

Multi-Stage Proxy Retrieval for Illiquid Bond Pricing

Arturo Oncevay*
JPMorgan AI Research

Joy Sain*
JPMorgan AI Research

Chen (Nichole) Ling
JPMorgan PricingDirect

Simerjot Kaur
JPMorgan AI Research

Charese Smiley
JPMorgan AI Research

Xiaomo Liu
JPMorgan AI Research

first.last@jpmorgan.com

Abstract

Identifying proxy issuers and bonds is a critical but largely manual step in illiquid corporate bond pricing, requiring analysts to navigate sparse public information and inconsistent industry signals. We present a modular pipeline for proxy retrieval that integrates issuer profiling with hierarchical industry classification and multi-stage retrieval over issuers. We evaluate alternative design choices across hierarchical classification and reranking stages to understand how structured representations and industry-aware signals influence performance. Experiments on a large internal dataset show that domain-adapted embeddings and explicit industry hierarchy signals are primary drivers of quality across stages. The evaluated system produces issuer-level proxy shortlists rather than final bond prices. Pilot deployment experience highlights the value of modular, cost-aware designs for robust proxy retrieval in enterprise settings.

1 Introduction

Accurate pricing of illiquid corporate bonds remains a persistent operational challenge in financial markets. In practice, analysts often rely on identifying comparable issuers with liquid bonds to serve as pricing proxies, a process that involves gathering fragmented public information, profiling issuers, and evaluating large candidate pools. These workflows are largely manual and depend heavily on analyst expertise, leading to variability in outcomes and significant time investment as coverage expands to thousands of issuers. The growing demand for scalable and transparent proxy identification has motivated the development of automated retrieval pipelines that augment, rather than replace, expert decision-making.

Prior research on bond pricing has primarily focused on direct price estimation or quantitative

modeling (Gu et al., 2020; Huang et al., 2023; Dolphin et al., 2024), while comparatively little attention has been given to the upstream task of identifying comparable issuers and instruments. Recent NLP and hybrid retrieval methods and LLM-augmented pipelines have enabled increasingly capable retrieval systems, and also emphasized the importance of multi-stage architectures in enterprise settings.

In this work, we present an evaluation of alternative implementations across a multi-stage proxy retrieval pipeline designed to support illiquid bond pricing by retrieving comparable issuer candidates, spanning three modules on **information discovery**, **issuer profiling** and **proxy matching**. Besides, we systematically compare design variants for **hierarchical text classification** (HTC) and **reranking** tasks, and analyze their impact on retrieval performance on an internal bond dataset, providing practical guidance on when LLM-centric configurations are sufficient and when hybrid retrieval pipelines offer stronger performance or lower operational cost. Beyond quantitative evaluation, we also share lessons learned from piloting the system in an enterprise environment, highlighting challenges related to data access, infrastructure evolution, and organizational compliance.

We focus on the upstream retrieval problem rather than direct bond valuation. The evaluated system returns ranked issuer-level proxy candidates intended to support analyst review; it does not produce final bond prices, and final CUSIP selection remains analyst-driven. Accordingly, our empirical claims concern retrieval quality, classification accuracy, and ranking performance, not downstream pricing accuracy or realized valuation impact.

Our main contributions are:

- Production-oriented issuer proxy retrieval pipeline to support illiquid bond pricing workflows.
- Modular framework to study pipeline design

*The authors contributed equally to this work

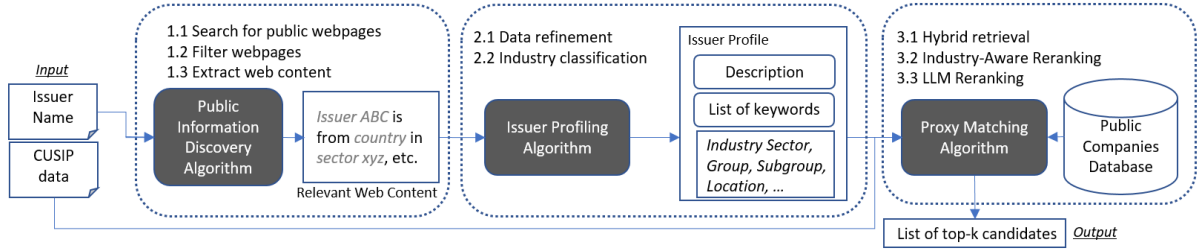


Figure 1: Pipeline for retrieving comparable issuer candidates to support illiquid bond proxy selection.

trade-offs across multi-stage retrieval systems, with a focus on HTC and reranking.

- Pilot deployment takeaways from building and evaluating an enterprise GenAI pipeline.

2 Related Work

Illiquid bond pricing has been widely studied in financial literature, focusing on liquidity effects, yield spreads, and predictive modeling approaches (Gu et al., 2020; Koziol and Sauerbier, 2003; (Meni) Abudy et al., 2018; Longstaff et al., 2004; Chen et al., 2007; Marcato, 2018; Goldstein and Hotchkiss, 2020; Baviera et al., 2021). These studies primarily address price estimation assuming comparable instruments are already known. In contrast, our work focuses on the upstream problem of discovering proxy issuers and bonds.

Financial NLP research has explored information extraction, document retrieval, and hierarchical industry classification using both traditional methods and language models (Chang et al., 2016; Tsai and Wang, 2012; Mavi et al., 2023; Cao et al., 2023). Recent financial-domain retrieval and RAG pipelines further highlight the role of structured retrieval workflows in enterprise settings (Zhao et al., 2024, 2025). These studies typically evaluate individual components or end-task performance, whereas our work integrates NLP components within a unified retrieval pipeline.

Recent advances in retrieval systems emphasize multi-stage architectures combining sparse and dense retrieval with reranking strategies. Despite rapid progress in LLM-driven retrieval and analysis, production workflows often require structured pipelines that balance effectiveness, efficiency, and operational constraints. Prior work examines LLM-based rerankers and reasoning-enhanced ranking strategies (Abdallah et al., 2025; Zhang et al., 2025), primarily focusing on ranking performance in isolation. Our work instead examines pipeline-level design decisions connecting discovery, classi-

fication, and proxy matching stages.

3 Background and Business Context

In production-like pilot settings, proxy identification must search large candidate spaces, often thousands of issuers per batch. The workflow remains largely manual and sensitive to public-information availability, and comparable selection depends heavily on analyst judgment, creating inconsistent outcomes and limited standardization.

Operational Cost: Manual review can take tens of minutes per issuer, scaling to several hundred analyst-hours for a single population; heterogeneity in issuer profiles and subjectivity in proxy choice drive large variance. These inefficiencies compound across the many products that rely on comparable bond identification. **Business and Client Considerations:** In internal pilots, usage increased rapidly, motivating scalable and transparent proxy identification as coverage expands. In parallel, clients increasingly expect systematic, transparent proxy generation with clearer justification and richer issuer context—raising the need for scalable, standardized, and explainable retrieval.

4 Proxy Retrieval Pipeline Methodology

Motivated by operational scale and standardization requirements described in Section 3, we design a modular proxy retrieval architecture that decomposes the task into structured stages: public information discovery, issuer profiling and hierarchical classification, and issuer-level proxy matching, with retrieved issuers supporting downstream analyst review of bond-level CUSIP candidates. This decomposition enables controllable trade-offs across effectiveness, computational cost, and interpretability, while allowing independent evaluation and optimization of components. Figure 1 provides a high-level overview of the architecture and Appendix Figure 3 includes a query example.

4.1 Public Information Discovery

This stage gathers publicly available information for the query issuer (or parent) associated with illiquid and liquid bonds, serving primarily as a high-recall candidate generation step. The process starts with a lightweight Website Search, where it formulates queries from issuer and parent names (e.g., ISSUER_NAME + ABOUT), with optional expansion to aliases, tickers, and sector+country variants. Then, Website Filtering prioritizes reliable sources (issuer site, filings, regulator registries, encyclopedic sources), and targets ABOUT US or equivalent canonical pages. Finally, Content Extraction parses HTML sections and collects structured fields (e.g., HQ location, activities, identifiers) and summaries from high-reliability sources, which are subsequently structured and refined in downstream profiling and classification stages.

4.2 Profiling and Industry Classification

This stage extracts structured issuer profiles from heterogeneous text and performs hierarchical industry classification. **Data refinement** leverages both LLMs and rule-based information extraction to normalize noisy web and document content into structured fields. We systematically tuned prompts for description summarization and keyword extraction, optimizing for the relevance of industry signals used in downstream classification and retrieval. We maintain multi-view features (e.g., description text, keywords, filings-derived cues, infobox attributes); embedding representations can be included as one view but are not relied upon exclusively, and we also evaluate text-only representations.

Industry classification assigns Sector, Group, and Subgroup labels in a three-level hierarchy.¹ Class imbalance poses a significant challenge in hierarchical text classification (HTC), particularly at the node level. Various strategies address this issue, including leveraging class definitions (Rivas Rojas et al., 2020), data augmentation (Wang et al., 2023), and model tuning (Cao et al., 2023).

We evaluate several approaches. First, a bottom-up kNN method with cosine similarity over sentence embeddings: we encode company descriptions and keywords using pretrained transformers and match them to encoded Subgroup definitions, with Group and Sector inherited accordingly. Second, a top-down kNN variant that uses dense class

centroids learned from a train set to predict Sector, then Group, then Subgroup. Third, a pure LLM top-down strategy, where the model sequentially selects the most relevant Sector, then Group, then Subgroup, guided by level-specific definitions. Fourth, a hybrid approach that first fine-tunes a pre-trained language model for Sector prediction and uses prompts to infer Group and Subgroup with an LLM. As a complementary approach, we tested LLM-based reranking of kNN candidates (providing the top- k plus definitions for reordering), which did not improve results over other approaches.

4.3 Multi-Stage Proxy Retrieval

This phase retrieves comparable issuers for proxy selection in illiquid bond pricing workflows, with three stages: (1) hybrid retrieval combining sparse and dense representations, (2) industry-aware soft reranking leveraging predicted industry hierarchies, and (3) neural and LLM-based reranking with calibrated score fusion over top candidates.

Stage 1: Hybrid Retrieval We combine BM25 sparse retrieval with dense semantic embeddings to leverage both lexical and semantic matching signals (Weller et al., 2026). For each query q and document d , we compute:

$$s_{\text{hybrid}}(q, d) = \alpha \cdot s_{\text{BM25}}(q, d) + (1 - \alpha) \cdot s_{\text{dense}}(q, d) \quad (1)$$

where s_{BM25} is the normalized BM25 score and s_{dense} is the cosine similarity between query and document embeddings. The dense score itself combines description and keyword embeddings:

$$s_{\text{dense}}(q, d) = \beta \cdot \text{sim}(e_q^{\text{desc}}, e_d^{\text{desc}}) + (1 - \beta) \cdot \text{sim}(e_q^{\text{kw}}, e_d^{\text{kw}}) \quad (2)$$

Stage 2: Industry-Aware Soft Reranking We leverage industry classification predictions to softly boost candidates with matching industry hierarchies. Given predicted industries for query q (sector, group, subgroup), we compute an industry match score for each candidate d :

$$s_{\text{industry}}(q, d) = w_s \cdot m_s + w_g \cdot m_g + w_{sg} \cdot m_{sg} \quad (3)$$

where m_s, m_g, m_{sg} are match scores for sector, group, and subgroup respectively, with weights w_s, w_g, w_{sg} summing to 1.0. We use rank-based matching where $m_i = 1/\text{rank}$ if the candidate's industry appears in the top- k predictions, else 0.

¹Our taxonomy is proprietary; e.g., ENTERPRISE SOFTWARE Subgroup - SOFTWARE Group - TECHNOLOGY Sector.

The final score fuses retrieval and industry signals:

$$s_2(q, d) = s_{\text{hybrid}}(q, d) + \lambda_{\text{industry}} \cdot s_{\text{industry}}(q, d) \quad (4)$$

Stage 3: Neural + LLM Reranking For the top- k candidates from Stage 2, we apply a neural cross-encoder reranker followed by an optional LLM reranking step.

Stage 3A: Cross-Encoder Fusion The cross-encoder produces a relevance score $s_{\text{ce}}(q, d)$ for each (query, document) pair. We min-max normalize both s_2 and s_{ce} within the candidate set and compute:

$$s_3(q, d) = \lambda_{\text{ce}} \cdot \tilde{s}_2(q, d) + (1 - \lambda_{\text{ce}}) \cdot \tilde{s}_{\text{ce}}(q, d) \quad (5)$$

Stage 3B: LLM Reranking For the top- K candidates from Stage 3A (we use $K=30$), we apply an LLM reranker that produces an ordering over the candidate set. The LLM receives each candidate’s name, description, keywords, and predicted industry context, and outputs a ranked list. The final ranking is determined directly by the LLM’s ordering, without further score interpolation.²

All fusion weights (α , β , $\lambda_{\text{industry}}$, λ_{ce}) are calibrated on a validation set.

CUSIP-Level Use in Analyst Workflow. The retrieval pipeline produces ranked comparable issuer candidates that can support downstream selection of specific liquid bonds (CUSIPs). In this work, we evaluate ranking quality at the issuer level, where supervised relevance labels are available and metrics such as NDCG and MRR can be computed (see §5.4). Final CUSIP selection remains analyst-driven and is outside the scope of our quantitative evaluation; therefore, we do not claim measured improvements in final bond pricing outcomes.

5 Experimental Setup

5.1 Models

We use `msmarco-roberta-base-v2` (Reimers and Gurevych, 2019) and `jina-embeddings-v3` (Sturua et al., 2024) as pretrained sentence embeddings, with 125M and 570M parameters, respectively. We fine-tune the former using `qrels` from our proprietary dataset. For industry classification, we fine-tune `RoBERTa-base` (Liu et al., 2019) as a sector

²When calibrating a fusion weight between pre-LLM and LLM rank scores, the optimal weight was consistently $\lambda_{\text{llm}} \approx 0$, indicating the LLM’s ordering alone is the stronger signal within the shortlisted candidate pool.

classifier. For different tasks, we use two LLMs at different cost tiers: `gpt-4.1` and `gpt-4.1-nano`. Finally, we leverage `jina-reranker-v3` (Wang et al., 2025) as a neural reranker in the last retrieval stage, which we fine-tune with our proprietary data.

5.2 Data

We use an internal dataset of 250 issuers each for validation and test sets, with annotated relevant candidate pairs at three relevance levels: +++ (highest: same subgroup, bond rating, and currency), ++ (medium: same subgroup, and bond rating or currency), + (low: same subgroup only).³ We obtain 1775 and 1545 relevance pairs for validation and test, respectively. Additionally, we source 560k pairs, including negative ones, for fine-tuning `msmarco-roberta-base-v2` and `jina-reranker-v3`. Finally, we use 5k companies with labeled and proportionally stratified sectors for fine-tuning `RoBERTa-base` as an industry sector classifier, preserving the natural class imbalance of the issuer population.

Evaluation independence and annotation. The validation and test query issuer sets are disjoint from each other and from the query issuers used for model fine-tuning. We also exclude held-out test query–candidate relevance pairs from the fine-tuning pool to avoid pair-level leakage between supervised training and final evaluation. Because the retrieval corpus represents the available universe of candidate issuers, some candidate issuers may appear across splits; however, no held-out test relevance judgments are used during fine-tuning or hyperparameter selection.

Graded relevance labels. Candidate issuer pairs were assigned graded relevance labels using domain criteria aligned with analyst proxy-selection practice. A subset of labels was independently reviewed by bond-pricing domain experts, with disagreements resolved through adjudication. We then applied rule-based consistency checks to the full labeled set to verify that each assigned relevance level was consistent with the underlying industry subgroup, bond rating, and currency attributes.

5.3 Fine-Tuning and Hyperparameters

Appendix A.1 presents the hyperparameters used for fine-tuning (i) `msmarco-roberta-base-v2` →

³This evaluation is stricter than typical business practice, where comparable issuers often share only the same group in the taxonomy. This decision ensures a more robust task.

Method	Sector			Group			Subgroup		
	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5
<i>Top-down: kNN with train centroids</i>									
kNN-centroid (msmarco)	56.8	74.4	90.8	47.6	55.2	60.8	24.0	32.8	38.8
kNN-centroid (FT-msmarco)	67.2	80.4	94.0	58.0	66.0	72.4	59.6	61.6	64.8
<i>Top-down: LLM-only</i>									
LLM (gpt-4.1-nano)	64.0	83.2	94.0	52.0	57.6	66.0	25.2	32.0	44.0
LLM (gpt-4.1)	73.2	85.6	94.0	66.0	74.4	80.0	38.8	53.2	68.4
<i>Top-down: Fine-tuned Sector Classifier + LLM</i>									
FT-RoBERTa + LLM (gpt-4.1-nano)	78.0	87.2	95.2	56.0	62.8	70.8	29.6	38.8	52.0
FT-RoBERTa + LLM (gpt-4.1)	78.0	87.2	95.2	66.4	72.0	79.6	40.8	54.0	68.4

Table 1: Industry Classification Results on Test Set: Accuracy at Different Taxonomy Levels (in %). Cells are color-coded by performance within each column (red=low, green=high). Best results in bold.

FT-msmarco; (ii) RoBERTa-base as sector classifier $\rightarrow$ FT-RoBERTa; and (iii) jina-reranker-v3 $\rightarrow$ FT-jina-reranker or FT-Reranker. Besides, Appendix A.2 includes the hyperparameter tuning with the validation set for α , β , $\lambda_{\text{industry}}$, λ_{ce} and industry weights ($w_s = 0.3$, $w_g = 0.2$, $w_{sg} = 0.5$).⁴ For all LLM usage, we set temperature to 0 to ensure deterministic outputs.

5.4 Metrics

We use **Query Success Rate (QSR)** for Public Information Discovery and Multi-Stage Retrieval, measuring the fraction of queries with at least one relevant document in the top- k results. For the latter module, we additionally report **Normalized Discounted Cumulative Gain (NDCG)**, which evaluates ranking quality by assigning higher weight to relevant documents at top positions, and **Mean Reciprocal Rank (MRR)**, which measures the inverse rank of the first relevant document. For industry classification, we report top-1, top-2, and top-5 accuracy at each hierarchy level.

Statistical significance. We also compute pairwise statistical significance tests between methods: **Wilcoxon signed-rank test** for NDCG@k and MRR (paired, non-parametric), and **McNemar’s test** for QSR@k (paired binary outcomes).

6 Results

6.1 Public Information Discovery

We measure the query success rate (QSR) to assess retrieval of relevant content across different types of websites. Table 3 in the Appendix presents QSR scores for 3,000 issuers using DuckDuckGo (DDG) and Google Search (GS). The highest QSR

was achieved from company pages, with DDG at 71.45% and GS at 89.16%. Wikipedia entries showed moderate retrieval rates, with DDG at 70.25% and GS at 68.62%. The variability in QSR scores underscores the need for diverse data sources to enhance coverage. DDG was chosen for the final pipeline due to its non-personalized search behavior and privacy characteristics.

6.2 Hierarchical Industry Classification

Table 1 summarizes results across all approaches, complemented by Appendix Table 4. Bottom-up kNN using node definitions is the weakest baseline (Appendix Table 4). Switching to top-down kNN with train-set centroids markedly improves accuracy (e.g., Subgroup top-1: 26.0% $\rightarrow$ 59.6% with FT-msmarco), but still trails LLM-based methods. Fine-tuning the embedding model consistently outperforms its off-the-shelf counterpart and also surpasses a larger off-the-shelf model (cf. Appendix Table 4), underscoring the value of domain-adapted representations.

LLM-only top-down classification is strong, with gpt-4.1 outperforming gpt-4.1-nano by 9–14 pp in top-1 accuracy across levels (Sector: 73.2% vs. 64.0%; Group: 66.0% vs. 52.0%; Subgroup: 38.8% vs. 25.2%), albeit at $\sim 4\times$ higher API cost under experimental setup. The best overall performance is achieved by the hybrid method (fine-tuned RoBERTa Sector classifier + LLM for Group/Subgroup), which matches or slightly exceeds the LLM-only results (e.g., Sector top-1 78.0%, Group 66.4%, Subgroup 40.8%) while improving cost efficiency by requiring one fewer LLM call per query. We use the predictions of FT-RoBERTa + gpt-4.1 for the industry-aware soft reranking in the next module.

⁴Industry experts suggested to focus on sector-level prediction, but our tuning prioritized subgroup matching.

Method	NDCG			MRR	QSR@10		
	@5	@10	@20		+++	++	+
BM25 + FT-msmarco	33.00	33.27	33.83	47.81	37.60	53.60	60.80
+ Industry Soft	37.07	37.14	36.83	51.80	37.20	54.00	60.40
+ Reranker (base)	20.81	22.93	27.20	34.27	29.60	46.40	53.20
+ FT-Reranker	32.50	32.77	33.02	45.51	39.20	58.00	63.20
+ Industry Soft + FT-Reranker	38.87	38.20	36.98	53.13	40.00	56.80	62.40
+ Industry Soft + gpt-4.1-nano	39.36	38.83	39.24	52.68	40.00	54.80	60.00
+ Industry Soft + gpt-4.1	40.02	39.96	39.66	53.14	41.60	56.40	60.80
+ Industry Soft + FT-Reranker + gpt-4.1-nano	41.20 ^{†,§}	40.22 ^{†,§}	39.53 [§]	53.09	41.20	54.80	60.40
+ Industry Soft + FT-Reranker + gpt-4.1	41.25[‡]	40.68	40.37[§]	53.73	40.80	55.20	60.40

Table 2: Proxy Retrieval Results on Test Set (in %). Cells are color-coded by performance within each column (red=low, green=high). Best results in bold. Relevance levels are +++ (highest), ++ (medium), + (low). Significance markers: [†] gpt-4.1-nano vs. gpt-4.1 ($p < 0.05$), [‡] w/o vs. w/ FT-Reranker ($p < 0.05$), [§] w/o vs. w/ LLM ($p < 0.05$).

6.3 Multi-Stage Proxy Retrieval

Table 2 shows company matching results on the test set (FT-msmarco); complementary trends with msmarco-v2 and jina-embeddings-v3 are in Appendix Table 5. We evaluate the multi-stage retrieval pipeline (§ 4.3) progressively adding each component to isolate its contribution.

Hybrid and industry-aware stages. Domain fine-tuning of the sentence encoder yields the single largest gain: FT-msmarco improves NDCG@10 from 19.89% to 33.27% (+13.4 pp) and MRR from 39.19% to 47.81% over msmarco-v2, confirming that retrieval quality is the primary bottleneck. Industry-aware soft reranking adds a further +4.01 pp NDCG@10 for msmarco-v2 and +3.87 pp for FT-msmarco with consistent QSR improvements, providing a reliable parameter-free boost across all embeddings.

Neural reranking. The off-the-shelf cross-encoder degrades performance sharply (e.g., NDCG@10 drops to 10.33% for msmarco-v2), confirming limited transfer without domain adaptation. Fine-tuning recovers and, when combined with industry soft reranking, yields additive gains (e.g., 38.20% NDCG@10 for FT-msmarco), indicating the two signals are complementary: industry structure captures coarse sector alignment while the cross-encoder captures finer-grained relevance.

LLM-based reranking. Adding an LLM stage over the best candidates provides further improvements, with the full pipeline achieving the highest scores (e.g., 41.25% NDCG@5 and 53.73% MRR for FT-msmarco). Comparing Industry Soft + LLM against Industry Soft + FT-Reranker + LLM shows that the intermediate cross-encoder remains beneficial: for FT-msmarco, its addition yields statisti-

cally significant NDCG gains ($p < 0.05$, Wilcoxon; marked [‡] in Table 2). Conversely, adding the LLM on top of Industry Soft + FT-Reranker produces significant NDCG improvements (marked [§]), yet the best QSR scores for FT-msmarco are achieved by configurations *without* LLM reranking: the LLM reshuffles the top-10 in ways that improve overall rank quality but occasionally displace isolated relevant hits beyond the cutoff.

Choice of LLM. Between the two models, gpt-4.1 attains the best average metrics, but gpt-4.1-nano is comparable—often within < 1 pp NDCG@10. Pairwise tests reveal few statistically significant differences (marked with [†] where present). Unlike industry classification, model selection can therefore be guided by budget considerations, as gpt-4.1-nano offers a more cost-efficient option with near-parity performance.

7 Error Analysis and Discussion

Finally, we analyze key challenges and limitations encountered during our pipeline stages through an error analysis on a held-out test set under our best configuration. A domain expert in bond pricing reviewed 25 queries and provided free-form annotations on 127 non-relevant candidates surfaced by the system. Errors are categorized as in Figure 2.

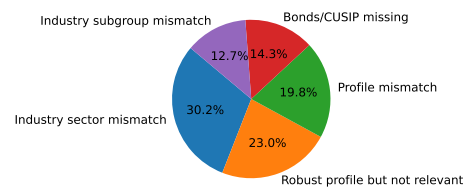


Figure 2: Error Type Distribution.

Industry sector and subgroup mismatches ac-

count for 43% of errors, primarily due to the complexity of multi-label classification and limited industry information in crawled descriptions. This highlights the need for more comprehensive data sources to accurately capture issuers' industry affiliations. **Profile mismatches**, comprising 19.8% of errors, often result from ambiguous names, suggesting the need for additional user input to clarify industry sectors and mitigate ambiguity. For instance, names like "Magnolia Inc." could be misinterpreted (e.g., florist vs. energy company) without detailed descriptions. In 23% of cases, **robust profiles were deemed less relevant**, indicating potential for our pipeline to offer insights beyond expert expectations, as some profiles may provide new perspectives for future matches. **Bonds/CUSIP missing** errors, at 14.3%, highlight issues in the matching stage, affected by errors propagated from earlier modules. This suggests future work to expand search parameters and tune reranker using feedback in a few-shot setting.

8 Deployment Challenges and Takeaways

Piloting the modular GenAI proxy-retrieval system surfaced non-model constraints that materially shaped the design and evaluation. In practice, data availability, governance, and infrastructure evolution influenced architectural choices.

Data Source Obstacles: Changes in external provider access models required repeated ingestion adjustments, motivating modular connectors that allow source substitution without disrupting downstream ranking. **Environment Setup and Integration:** Pilot integration of LLM components introduced friction around authentication, deployment workflows, and infrastructure coordination; standardized integration patterns reduced operational overhead. **Compliance and Governance Evolution:** During pilots, evolving enterprise GenAI governance required additional review cycles to clarify acceptable data usage and logging; transparent intermediate stages improved auditability and collaboration with oversight stakeholders. **Infrastructure and Policy Changes:** Infrastructure updates and evolving data-access controls periodically altered retrieval constraints; modular staging limited blast radius by enabling targeted updates to individual components.

In summary, operationalizing retrieval-centric GenAI systems depends on engineering practices that anticipate technical and organizational con-

straints. See Appendix A.4 for more details.

9 Conclusion

We introduced a modular, production-oriented issuer proxy retrieval pipeline that supports illiquid bond pricing workflows by combining issuer profiling, hierarchical classification, and hybrid retrieval. Results show that domain-adapted representations and industry-aware reranking improve ranking performance while enabling practical trade-offs between accuracy and cost.

The evaluation demonstrates improvements in issuer-level proxy retrieval, while measuring downstream impact on final CUSIP selection or bond pricing remains outside the scope of this work. Future work will explore agentic extensions of the pipeline, coordinating iterative search, evidence aggregation, and profile refinement to better support analyst review of issuer-level proxy candidates.

Limitations

A key limitation of this work is the lack of publicly available datasets for illiquid bond pricing, and in particular, for classifying issuers into industry sectors or retrieving comparable issuers, making it difficult to benchmark across industries. While we use proprietary datasets for building our pipeline and evaluation, data privacy constraints prevent public release. We also do not disclose or directly benchmark against certain analyst-style rule-based workflows, as these may encode proprietary business heuristics and operational practices.

Additionally, the pipeline-based architecture introduces challenges such as cascading errors, where failures in early stages impact later stages. For instance, not being able to retrieve the correct official website of the issuer, will classify it into an incorrect industry, and match with very different public bond issuers.

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JP-MorganChase and its affiliates ("JP Morgan") and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation

for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

©2026 JPMorganChase. All rights reserved

©JPMorgan Chase & Co. All Rights Reserved. JPMorgan Chase Bank, N.A. (“JPMCB”) is a Member FDIC. All services are subject to applicable laws, regulations, and service terms. Not all products and services are available in all geographic areas, and eligibility for particular products and services is subject to final determination by J.P. Morgan and/or its affiliates or subsidiaries. This document is provided for information purposes only and is not intended as a recommendation or an offer or solicitation for the purchase or sale of any security or financial instrument. Although the information contained herein has been obtained from sources believed to be reliable, its accuracy and completeness cannot be guaranteed.

PricingDirect Inc. (“PricingDirect”) is a wholly owned subsidiary of J.P. Morgan Chase & Co. PricingDirect does not provide any accounting, regulatory, tax, investment, or legal advice. Recipients must make an independent assessment of any legal, credit, tax, regulatory, and accounting issues and determine with their own professional advisors any suitability or appropriateness implications and consequences of any transaction in the context of their particular circumstances. PricingDirect is neither a broker dealer nor a member of any exchanges or self-regulatory organizations, does not hold securities or trade, and its pricing services are not intended to be, nor should they be considered, an offer to purchase or sell securities or as a representation that a purchase or sale could be accomplished at such price. Research services referenced are products of J.P. Morgan Securities, LLC (“JPMS, LLC”, Member SIPC, FINRA, the NYSE, and most major exchanges) and/or JPMCB. Securities products and trading services are offered through JPMS, LLC in the United States. J.P. Morgan is a marketing name for the Investor Services businesses of JPMorgan Chase Bank, N.A. and its affiliates worldwide. JPMCB is regulated by the Office of the Comptroller of the Currency in the U.S.A., by the Prudential Regulation Authority in the U.K., and subject to regulation by the Financial Conduct Authority and to limited regulation by the

Prudential Regulation Authority, as well as the regulations of the countries in which it or its affiliates undertake regulated activities. Details about the extent of our regulation by the Prudential Regulation Authority, or other applicable regulators, are available from us on request. All trademarks, service marks, trade names, and logos appearing in this report are the property of their respective owners. Their use herein is solely for identification and reference purposes and does not imply any affiliation, endorsement, or sponsorship.

©2026 PricingDirect Inc. All rights reserved. PricingDirect®, TransparencyDirect®, and PricingStudio® are registered trademarks of JPMorgan Chase & Co. J.P. Morgan Securities LLC is a registered broker dealer and a member of the NYSE, FINRA, and SIPC. v. 20230224

References

- Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. [How good are LLM-based rerankers? an empirical analysis of state-of-the-art reranking models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5693–5709, Suzhou, China. Association for Computational Linguistics.
- Roberto Baviera, Aldo Nassigh, and Emanuele Nastasi. 2021. [A closed formula for illiquid corporate bonds and an application to the european market](#). *Journal of International Financial Markets, Institutions and Money*, 71:101283.
- Lele Cao, Vilhelm von Ehrenheim, Astrid Berghult, Cecilia Henje, Richard Anselmo Stahl, Joar Wandborg, Sebastian Stan, Armin Catovic, Erik Ferm, and Hannes Ingelbag. 2023. [A scalable and adaptive system to infer the industry sectors of companies: Prompt + model tuning of generative language models](#). In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 55–62, Macao. -.
- Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. 2016. [Measuring the information content of financial news](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225, Osaka, Japan. The COLING 2016 Organizing Committee.
- Long Chen, David A Lesmond, and Jason Wei. 2007. Corporate yield spreads and bond liquidity. *The journal of finance*, 62(1):119–149.
- Rian Dolphin, Joe Dursun, Jonathan Chow, Jarrett Blankenship, Katie Adams, and Quinton Pike. 2024. [Extracting structured insights from financial news:](#)

- An augmented llm driven approach. *Preprint*, arXiv:2407.15788.
- Michael A. Goldstein and Edith S. Hotchkiss. 2020. Providing liquidity in an illiquid market: Dealer behavior in us corporate bonds. *Journal of Financial Economics*, 135(1):16–40.
- Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273.
- Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.
- Christian Koziol and Peter Sauerbier. 2003. Valuation of bond illiquidity: An option-theoretical approach.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Francis A Longstaff, Sanjay Mithal, and Eric Neis. 2004. Corporate yield spreads: Default risk or liquidity? new evidence from the credit-default swap market. Working Paper 10418, National Bureau of Economic Research.
- Gianluca Marcato. 2018. Liquidity Pricing of Illiquid Assets. ERES eres2018_215, European Real Estate Society (ERES).
- Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. 2023. Retrieval-augmented chain-of-thought in semi-structured domains. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 178–191, Singapore. Association for Computational Linguistics.
- Menachem (Meni) Abudy, Hadar Binsky, and Alon Raviv. 2018. The effect of liquidity on non-marketable securities. *Finance Research Letters*, 26:139–144.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo. 2020. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2252–2257, Online. Association for Computational Linguistics.
- Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.
- Ming-Feng Tsai and Chuan-Ju Wang. 2012. Visualization on financial terms via risk ranking from financial reports. In *Proceedings of COLING 2012: Demonstration Papers*, pages 447–452, Mumbai, India. The COLING 2012 Organizing Committee.
- Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for listwise document reranking. *Preprint*, arXiv:2509.25085.
- Yue Wang, Dan Qiao, Juntao Li, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2023. Towards better hierarchical text classification with data generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7722–7739, Toronto, Canada. Association for Computational Linguistics.
- Orion Weller, Michael Boratko, Iftexhar Naim, and Jinhyuk Lee. 2026. On the theoretical limitations of embedding-based retrieval. In *The Fourteenth International Conference on Learning Representations*.
- Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. 2025. REARANK: Reasoning re-ranking agent via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2458–2471, Suzhou, China. Association for Computational Linguistics.
- Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. FinRAGBench-V: A benchmark for multimodal RAG with visual citation in the financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4215–4249, Suzhou, China. Association for Computational Linguistics.
- Yiyun Zhao, Prateek Singh, Hanoz Bhathena, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram, and Saket Sharma. 2024. Optimizing LLM based retrieval augmented generation pipelines in the financial domain. In *Proceedings of the 2024 Conference of the North*

A Appendix

A.1 Fine-Tuning Details

A.1.1 Sentence Embedding Model

We fine-tune `msmarco-roberta-base-v2` on domain-specific query-document relevance pairs using Multiple Negatives Ranking Loss (MNRL) (Henderson et al., 2017). Training examples consist of one query, one positive document (e.g., comparable issuer from the same industry subgroup or group), and five randomly sampled negatives per positive. We use AdamW optimizer with learning rate 2×10^{-5} , weight decay 0.01, batch size 16, and train for 3 epochs with 10% linear warmup. Evaluation is performed every 500 steps on the validation set using an Information Retrieval evaluator, and the best model is selected based on validation NDCG@10.

A.1.2 Sector Classifier

We fine-tune RoBERTa-base for sector classification on over 5k labeled issuers with proportionally stratified class distribution, preserving the natural class imbalance. To address imbalance, we apply random oversampling (minimum 50 samples per class) and Focal Loss (Lin et al., 2020) with $\gamma = 2.0$ and inverse-frequency class weights. Input text combines issuer name, description, and keywords separated by [SEP] tokens, truncated to 256 tokens. The classification head consists of two linear layers with LayerNorm, GELU activation, and dropout (0.2).

We perform a focused grid search over learning rate $\{2 \times 10^{-5}, 3 \times 10^{-5}\}$ and focal gamma $\{2.0, 2.5\}$ on the validation set. The best configuration uses learning rate 2×10^{-5} and $\gamma = 2.0$. Final training uses AdamW optimizer with weight decay 0.01, batch size 64, linear warmup (10% of total steps), gradient clipping (max norm 1.0), and early stopping (patience 5 epochs based on validation macro-F1). We train for up to 15 epochs and select the model with highest validation macro-F1.

A.1.3 Reranker

We fine-tune the cross-encoder reranker `jina-reranker-v3` on domain-specific query-document relevance pairs with graded relevance

labels. Each training instance is a (query, document) pair, where both texts are formatted as “Company: {} \n Description: {} \n Keywords: {}” and truncated to a maximum length of 512 tokens. Relevance labels are mapped to continuous regression targets: $0 \rightarrow 0.0$, $2 \rightarrow 0.5$, $3 \rightarrow 0.75$, and $4 \rightarrow 1.0$. We split the proprietary data by query ID into 90% training and 10% validation with no query overlap, and subsample up to 10 negatives per query while retaining all positives.

For parameter-efficient adaptation, we apply LoRA (Hu et al., 2022) to attention projection layers (rank $r = 16$, $\alpha = 32$, dropout 0.1) and fine-tune using AdamW with learning rate 2×10^{-5} , batch size 16, and 10% linear warmup for 5 epochs. Evaluation is run every 4000 steps on the validation set, reporting NDCG@10 and MRR, and we select the final checkpoint based on the best validation NDCG@10.

A.2 Hyperparameter Tuning

We tune hyperparameters on the validation set to optimize retrieval performance. Figure 4 shows the tuning curves for all parameters across all embedding models.

Stage 1: Baseline Retrieval We first tune α (BM25 weight in hybrid retrieval) and β (description weight in dense embeddings). For `msmarco-v2`, optimal values are $\alpha = 0.5$ and $\beta = 0.6$, while the fine-tuned version achieves best performance with $\alpha = 0.2$ and $\beta = 0.6$, and `jina-v3` with $\alpha = 0.3$ and $\beta = 0.6$. The lower optimal α for `msmarco-v2-ft` indicates its stronger semantic embeddings require less reliance on lexical matching.

Stage 2A: Industry Hierarchy Weight Tuning

To optimize the industry matching component in soft reranking, we tune the relative weights ($w_{\text{sector}}, w_{\text{group}}, w_{\text{subgroup}}$) for the three-level industry hierarchy. Testing 24 weight configurations on the validation set, we find that the optimal weights are (0.3, 0.2, 0.5)—heavily favoring the most specific subgroup level—achieving NDCG@10 of 0.399 (test: 0.374). This is notably different from the intuitive sector-heavy configuration (0.5, 0.3, 0.2), which achieves NDCG@10 of 0.392. The subgroup-only configuration (0.0, 0.0, 1.0) also performs competitively (NDCG@10: 0.398), suggesting that fine-grained industry classifications provide stronger discriminative signals for company matching than broader categories.

Stage 2B: Industry-Aware Reranking We tune $\lambda_{\text{industry}}$ (industry boost weight) for soft reranking with predicted industry hierarchies. The optimal values differ between models: $\lambda = 0.4$ for msmarco-v2 and $\lambda = 0.2$ for both msmarco-v2-ft and jina-v3. The lower λ for the later two models suggests its embeddings already capture some industry-related semantic patterns, requiring less explicit industry boosting.

Stage 3: Fusion Weight Tuning with FT-Reranker We additionally tune λ_{ce} (cross-encoder fusion weight) to optimally combine industry soft reranking and the fine-tuned neural reranker. All models achieve their best performance at $\lambda_{\text{ce}} = 0.7$, indicating that the fine-tuned cross-encoder provides complementary signals to explicit industry matching. This effect is not observed for the base neural reranker, which does not improve over industry soft reranking and is therefore excluded from fusion weight analysis.

A.3 Complementary Results

Public Information Discovery Table 3 presents the results for the Public Information Discovery module.

	DuckDuckGo	Google Search
Company/Other pages	71.45%	89.16%
Wikipedia entry	70.25%	68.62%
Infobox Industries	51.80%	47.67%
No Information	15.49%	3.61%

Table 3: Query Success Rate results for Discovery

Industry Classification Table 4 presents complementary results for jina-embeddings-v3 and the bottom-up approach of LLM reranking for kNN predictions.

Proxy Retrieval Table 5 presents complementary results for proxy matching using jina-embeddings-v3 in the initial stage.

A.4 Deployment Challenges in Detail

This section provides additional implementation context for the pilot deployment observations summarized in Section 8. The main challenges were not model-specific; they related primarily to data access, enterprise integration, governance review, and infrastructure constraints.

Data source changes. Early prototypes relied on automated web collection from a primary search provider to construct issuer profiles. Changes to

provider access policies and subscription interfaces required adjustments to the ingestion layer and evaluation of alternative sources. This motivated a modular ingestion design, so that source changes could be handled without redesigning downstream profiling or retrieval components.

Enterprise integration. Integrating GenAI components into enterprise environments introduced coordination needs around authentication, service accounts, deployment workflows, access patterns, and monitoring. Differences between development and pilot deployment environments required additional configuration work to align the system with organizational standards.

Governance and compliance. As review procedures and documentation expectations evolved, design decisions around data handling, logging, model usage, and evaluation practices required iterative alignment with oversight stakeholders. Transparent intermediate stages helped make the system easier to review and audit.

Infrastructure constraints. Infrastructure updates and data-access controls affected network routing, approved external data sources, and retrieval configuration. Configurable ingestion layers and loosely coupled pipeline stages helped isolate these changes and reduce disruption to the broader system.

Overall, these experiences reinforced the importance of modularity, auditability, and configurable components when moving retrieval-centric GenAI systems from prototype development to controlled pilot deployment in enterprise settings.

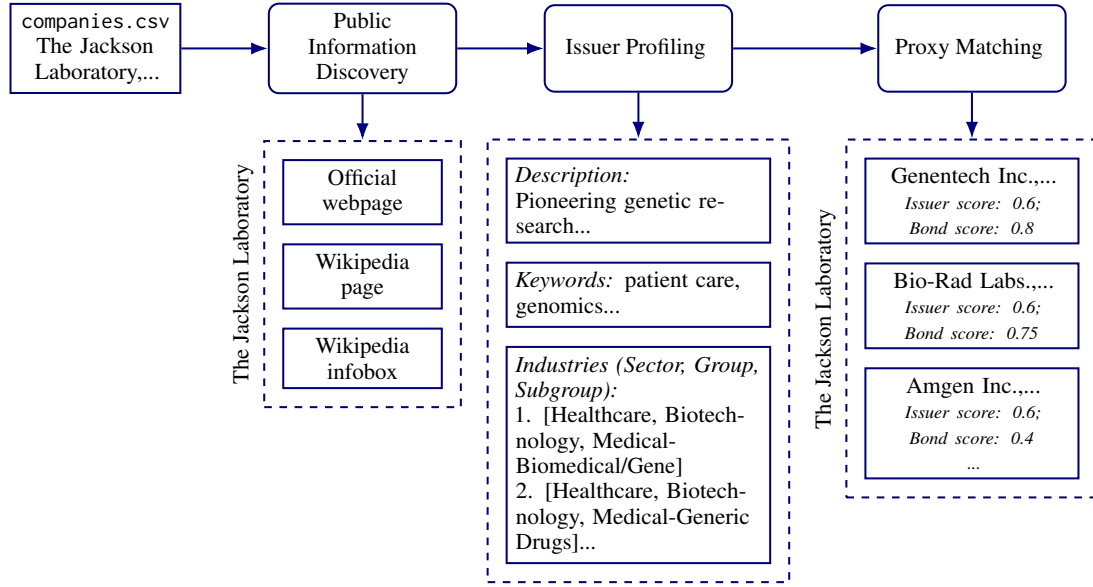


Figure 3: Example workflow for identifying pricing proxies for illiquid corporate bonds.

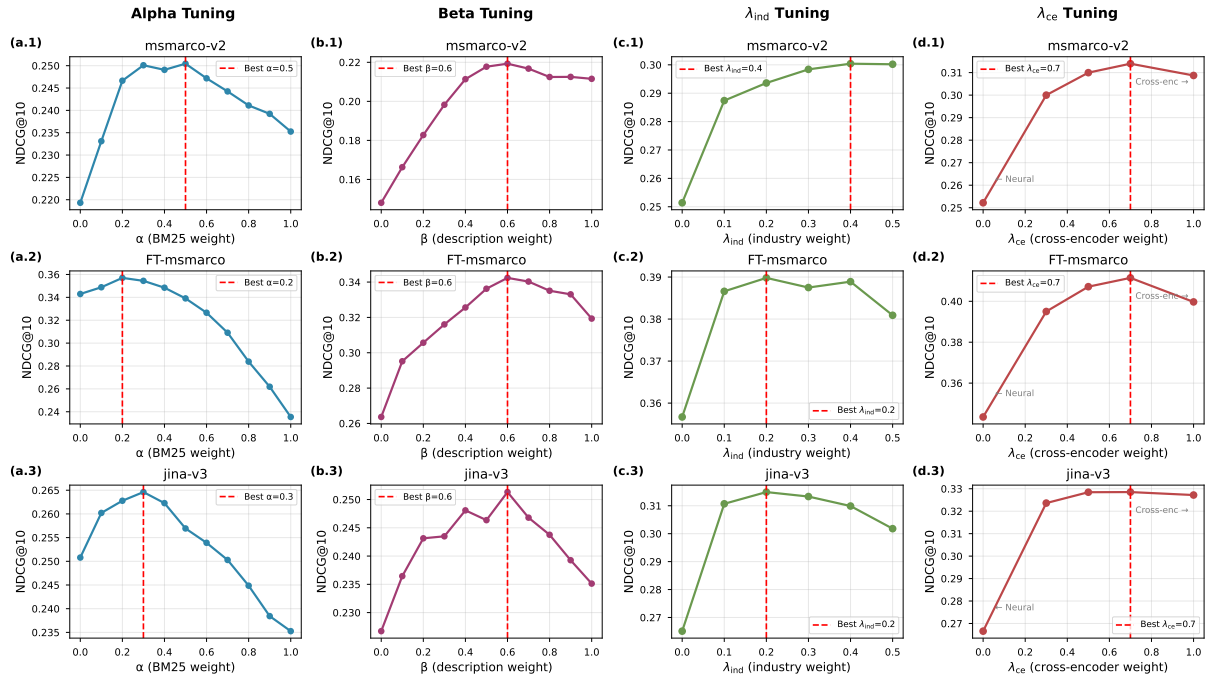


Figure 4: Hyperparameter tuning curves on validation set for three embedding models across rows. (a.1-a.3) Alpha tuning for BM25 weight in hybrid retrieval. (b.1-b.3) Beta tuning for description weight in dense embeddings. (c.1-c.3) Lambda tuning for industry boost weight in soft reranking. (d.1-d.3) Lambda tuning for cross-encoder fusion. All metrics shown are NDCG@10 on the validation set.

Method	Sector			Group			Subgroup		
	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5	Top-1	Top-2	Top-5
<i>Bottom-up: kNN with definitions</i>									
kNN-def (msmarco)	50.8	59.6	68.8	36.0	46.0	56.8	17.6	26.4	37.2
kNN-def (FT-msmarco)	66.0	73.2	78.0	52.0	60.4	69.2	26.0	36.0	51.6
kNN-def (jina-v3)	56.8	66.0	72.8	44.0	53.2	63.2	22.4	32.0	46.0
<i>Bottom-up: kNN + LLM reranking</i>									
kNN-def (msmarco-v2) + gpt-4.1-nano	64.0	69.6	72.8	48.0	55.6	64.8	28.4	34.4	46.0
kNN-def (msmarco-v2) + gpt-4.1	71.6	74.4	79.2	62.0	66.4	72.4	34.4	46.0	56.8
kNN-def (FT-msmarco) + gpt-4.1-nano	73.6	76.0	78.4	60.0	65.2	73.2	34.0	44.8	58.8
kNN-def (FT-msmarco) + gpt-4.1	76.4	80.8	83.2	65.6	71.6	78.4	38.8	52.0	67.6
kNN-def (jina-v3) + gpt-4.1-nano	66.8	71.6	74.4	52.4	58.8	66.0	28.4	36.8	50.0
kNN-def (jina-v3) + gpt-4.1	71.6	76.8	80.0	61.2	69.2	74.0	39.2	49.2	60.8
<i>Top-down: kNN with train centroids</i>									
kNN-centroid (jina-v3)	60.0	78.4	91.2	49.2	62.4	70.0	27.6	36.4	46.4

Table 4: Complementary results for Industry Classification Results on Test Set: Accuracy at Different Taxonomy Levels (in %). Cells are color-coded by performance within each column (red=low, green=high).

Method	NDCG			MRR	QSR@10		
	@5	@10	@20		+++	++	+
BM25 + msmarco-v2	21.25	19.89	19.14	39.19	31.60	44.40	50.80
+ Industry Soft	25.91	23.90	21.52	42.97	32.80	47.20	52.80
+ Reranker (base)	9.34	10.33	13.01	20.61	23.20	36.80	42.40
+ FT-Reranker	22.95	21.06	19.12	36.58	33.60	48.00	53.20
+ Industry Soft + FT-Reranker	26.79	24.23	21.55	43.74	32.40	47.20	53.20
+ Industry Soft + gpt-4.1-nano	29.43	26.87	24.62	45.04	34.00	49.60	54.80
+ Industry Soft + gpt-4.1	30.71	27.82	25.53	48.24 [†]	35.60	51.20	56.40
+ Industry Soft + FT-Reranker + gpt-4.1-nano	30.41 [§]	27.54 [§]	24.98 [§]	47.44 ^{†,§}	33.60	49.60	55.20
+ Industry Soft + FT-Reranker + gpt-4.1	30.96 [§]	28.08 [§]	25.64 [§]	47.99	35.60 [§]	50.80	56.40
BM25 + jina-embeddings-v3	22.89	22.09	21.62	41.06	31.20	48.00	54.80
+ Industry Soft	29.25	27.39	25.15	47.24	33.20	52.00	58.80
+ Reranker (base)	11.21	12.52	15.48	21.98	25.60	42.00	48.40
+ FT-Reranker	24.05	22.90	21.40	37.77	34.40	52.00	58.00
+ Industry Soft + FT-Reranker	29.96	27.77	25.17	47.79	34.00	52.80	60.00
+ Industry Soft + gpt-4.1-nano	33.04	30.65	28.48	49.70	35.60	54.00	59.60
+ Industry Soft + gpt-4.1	34.12	31.87	29.57	52.36	35.60	54.80	60.80
+ Industry Soft + FT-Reranker + gpt-4.1-nano	33.37 [§]	30.91 [§]	28.24 [§]	50.20	36.00	54.40	60.80
+ Industry Soft + FT-Reranker + gpt-4.1	35.23 ^{†,§}	32.23 [§]	29.60 ^{†,§}	52.97 [§]	36.40	55.20	61.20

Table 5: Complementary Proxy Retrieval Results on Test Set (in %). Cells are color-coded by performance within each column (red=low, green=high). Best results in bold. Relevance levels are +++ (highest), ++ (medium), + (low). Significance markers ($p < 0.05$): [†] gpt-4.1-nano vs. gpt-4.1, [‡] w/o vs. w/ FT-Reranker, [§] w/o vs. w/ LLM.