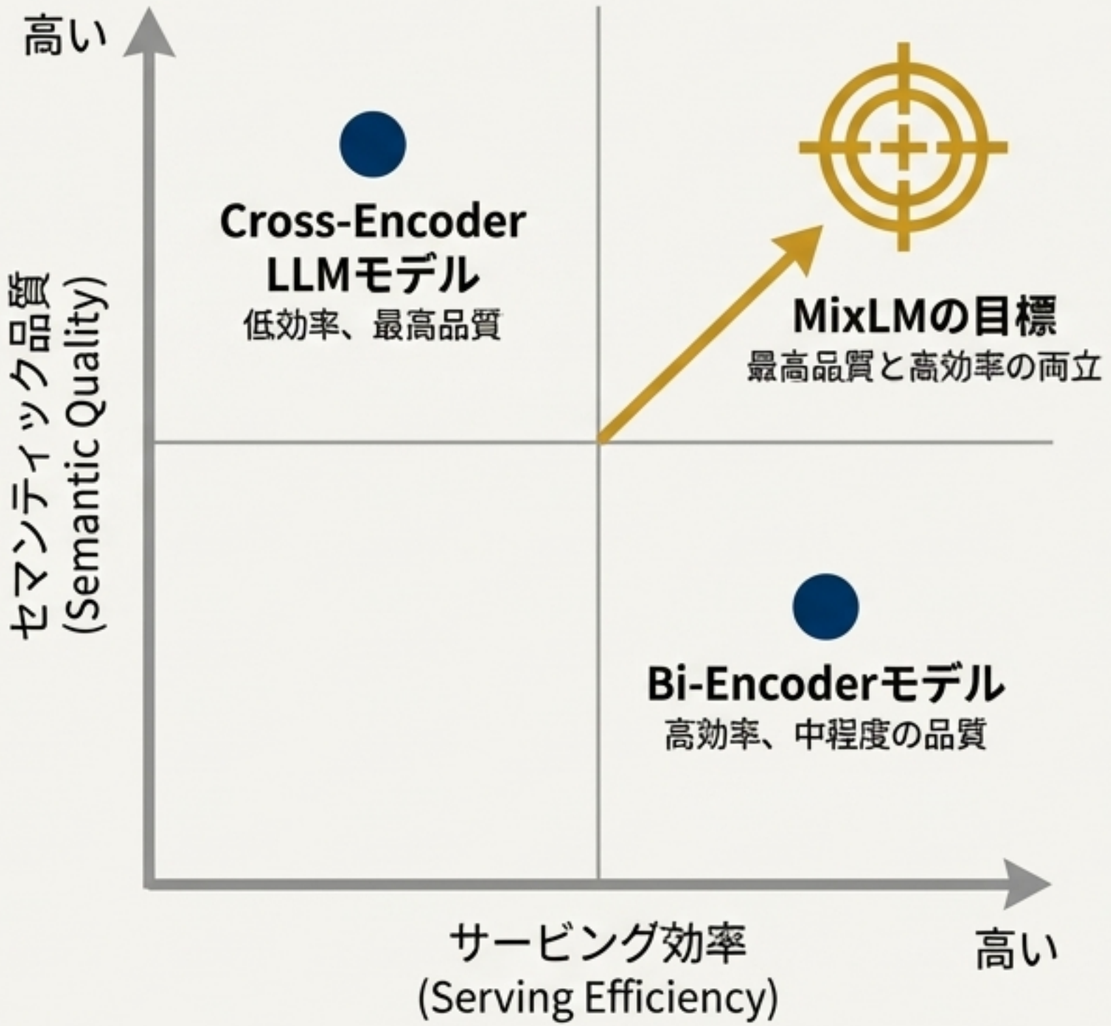
98%

関連性品質の維持

フルテキストモデルのNDCG@10スコア（0.9432）に対し、MixLMはNDCG@10スコア0.9239を達成。計算コストを劇的に削減しつつ、ランキング品質をほぼ維持。

MixLMは、品質を犠牲にすることなく、LLMランキングを経済的に実行可能にし、 プロダクション規模での展開を実現します。

LLMランキングにおけるジレンマ：品質と効率のトレードオフ



Cross-Encoder型LLMは、クエリとアイテムの文脈を深く理解し、卓越したランキング精度を達成します。しかし、その計算コストは非常に高く、特に長いテキスト入力を処理する際には、レイテンシとスループットの要件を満たすことが困難です。これが、大規模な本番システムへの導入を妨げる最大の障壁となっていました。

ボトルネックはコンテキスト長：アイテム説明文が計算コストを増大させる

クエリ (Query)

アイテム説明文 (Item Description)

中央値900トークン

LinkedInの求人情報（アイテム）のテキスト長。最大で2100トークンに達することもあります。

二次関数的な計算量増加

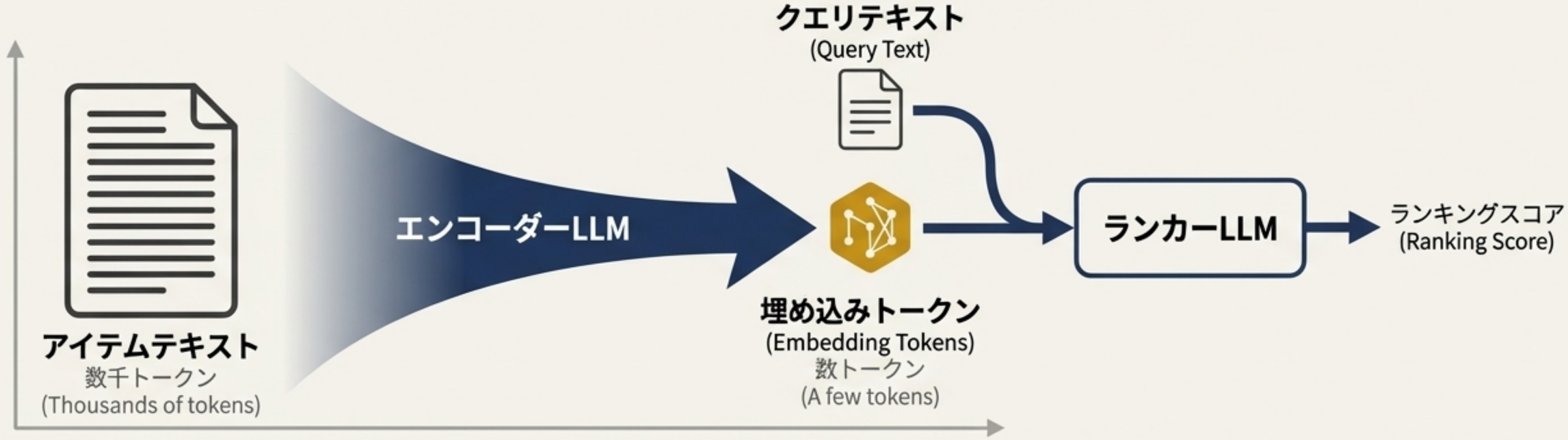
Transformerの自己注意機構（Self-Attention）の計算量は（Self-Attention）の計算量は、入力コンテキスト長に対して二次関数的に増加します。

毎秒315万アイテム

LinkedInの求人検索で処理が必要なアイテム数。このトラフィックを捌くには、ランキング処理の効率が極めて重要です。

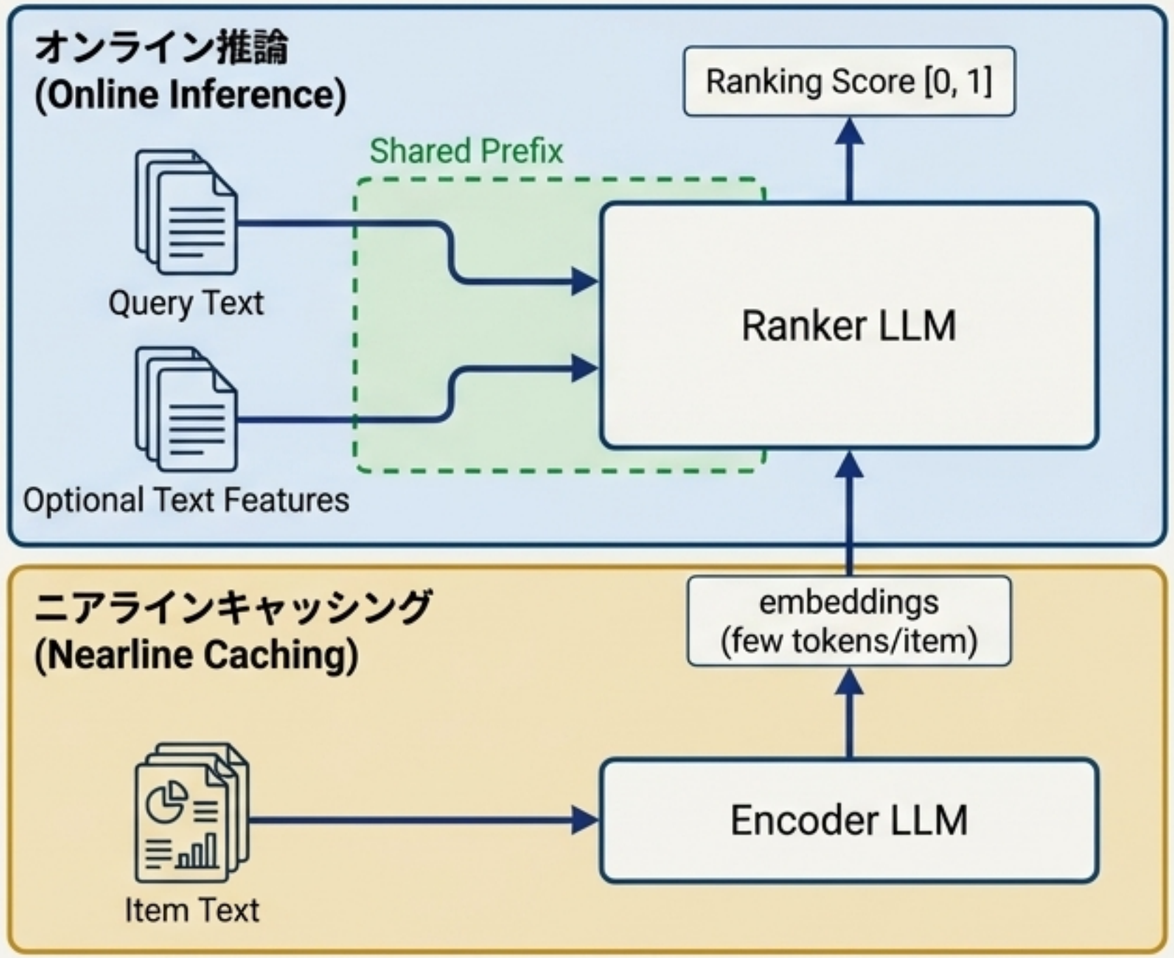
長大なプロンプトは、許容不可能なレイテンシとコストを引き起こし、LLMランキングの全面的な展開を阻害していました。

MixLMのコンセプト：長文テキストを少数の「埋め込みトークン」に圧縮する



- MixLMは、アイテムの長大なテキスト情報を、事前に計算された少数の埋め込みトークンに圧縮します。
- オンライン推論時には、このコンパクトな埋め込みとユーザークエリのテキストを「混合（Mix）」してランカーLLMに入力します。
- これにより、Cross-Encoderの持つ豊かなインタラクションを維持しつつ、計算量を劇的に削減します。

MixLMアーキテクチャ：オフラインのエンコーダーとオンラインのランカー



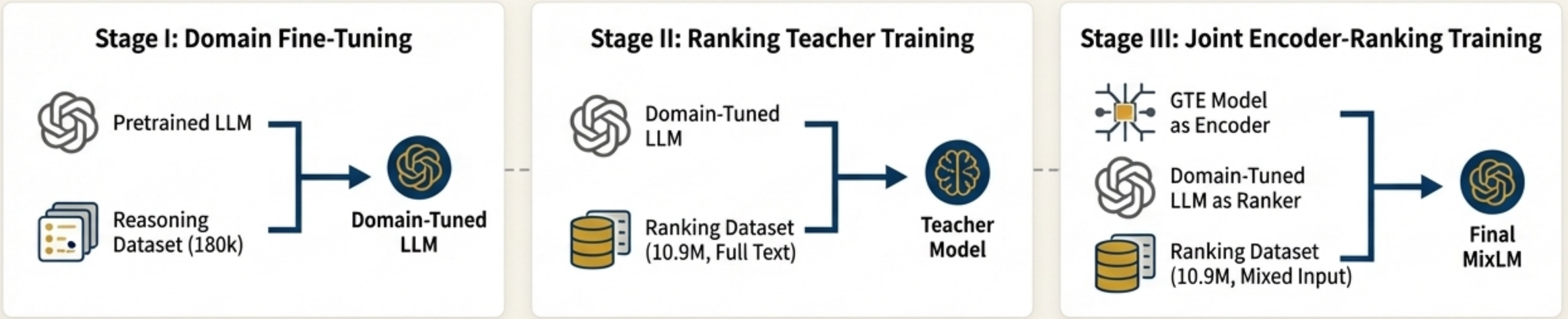
エンコーダーLLM (Encoder LLM) - オフライン/ニアライン

カタログ内の全アイテムのテキスト情報を処理し、コンパクトな埋め込み表現を生成。生成された埋め込みはニアラインキャッシュに保存される。

ランカーLLM (Ranker LLM) - オンライン

リアルタイムでユーザークエリと、キャッシュから取得したアイテム埋め込みを組み合わせた混合入力を処理し、関連性スコアを予測する。

3段階のトレーニング戦略でエンコーダーとランカーを連携させる



目的：ドメイン知識と推論能力の獲得。

目的：高品質なフルテキストベースのランキング教師モデルの作成。

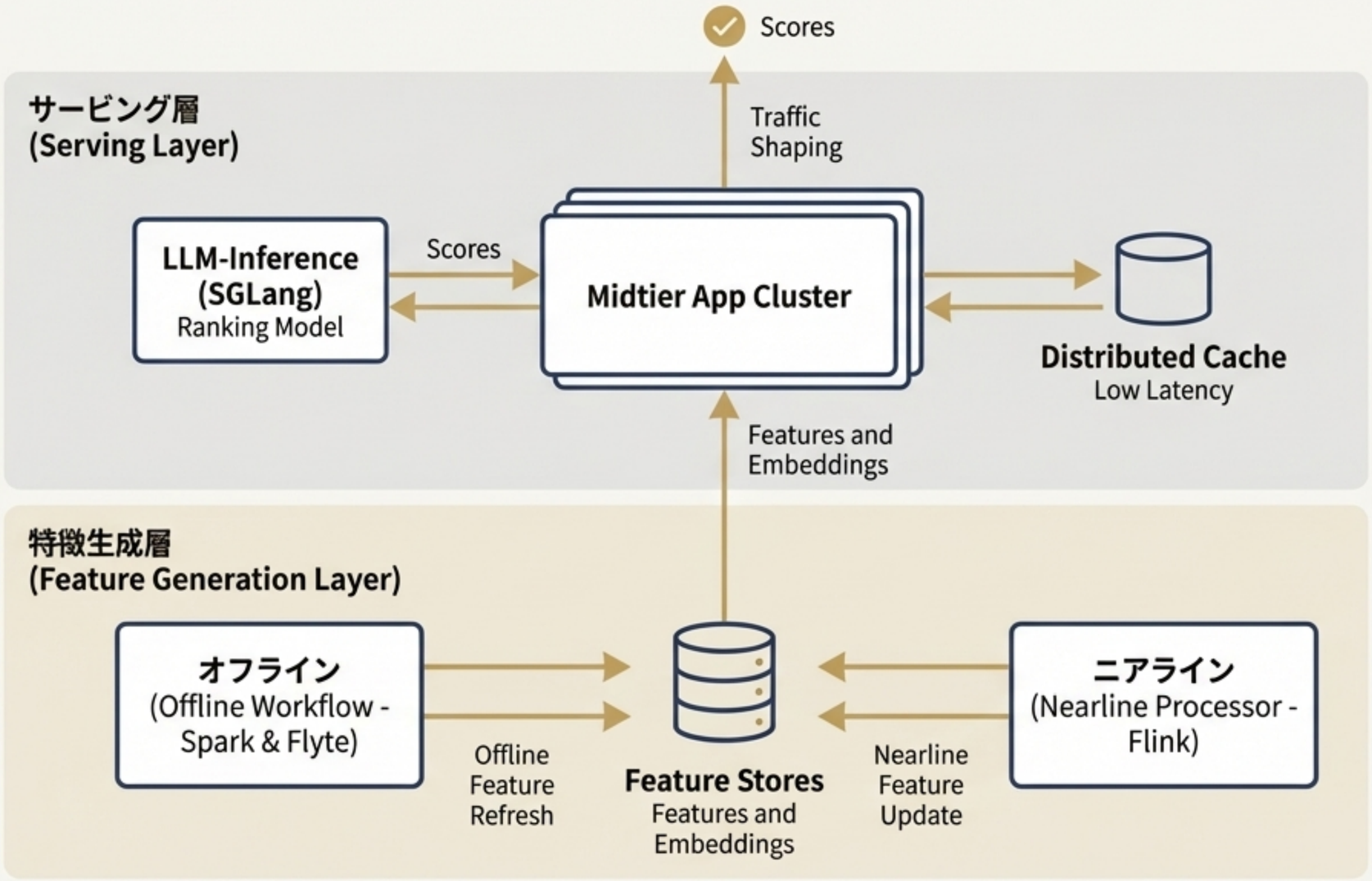
目的：エンコーダーとランカーの共同最適化。

SFT Loss
MixLMの予測を正解ラベルに近づける。

Distillation Loss
フルテキストの教師モデルの知識をMixLMに蒸留し、学習を安定させる。

Self-Alignment Loss
エンコーダーの出力埋め込みとランカーの内部表現空間を整列させ、過学習を抑制する。

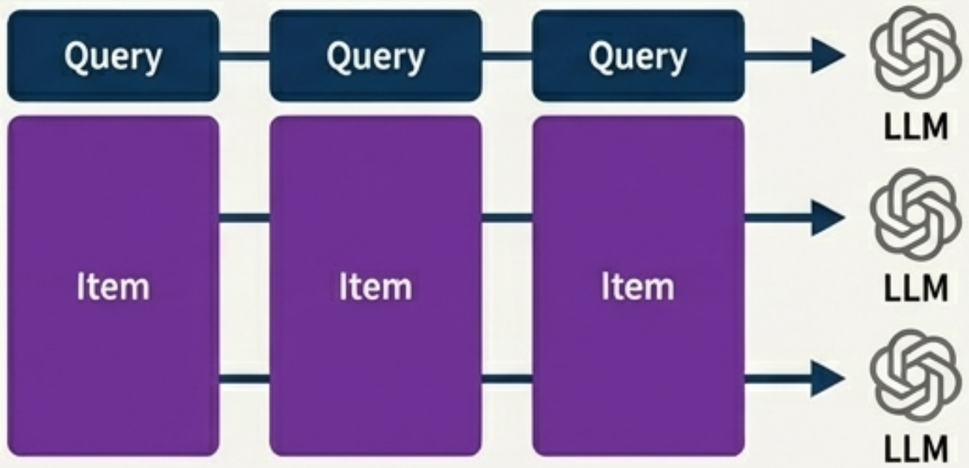
本番環境に対応したサービングシステム



- オフラインとニアラインを組み合わせたハイブリッドシステムにより、網羅性と最新性を両立。
- 事前計算された特徴量を効率的に再利用することで、リアルタイム推論の応答性を確保。
- SGLangなどの最適化された推論エンジンを活用。

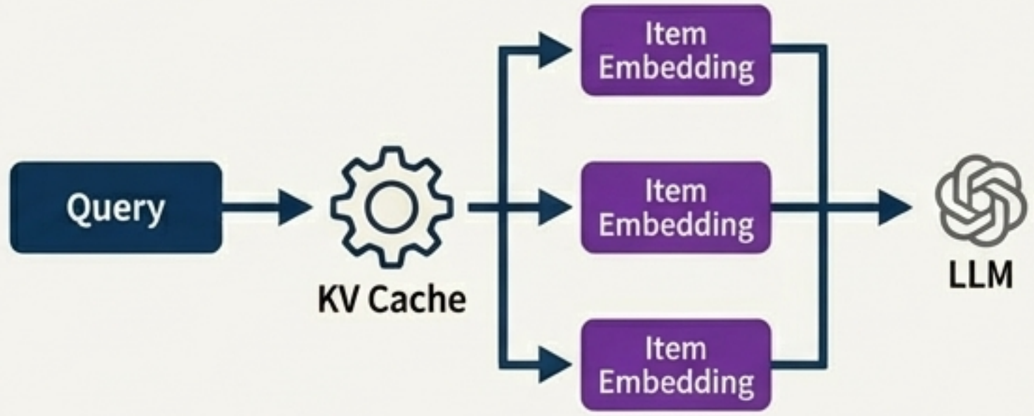
推論の最適化：共有プレフィックスの計算が10倍高速化の鍵

ナイーブなアプローチ (Naïve Approach)



$$\text{Cost} \propto N_{\text{items}} * (T_{\text{query}} + T_{\text{item}})^2$$

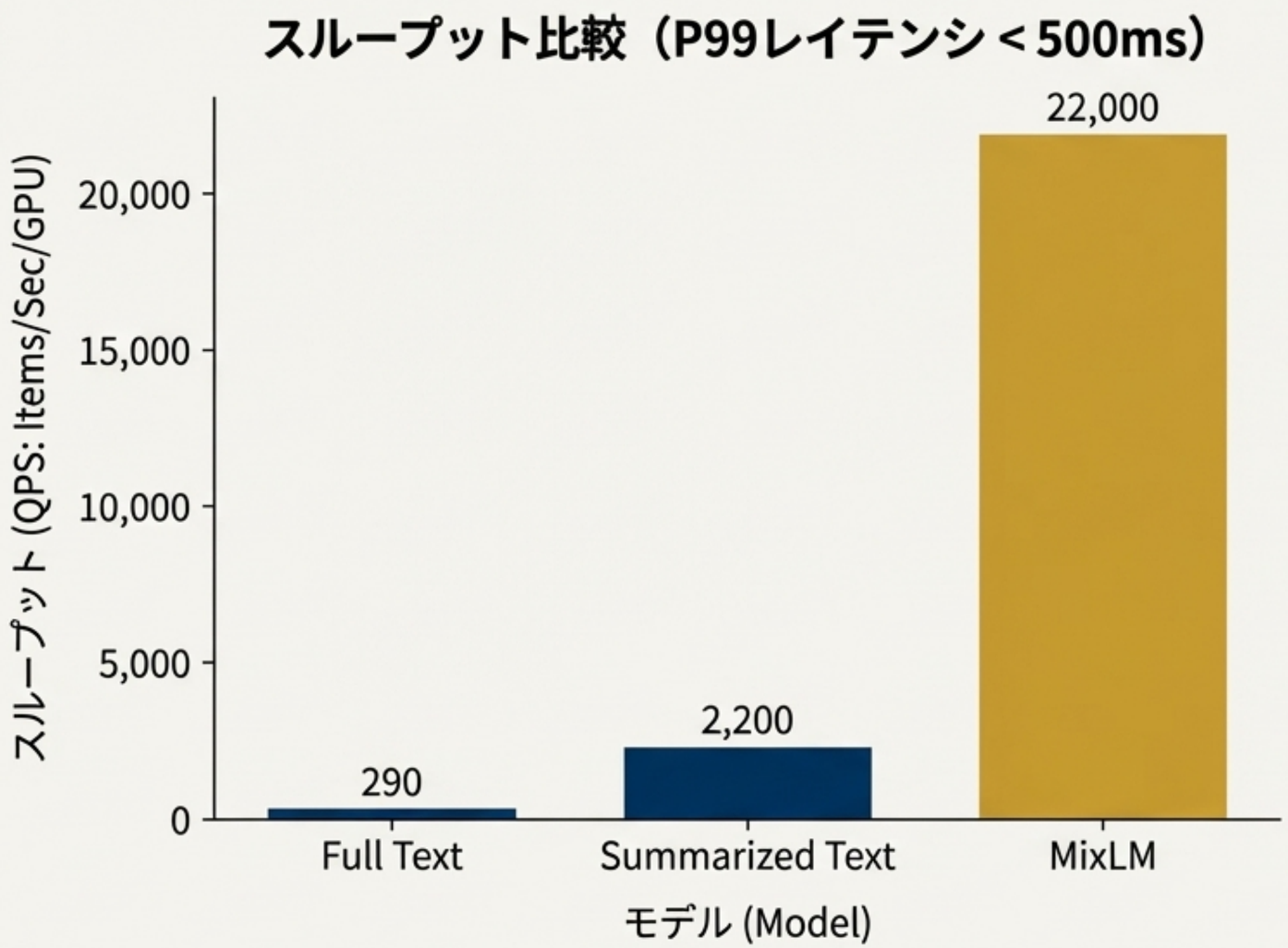
償却プリフィル (Amortized Prefill with MixLM)



$$\text{Cost} \propto T_{\text{query}}^2 + N_{\text{items}} * (2 * T_{\text{item\_emb}} * T_{\text{query}} + T_{\text{item\_emb}}{}^2)$$

検索リクエストでは、通常1つのクエリに対して数百の候補アイテムをランキングします。
MixLMではアイテム表現がわずか数トークンに圧縮されるため、計算負荷の大部分は全アイテムで共通の「クエリ部分（共有プレフィックス）」に移ります。
償却プリフィル（Amortized Prefill）では、この共有プレフィックスのKVキャッシュをバッチ内で再利用します。
これにより、冗長な計算が劇的に削減され、スループットが大幅に向上します。

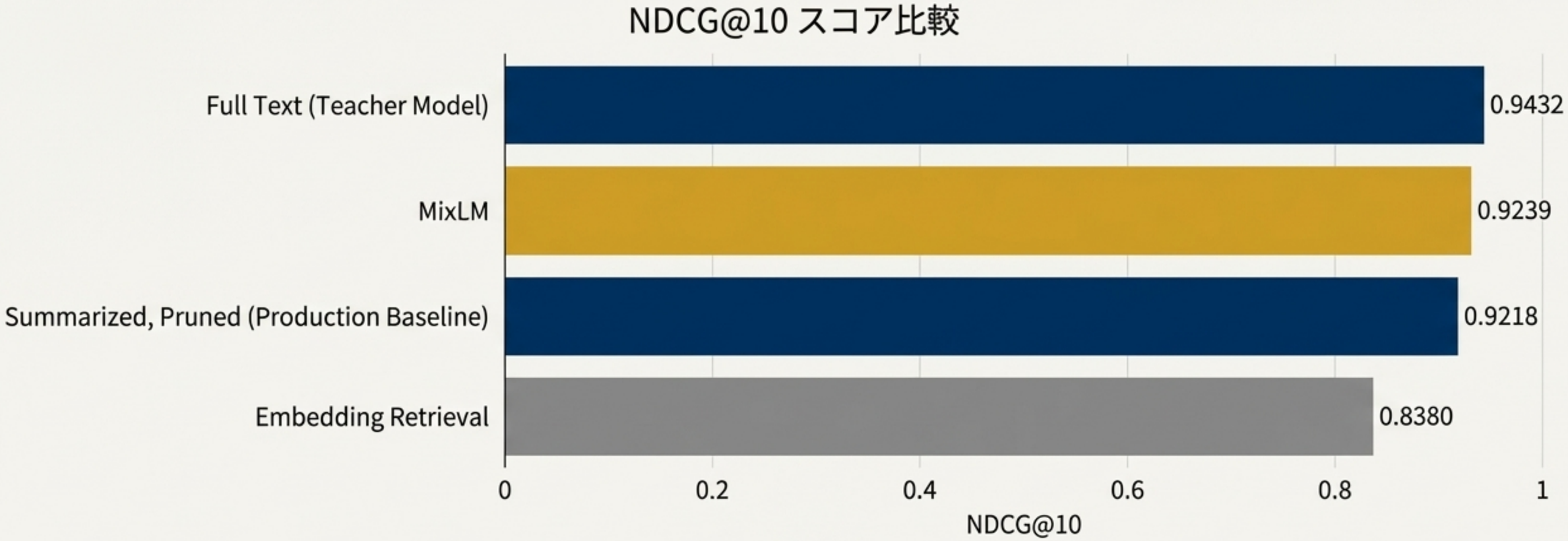
パフォーマンス対決：スループットとレイテンシ



MixLMは、要約テキストベースラインの10.0倍、フルテキストベースラインの75.9倍のスループットを達成。

アイテム説明文の圧縮と推論最適化の組み合わせにより、計算効率が飛躍的に向上しました。

品質の検証：効率向上のための関連性低下は最小限



MixLMのNDCG@10スコア（0.9239）は、既存の本番ベースライン（0.9218）と同等です。
また、計算コストが遥かに高いフルテキストの教師モデル（0.9432）との差もわずか1.8ポイントに留まっています。
MixLMは、品質と効率のトレードオフにおいて「スイートスポット」を達成しました。

最終的な証明：オンラインA/Bテストが示したビジネスインパクト

+0.47%

Daily Active Users (DAU) 増加

実験グループ（Experiment Group)：セマンティック求人検索（MixLM搭載）vs. 従来の求人検索

MixLM による10倍のスループット向上が、LLMベースのセマンティック検索を初めて全トラフィックに展開することを可能にしました。

この技術的な成功が、LinkedInの最重要ビジネス指標であるDAUの有意な増加に直接結びつきました。

MixLMは、優れたエンジニアリングが直接的なビジネス価値を生み出すことを証明しました。

何が効果的だったか：Ablation Studyから得られた主要な知見

データとトークン数のスケーリング (Scaling Data & Tokens)

訓練データ量とアイテムあたりの埋め込みトークン数を増やすと、NDCG@10は一貫して向上。スケールアップが品質向上に直結することを確認。（ただし本番ではレイテンシ制約から1トークンを採用）

ドメイン特化ファインチューニングの重要性 (Importance of Domain Fine-Tuning)

汎用的な事前学習済みモデルではなく、ドメイン固有のデータでファインチューニングしたモデルをランカーのベースにすることで、性能が大幅に向上（NDCG@10で+0.0185）。

蒸留損失の貢献 (Contribution of Distillation Loss)

補助的な損失関数の中で、教師モデルからの蒸留損失が最も性能向上に貢献（NDCG@10で+0.009）。SFT Lossのみの場合と比較して、学習の安定化と高品質化に不可欠。

カリキュラム学習の有効性 (Effectiveness of Curriculum Learning)

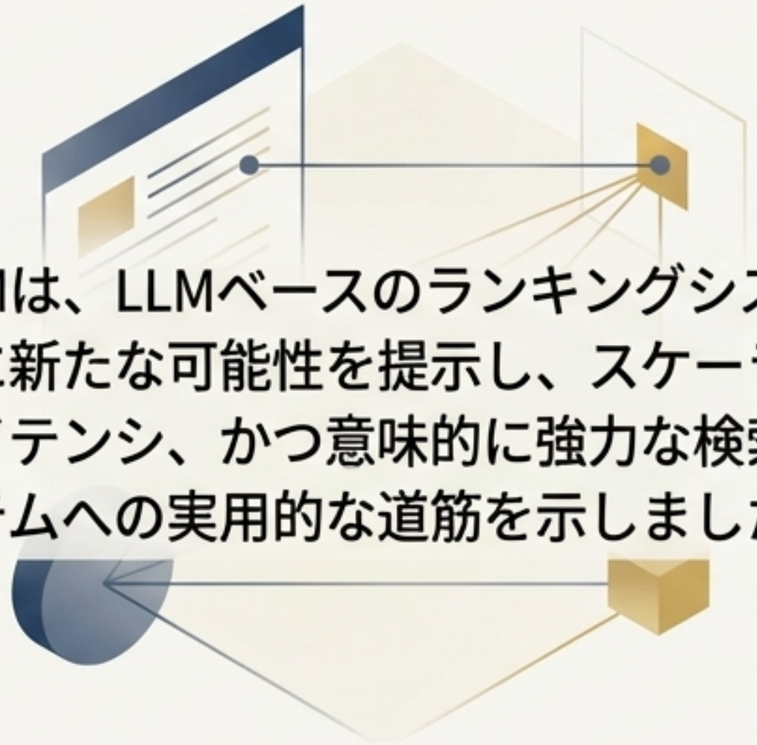
最初にアライメントを重視し、次に関連性予測タスクに焦点を当てる2段階のカリキュラム学習が、最も効果的であった（NDCG@10で+0.0020）。

結論：MixLMは、高品質なLLMランキングを実用的かつ経済的に実現する

サマリー（Summary）

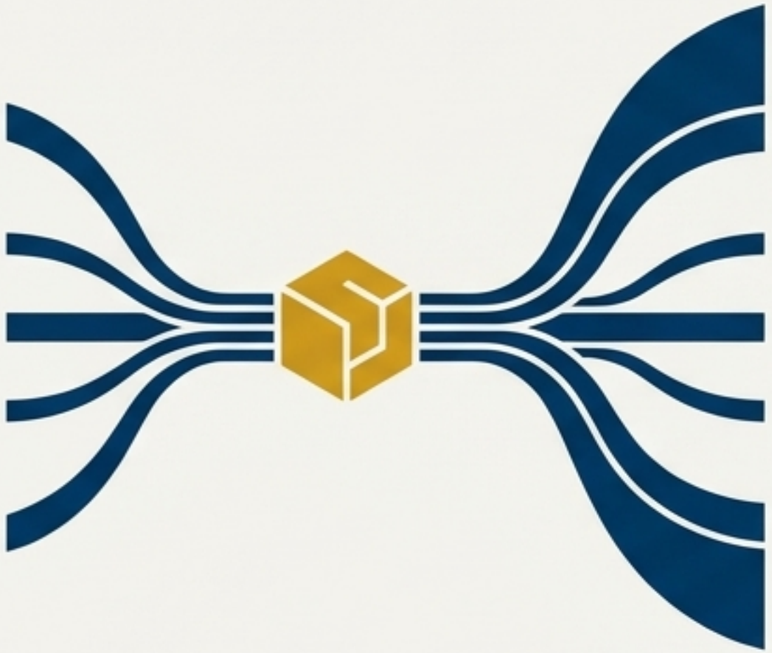
- MixLMは、アイテムのテキスト表現をコンパクトな埋め込みに置き換える「テキストと埋め込みの混合インタラクション」を導入しました。
- これにより、Cross-Encoder LLMの持つ高いセマンティック理解能力を維持しながら、本番環境で求められる厳しい効率要件をクリアしました。
- 高度なトレーニング戦略とサービングシステムの最適化により、スループットを10倍に向上させ、LinkedInの主要なビジネス指標を押し上げることに成功しました。

インパクト（Impact）



MixLMは、LLMベースのランキングシステムの設計に新たな可能性を提示し、スケーラブルで低レイテンシ、かつ意味的に強力な検索・推薦システムへの実用的な道筋を示しました。

次のステップへ



メンバーヒストリーのモデリング (Modeling Member History)

共同学習させた埋め込みを、ユーザーの行動履歴やプロファイルのモデリングに応用する。

検索タスクへの応用 (Exploring for Retrieval Tasks)

MixLMのコンセプトを、現在のランキングだけでなく、より広範な検索（リトリーバル）タスクにも拡張する可能性を探る。

MixLMで培ったテキストと埋め込みの融合というパラダイムは、
次世代のパーソナライズされたAI体験を構築するための基盤となります。