NLP-Driven Proxy Retrieval for Illiquid Bond Pricing

Arturo Oncevay[*]
JPMorgan AI Research

Joy Sain[*]
JPMorgan AI Research

Simerjot Kaur
JPMorgan AI Research

Nichole Ling
JPMorgan PricingDirect

Charese Smiley
JPMorgan AI Research

Xiaomo Liu
JPMorgan AI Research

Manuela Veloso
JPMorgan AI Research

Abstract

In finance, pricing illiquid bonds is a complex challenge due to their infrequent trading and market data scarcity. This paper presents a novel generative AI and NLP-based framework to retrieve liquid proxy bonds, enhancing scalability. Our end-to-end pipeline comprises three modules: (i) Public Information Discovery, (ii) Profiling, and (iii) Matching. Using web data and Large Language Models (LLMs), we generate descriptive summaries and keywords for illiquid bonds and match them with liquid candidates, reducing manual effort. Rigorous evaluation achieved a 71.4% query success rate, and the scalable solution, ∼9x faster than a manual approach, has been well-received by industry experts. We are now deploying this pipeline to production, aiming to improve the process of illiquid bond pricing.

1 Introduction

Accurate bond pricing is crucial for financial markets, guiding investment decisions. Key factors influencing bond prices include interest rates, credit risk, and market conditions. While liquid bonds are easily priced due to available data, illiquid bonds pose challenges due to infrequent trading and data scarcity [Gu et al., 2020, Koziol and Sauerbier, 2003], leading to potential inaccuracies in valuations and risk assessments. Accurate valuations are essential for regulatory compliance and asset allocation. A common approach is to identify comparable issuers and use their liquid bonds as proxies, ensuring valuations reflect market sentiment and conditions.

Existing bond pricing models often focus on direct price estimation using ML [Huang et al., 2023, Dolphin et al., 2024] but overlook the complexity of identifying suitable proxies. Prior research highlights liquidity's role in valuation [(Meni) Abudy et al., 2018, Longstaff et al., 2004, Chen et al., 2007, Marcato, 2018, Goldstein and Hotchkiss, 2020, Baviera et al., 2021], yet many models rely solely on quantitative data, neglecting issuer-specific insights. Despite advancements in financial NLP [Chang et al., 2016, Tsai and Wang, 2012, Mavi et al., 2023], no studies address CUSIP-level proxy identification using generative AI (see more related work in §B.3 in the Appendix).

We propose a generative AI and NLP-driven methodology that automates the discovery and profiling of comparable issuers for pricing illiquid bonds. Our pipeline consists of three modules: (i) Public Information Discovery, gathering issuer data from the web; (ii) Issuer Profiling, extracting and summarizing issuer-specific details with generative models; and (iii) Proxy Matching, identifying

[*]The authors contributed equally to this work

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

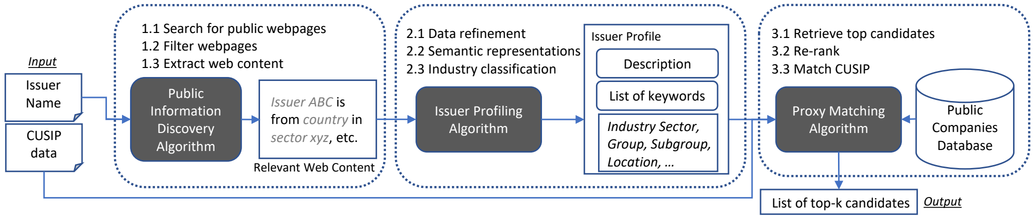

Figure 1: End-to-end pipeline to find a comparable for the issuer of illiquid bond.

liquid proxies using a three-stage retrieval approach. Our work provides a comprehensive evaluation, including quantitative metrics and human assessment, and advances AI-driven bond pricing.

2 Methodology

To identify comparable issuers for illiquid bonds, we propose an end-to-end pipeline leveraging NLP and generative AI to automate this process, illustrated in Fig. 1, and structured into three modules (Figs. 4, 5 and 6 in the Appendix include more details per module):

2.1 Public Information Discovery

This module systematically gathers publicly available information on issuers of illiquid and liquid bonds, focusing on unique CUSIPs linked to parent companies. The process begins with **Website Search**, wherein search queries are formulated using the issuer's name (e.g., ISSUER_NAME + ABOUT) and retrieving results via a search engine. If the initial results yield insufficient information, we expand the query scope by including the parent company's name.

We then perform **Website Filtering** to ensure accuracy and reliability. We prioritize reliable sources such as the company's official website and Wikipedia. The filtering algorithm specifically targets ABOUT US pages, selecting relevant sites for extraction.

Finally, in **Content Extraction**, relevant information is retrieved and processed. For company websites, HTML parsing isolates key sections from ABOUT US pages. For Wikipedia, the MediaWiki API is used to collect summaries, extract infobox data, and parse additional content.

2.2 Issuer Profiling

In this module, we use an LLM[*] to build an issuer profile based on the gathered data. Generative models, such as LLMs, are robust models for summarization, information extraction [Chang et al., 2024], and do not require task-specific labeled data [Huang et al., 2023, Dolphin et al., 2024].

First, we focus on **Data Refinement** of the raw web-extracted data to ensure relevance. Using various prompts, we extract key details such as the issuer's location of operation, a concise summary of its business activities, and industry-related keywords that aid in classification and market positioning.

Second, we obtain the **Semantic Representation** of the extracted information. We convert the keywords obtained from the Wikipedia infobox into complete sentences that represent the issuer's areas of operation. For instance, CLOUD COMPUTING and DATA ANALYTICS become "THE COMPANY SPECIALIZES IN CLOUD COMPUTING AND DATA ANALYTICS". We then generate vector representations using a pretrained Transformer-based sentence embedding model [Reimers and Gurevych, 2019]. The embeddings capture different aspects of an issuer's profile, forming a multi-view semantic representation using LLM-generated summaries, extracted keywords from the company website or Wikipedia, plus mean and max-pooling vector operations to further refine the representations.

The third step is the **Industry Classification** of the issuers into a three-level hierarchy of Sectors, Groups, and Subgroups, which is predefined from our business.[*] For this hierarchical classification

[*]For all our work, we used GPT-4o [OpenAI et al., 2024] with temperature value at 0.

[*]Due to proprietary reasons, we cannot disclose details of the full taxonomy. As an example, the ENTERPRISE SOFTWARE subgroup belongs to the SOFTWARE Group and the TECHNOLOGY Sector

task, we follow a bottom-up approach and use a Nearest-Neighbor classifier plus cosine similarity as the distance metric, comparing Subgroup definitions with issuer representations. To do so, we also leverage LLMs to obtain definitions for each Subgroup. The top-5 predictions from each feature embedding (e.g. keywords or summary) are aggregated, ensuring a more robust classification.

2.3 Proxy Matching

In the final module, our goal is to identify a list of comparable candidates and relevant CUSIPs of issuers with liquid bonds. This process involves a three-stage retrieval system that combines semantic representation and weighted ranking to derive the most relevant and comparable candidates.

The first stage is **Retrieve Top Candidates**, which involves calculating the cosine similarity between the embedding representations of the query issuer's profile (description and keywords) with those of issuers with liquid bonds from a large bonds database. In other words, for each query, we retrieve the top-k candidates based on a semantic similarity approach. We use a parameter α (ranging from 0 to 1) to weight the contributions of the description and keywords in the similarity calculation.

The second retrieval stage is **Re-rank**, where we use a reranker model for the retrieved top-k candidates. Typically, a reranker is a Transformer-based model fine-tuned for reordering inputs based on textual descriptions. Lacking annotated data for training, we leverage an LLM as a zero-shot reranker. The LLM receives a structured (JSON formatted) prompt, including a task description, query inputs (issuer name, description summary, keywords), and the initial ranked list of top-k candidates' details, and returns a reranked list.

The final retrieval stage is **CUSIP Matching**, where we expand the retrieval system to the CUSIP level to find the best proxies for illiquid issuers. This involves analyzing various categorical market factors such as location, currency, and industry sector, each assigned a weight to reflect its importance.[*]. The weighted score determines the degree to which each candidate matches user-specified criteria, and the top-k comparable issuers' CUSIP entries are ranked based on the maximum weighted score.

3 Evaluation and Results

Each component of the pipeline has a distinct objective and requires specific evaluation, yet the overall evaluation is interconnected, especially for the last module of Proxy Matching.

Public Information Discovery We measure the query success rate (QSR) to assess the retrieval of relevant content across different type of websites. Table 1 in the Appendix presents QSR scores for 3,000 issuers using DuckDuckGo (DDG) and Google Search (GSearch). The highest QSR was achieved from company pages, with DDG at 71.45% and GSearch at 89.16%. Wikipedia entries showed moderate retrieval rates, with DDG at 70.25% and GSearch 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 strong privacy focus and unbiased results from non-personalized searches.

Issuer Profiling Evaluating issuer profiling, a generation task, is challenging due to the subjective nature of outputs. Therefore, we assess performance through a downstream task on text classification, generating descriptions and keywords to classify companies into industry sectors, groups, and subgroups. Table 2 in the Appendix reports accuracy scores using the MSMARCO-ROBERTA-BASE-V2 pretrained model [Reimers and Gurevych, 2019]. An ablation study highlights the significance of feature views, including summaries, keywords, and infobox industries. At the sector level, accuracy reaches 83.02% for top-2 and 87.98% for top-5 predictions. Group classification achieves 70.81% for top-2 and 79.51% for top-5 predictions, demonstrating the effectiveness of combined features. Subgroup classification, being more specific, has lower accuracy at 58.44% for top-5 predictions, consistent with literature on hierarchical text classification [Rivas Rojas et al., 2020, Cao et al., 2023]. The "Company Keywords" feature is pivotal, with its removal leading to a decline in performance.

Proxy Matching and Pipeline Evaluation Proxy Matching is a retrieval task where we assess the success of candidates in our top-k list using a human-annotated dataset. We conducted annotations on two batches: 25 query issuers with 236 annotations of potential candidates for validation, and 200

[*]Due to proprietary reasons, we cannot disclose more details about the specific variables and the weights.

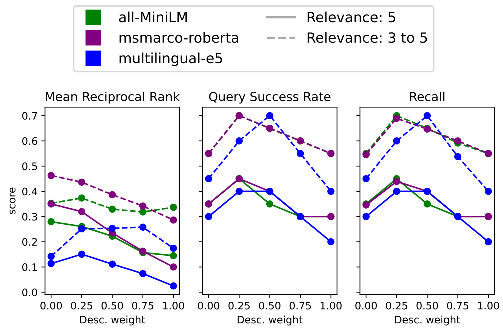

Figure 2: Benchmarking and tuning of α (x-axis, "description weight") in the validation set.

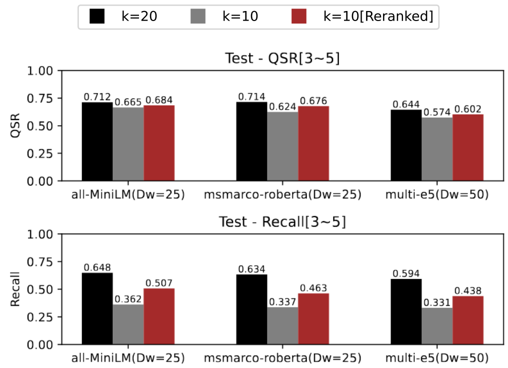

Figure 3: Reranking results (top-k=10) for the best setting per embedding model in the test set.

query issuers with 3,639 annotations for testing. Annotators scored candidates from 1 to 5. A score of 5 often indicates a proxy already in production. Experts provided comments, and discussions led to the decision to treat scores of 3 to 5 as relevant. This process ensures that our system aligns with expert judgment in identifying suitable proxies.

In the tuning phase, we compared three lightweight (<120M params.) sentence embedding models: MSMARCO-ROBERTA-BASE-V2, ALL-MINILM-L6-V2 [Reimers and Gurevych, 2019], and MULTILINGUAL-E5-SMALL [Wang et al., 2024], tuning the α parameter from 0 to 1 using the validation set (see Fig. 2). We focused on QSR and Recall, setting k to 20. Results showed similar performance across models, with α=0.25 optimal for the first two models and α=0.5 for the third. Furthermore, with the test set, we observe that the reranking model improved Recall, indicating the potential of using an LLM as a zero-shot reranker (see Fig. 3). Overall, the highest QSR for relevant scores (3-5) was 71.4%, demonstrating the system's effectiveness in identifying suitable proxies and expanding the pool of potential candidates for illiquid bond pricing. We note that the success of CUSIP matching depends on the initial selection of candidates, as relevant candidates are likely to have CUSIP-level details that align well with the weighted score of categorical variables. Finally, regarding **speed and efficiency**, our pipeline processes each issuer in approximately 1.7 minutes, which is about 9 times faster than manual expert analysis, which takes around 15 minutes per issuer.

4 Error Analysis and Discussion

Finally, we analyze the key challenges and limitations encountered during our pipeline stages using expert feedback on 127 non-relevant answers from the test set, employing the MSMARCO-ROBERTA-BASE-V2 model with $\alpha = 0.25$. Errors were categorized into groups (see Fig. 7 in the Appendix):

Industry sector and subgroup mismatches account for 43% of errors, primarily due to the complexity of multi-label classification and insufficient 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 (either as a florist or 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 the reranker using feedback in a few-shot setting. Additionally, **website restrictions** issues, not included in expert feedback, impede data extraction due to cookies and security measures, necessitating alternative data sources or methods to bypass scraping obstacles.

5 Conclusion

The integration of generative AI and NLP techniques in bond pricing marks a significant advancement in financial technology. By automating the identification and profiling of comparable companies and enabling expert-assisted decisions, our approach addresses previous limitations, enhancing efficiency and accuracy. This research contributes to more transparent financial markets and sets a new standard for collaboration between human expertise and technology. Future work will focus on enhancing data collection by incorporating alternative sources and overcoming web scraping restrictions, as well as implementing a multi-label classification system for industry sectors and integrating user feedback to refine issuer descriptions, thereby strengthening our illiquid bond pricing framework.

Limitations The work faces limitations due to the lack of publicly available datasets, reliance on proprietary data and constraints in comparing diverse models. More details are in §A in the Appendix.

References

Roberto Baviera, Aldo Nassigh, and Emanuele Nastasi. A closed formula for illiquid corporate bonds and an application to the european market. *Journal of International Financial Markets, Institutions and Money*, 71:101283, 2021. ISSN 1042-4431. doi: https://doi.org/10.1016/j.intfin.2021.101283. URL https://www.sciencedirect.com/science/article/pii/S1042443121000020.

Lele Cao, Vilhelm von Ehrenheim, Astrid Berghult, Cecilia Henje, Richard Anselmo Stahl, Joar Wandborg, Sebastian Stan, Armin Catovic, Erik Ferm, and Hannes Ingelhag. A scalable and adaptive system to infer the industry sectors of companies: Prompt + model tuning of generative language models. In Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 55–62, Macao, 20 August 2023. -. URL https://aclanthology.org/2023.finnlp-1.5.

Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. Measuring the information content of financial news. In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1303.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Long Chen, David A Lesmond, and Jason Wei. Corporate yield spreads and bond liquidity. *The journal of finance*, 62(1):119–149, 2007.

Rian Dolphin, Joe Dursun, Jonathan Chow, Jarrett Blankenship, Katie Adams, and Quinton Pike. Extracting structured insights from financial news: An augmented llm driven approach, 2024. URL https://arxiv.org/abs/2407.15788.

Yiyang Geng. Machine learning in the chinese corporate bond market. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 84–90, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400717147. doi: 10.1145/3675417.3675432. URL https://doi.org/10.1145/3675417.3675432.

Michael A. Goldstein and Edith S. Hotchkiss. Providing liquidity in an illiquid market: Dealer behavior in us corporate bonds. *Journal of Financial Economics*, 135(1):16–40, 2020. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2019.05.014. URL https://www.sciencedirect.com/science/article/pii/S0304405X19301394.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273, 02 2020. ISSN 0893-9454. doi: 10.1093/rfs/hhaa009. URL https://doi.org/10.1093/rfs/hhaa009.

Olivier Guéant and Iuliia Manziuk. Deep reinforcement learning for market making in corporate bonds: Beating the curse of dimensionality. *Applied Mathematical Finance*, 26(5):387–452, 2019. doi: 10.1080/1350486X.2020.1714455. URL https://doi.org/10.1080/1350486X.2020.1714455.

Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

Christian Koziol and Peter Sauerbier. Valuation of bond illiquidity: An option-theoretical approach, 2003. URL http://dx.doi.org/10.2139/ssrn.424282.

Francis A Longstaff, Sanjay Mithal, and Eric Neis. Corporate yield spreads: Default risk or liquidity? new evidence from the credit-default swap market. Working Paper 10418, National Bureau of Economic Research, April 2004. URL http://www.nber.org/papers/w10418.

Seyed Mohammad Mansouri and Dalibor S. Eterovic. Machine learning and the cross-section of emerging market corporate bond returns, 2023. URL http://dx.doi.org/10.2139/ssrn.4632924.

Gianluca Marcato. Liquidity Pricing of Illiquid Assets. ERES eres2018_215, European Real Estate Society (ERES), January 2018. URL https://ideas.repec.org/p/arz/wpaper/eres2018_215.html.

Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. Retrieval-augmented chain-of-thought in semi-structured domains. In Daniel Preoțiuc-Pietro, Catalina Goanta, Ilias Chalkidis, Leslie Barrett, Gerasimos Spanakis, and Nikolaos Aletras, editors, *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 178–191, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nllp-1.18. URL https://aclanthology.org/2023.nllp-1.18.

Menachem (Meni) Abudy, Hadar Binsky, and Alon Raviv. The effect of liquidity on non-marketable securities. *Finance Research Letters*, 26:139–144, 2018. ISSN 1544-6123. doi: https://doi.org/10.1016/j.frl.2017.12.017. URL https://www.sciencedirect.com/science/article/pii/S1544612317307109.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine

Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Nils Reimers and Iryna Gurevych. 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, 11 2019. URL http://arxiv.org/abs/1908.10084.

Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2252–2257, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.205. URL https://aclanthology.org/2020.acl-main.205.

Ming-Feng Tsai and Chuan-Ju Wang. Visualization on financial terms via risk ranking from financial reports. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012: Demonstration Papers*, pages 447–452, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://aclanthology.org/C12-3056.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024. URL https://arxiv.org/abs/2402.05672.

A Limitations

A key limitation of this work is lack of publicly available datasets for illiquid bond pricing, and in particular, for classifying issuers into industry sectors or retrieve 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.

While we compare different pretrained sentence embedding models and use GPT-4o as the official LLM for all the pipeline, due to compute and production environment constraints we have not been able to compare with a more diverse set of embedding models or LLMs.

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.

B Supplementary Material

B.1 Methodology

Our pipeline is composed by three modules. Each of them is detailed in Figures 4, 5 and 6.

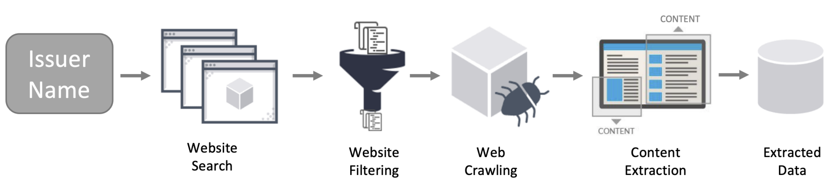

Figure 4: Public Information Discovery Architecture

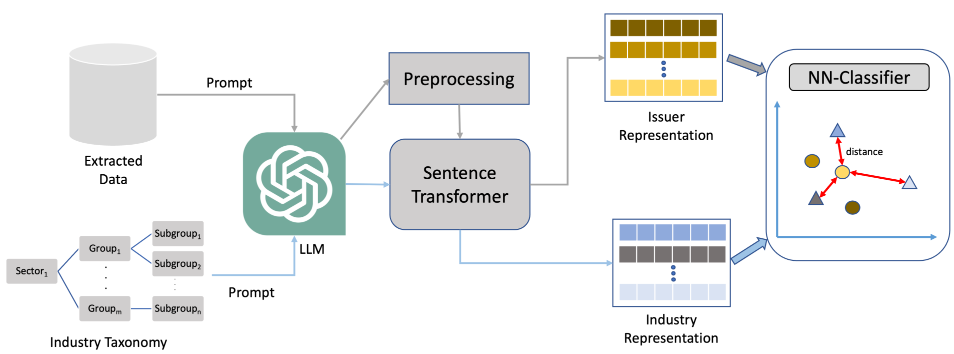

Figure 5: Issuer Profiling Architecture

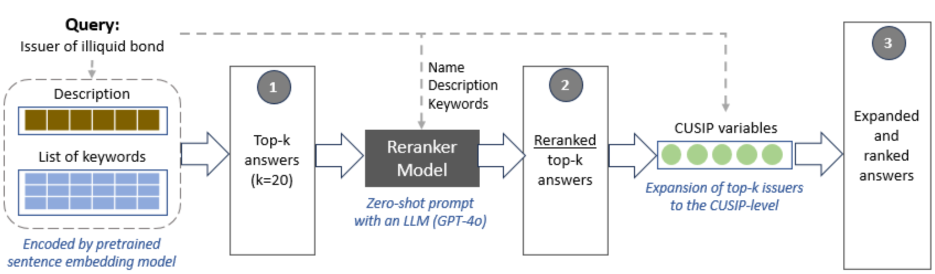

Figure 6: Proxy Matching Architecture: Three-Stage Retrieval

B.2 Results

Table 1 presents the results for the Public Information Discovery module, and Table 2 shows the results for the industry classification task in the Issuer Profiling 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 1: Query Success Rate results for Discovery

	Top-2 Accuracy			Top-3 Accuracy			Top-5 Accuracy		
Features	Sector	Group	Subgroup	Sector	Group	Subgroup	Sector	Group	Subgroup
All Features	**83.02%**	**70.81%**	**46.47%**	**85.13%**	**74.98%**	**51.58%**	**87.98%**	**79.51%**	**58.44%**
- Company Summary	81.84%	68.86%	44.75%	83.92%	73.31%	50.04%	86.89%	78.02%	56.95%
- Wikipedia Summary	82.00%	68.86%	44.61%	84.12%	73.36%	49.88%	87.08%	78.46%	56.98%
- Company Keywords	81.76%	68.48%	44.12%	84.14%	72.89%	48.94%	87.35%	77.83%	55.99%
- Wikipedia Keywords	82.63%	69.96%	45.13%	84.83%	74.1%	50.34%	87.68%	78.88%	57.23%
- Infobox Industries	82.22%	70.04%	45.21%	84.58%	74.43%	50.59%	87.43%	78.93%	57.64%
- Max-Pooled All Features	83.02%	70.51%	46.09%	85.1%	74.76%	51.36%	87.96%	79.29%	58.22%
- Mean-Pooled All Features	82.94%	70.7%	46.26%	85.08%	74.87%	51.47%	87.93%	79.48%	58.33%

Table 2: Issuer Profiling Results for Industry Classification and Ablation Study for Top-N Predictions

Finally, Figure 7 includes the error type distribution in the error analysis.

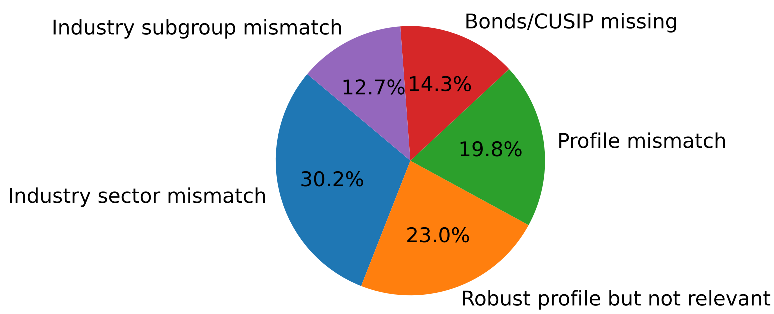

Figure 7: Error Type Distribution.

B.3 Related Work

The challenge of pricing illiquid bonds has been a focal point in financial research, with various studies exploring different methodologies. Direct pricing methods have been examined in works such as Gu et al. [2020], Koziol and Sauerbier [2003], which highlight the complexities and limitations inherent in valuing bonds that lack frequent trading data.

Several studies have underscored the critical role of liquidity in bond pricing. For instance, (Meni) Abudy et al. [2018], Longstaff et al. [2004], Chen et al. [2007], Marcato [2018], Goldstein and Hotchkiss [2020], and Baviera et al. [2021] have shown that failing to account for liquidity can lead to significant pricing errors. These models, while insightful, often rely solely on quantitative data, potentially overlooking qualitative factors that experts consider crucial in financial decision-making.

The application of ML methods to bond pricing has also garnered significant attention. Studies such as Guéant and Manziuk [2019], Mansouri and Eterovic [2023], Geng [2024] have explored various ML techniques to enhance the accuracy of bond pricing models. These approaches have demonstrated promise in improving pricing precision but often bypass the nuanced process of identifying and analyzing comparable liquid bonds. This oversight can result in models that fail to capture the specificities associated with different issuers, leading to less accurate pricing.

Moreover, while there has been some work on using NLP methods in finance for information retrieval [Chang et al., 2016, Tsai and Wang, 2012, Mavi et al., 2023], these studies do not address the specific challenge of finding suitable proxies for illiquid bonds. The closest related work to one of our components, Issuer Profiling, is by Cao et al. [2023], who employed generative models to classify industry sectors. However, their focus was not on bond pricing or proxy identification.

Our approach aims to fill this gap by leveraging generative AI and NLP techniques to gather relevant market data, analyze issuer-specific and CUSIP-level features, and combine these insights to retrieve relevant CUSIP-level proxies for pricing illiquid bonds.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are reflected in the rest of the paper, where we evaluated our approach in both a quantitative and qualitative way.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We include a Limitations sections after the bibliography: §A.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: We partially disclose all the information to reproduce the main experimental result. We include models and parameters, but due to proprietary reasons, we cannot disclose information related to the industry taxonomoy, CUSIP variables and weights, or annotated data for evaluation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to proprietary reasons, we cannot disclose information related to the industry taxonomoy, CUSIP variables and weights, or annotated data for evaluation. However, we provide details that can be followed by other researchers or practitioners.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: Due to proprietary reasons, we cannot disclose information related to annotated data splits for evaluation. However, we have included information about parameter tuning.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While we performed a benchmark across different sentence embedding models for parameter tuning, we consider their results as comparable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Due to propietary reasons, we cannot disclose information on the hardware use for the experimentation. However, we note that we do not use GPUs for our pipeline.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the Introduction, we noted the motivation and impact of about work on illiquid bond pricing, and its relevance for financial applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks as we are not releasing new models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cited the creators or original owners of assets in our paper: embedding models and LLMs.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects. The experts that have contributed in our work are co-authors of the paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We incorporated LLMs as part of our proposed pipeline. This is described in the Methodology section.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.