# Identifying Financial Risk Information with Contrastive Reasoning

**Ali Elahi***
Department of Computer Science
University of Illinois Chicago; Surlamer Investments
Chicago, IL
aelahi6@UIC.edu

## Abstract

In specialized domains, humans often compare new problems against similar examples, highlight nuances, and draw conclusions rather than analyzing information in isolation. When applying reasoning in specialized contexts with LLMs on top of a RAG, the pipeline can capture contextually relevant information, but it is not designed to retrieve comparable cases or related problems.

While retrieval augmentation is effective at extracting factual information, its outputs in specialized reasoning tasks often remain generic, reflecting broad facts rather than context-specific insights. In finance, it results in generic risks that are true for the majority of companies. To address this limitation, we propose a peer-aware comparative inference layer.

Our contrastive approach outperforms the baseline RAG in text generation metrics, such as ROUGE and BERTScore, in comparison with human-generated equity research and risk.

## 1 Introduction

Retrieval-based systems access external documents to provide relevant information for downstream tasks [2]. Retrieval-Augmented Generation (RAG) combines a similarity-based retrieval with a language model inference layer to refine and synthesize retrieved content, providing a robust framework for summarization, question answering, and information extraction. Limitation arises from RAG's performance in retrieving critical information in specialized domains such as finance, particularly when extracting company risk factors from lengthy filings. A contextually rich passage detected by cosine similarity can still be irrelevant or uninformative, while a less rich section can convey more important details. In other words, the salience of information does not necessarily align with the semantic similarity index. We propose a contrastive approach: retrieving a broader set of relevant information for both the target firm and comparable peers, then prompting LLMs to generate a contrastive analysis that highlights nuances and distinctions.

### 1.1 Financial Risk

In financial equity research, risk identification refers to finding the key factors that could significantly affect a company's performance, whether by helping it grow or causing it to lose value. These factors shape how analysts, investors, and decision-makers judge the company's prospects and determine the value of its securities. A primary source for this information is the periodic reports public companies file with the U.S. Securities and Exchange Commission (SEC), including the annual document 10-K, the quarterly document 10-Q, and earnings call transcripts, records of quarterly calls where executives

*The research was done during Summer 2025 at Surlamer Investments.

discuss results, highlight challenges, and answer questions from analysts, often revealing insights not found in the written filings. These documents contain sections dedicated to describing the company's operations and financial condition.

For example, many filings include standard sections on topics like cybersecurity, regulatory compliance, or supply chain risk. These appear across most companies' reports, regardless of their actual relevance. A retrieval system or RAG model will reliably extract these sections simply because they are explicitly labeled, even if they are low-priority or non-differentiating. The challenge is to extract the relevant information and assess its relative importance in the context of the specific company and its peers in the same sub-sector.

## 2 Methodology

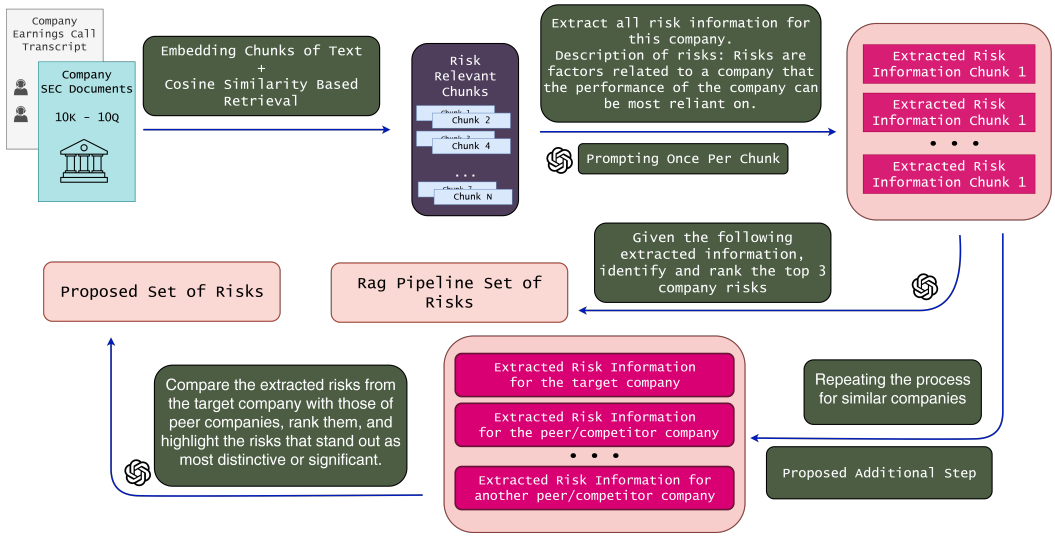


Figure 1: Baseline approach for risk identification based on RAG and summarization mythologies generates the "Final Set of Risks"; our proposed additional step will flag the most important items based on using a comparative approach and generate the "Proposed Set of Risks".

We propose an additional stage on top of RAG to identify a company's most significant and distinctive risks. First, the system extracts a broad set of potential risks from filings and earnings call transcripts. It then compares these with risk factors from peer companies in the same industry. By contrasting common versus unique or unusually emphasized risks, the system highlights those most critical in context.

Figure 1 summarizes the proposed risk identification pipeline. In the first stage, text chunks most relevant to the risk query (Appendix A – Risk Query) are extracted using cosine similarity of text embeddings. Each chunk is then processed by an LLM with a prompt (Appendix A – Risk Information Extraction Prompt) to retrieve relevant risk information. A subsequent prompt (Appendix A – Risk Aggregation Prompt) aggregates information across chunks, removes duplicates, and categorizes content into distinct risk topics. Finally, a contrastive prompt (Appendix A – Contrastive Risk Identification Prompt) compares the aggregated risk information for the target company and its peers to identify risks that are most specific or significant for the target company.

In the baseline approach, after the aggregation method, we prompt the LLM to generate the most important risks for the company.

### 2.1 Model sizes and Temperatures

We used a variety of OpenAI LLMs for our experiments. The described pipeline consists of three stages: information extraction, risk aggregation, and risk identification and ranking, the latter being the stage that differs between the baseline and contrastive approaches. The first two stages use

| | | BERTScore | | ROUGE-1 | | ROUGE-2 | | ROUGE-L | |
|---|---|---|---|---|---|---|---|---|---|
| **Models** | **Algorithms** | **Recall** | **F1** | **Recall** | **F1** | **Recall** | **F1** | **Recall** | **F1** |
| GPT 4o | Baseline | 0.1398 | 0.1417 | 0.1872 | 0.1889 | 0.0225 | 0.0239 | 0.0713 | 0.0702 |
| | Contrastive | **0.1401** | **0.1436** | **0.1904** | **0.1923** | **0.0231** | **0.0246** | **0.0719** | **0.0709** |
| GPT 4.1 | Baseline | 0.1504 | 0.1490 | 0.1941 | 0.1907 | 0.0228 | 0.0237 | 0.0718 | 0.0689 |
| | Contrastive | **0.1561** | **0.1548** | **0.2177** | **0.2025** | **0.0264** | **0.0256** | **0.0783** | **0.0705** |
| GPT O3 | Baseline | 0.1673 | 0.1570 | 0.1913 | 0.1913 | 0.0191 | 0.0201 | 0.0654 | 0.0644 |
| | Contrastive | **0.1692** | **0.1605** | **0.2116** | **0.2011** | **0.0210** | **0.0210** | **0.0702** | **0.0657** |

Table 1: BERTScore Score Recall and F1 Scores. Recall has a higher importance since we do not want to miss any true risks.

GPT-4.1-mini, the aggregation stage uses GPT-4.1, and for the final stage, we compare three models: O3 (a recent OpenAI reasoning model with reasoning level set to medium), 4o (an older OpenAI model released in late 2024), and 4.1 (OpenAI's most capable non-reasoning model).

### 2.2 Peers

Peers are companies in the same industry with similar business characteristics, such as scale (e.g., market capitalization), sector, products, or services, and whose stock returns tend to be correlated. In this work, we constructed our own peer sets for each company, though they can also be defined using methods such as GICS [5] sectors or Bloomberg's peers tool.

## 3 Experiments and Results

In our analysis, we generated risks for S&P 500 companies using the latest 10-Ks, 2025 10-Qs, and earnings call transcripts, which are all publicly available data. Relevant text chunks from these documents were extracted as input for the LLMs, and the output is a ranked list of risks.

### 3.1 Evaluation Methods

For the financial risk identification task, we evaluate performance using standard NLP summarization and text generation metrics, including ROUGE-1, ROUGE-2, ROUGE-L [3], and BERTScore [11]. Instead of using the input text as the reference, we compare the outputs against the risks and insights documented in human-written equity research reports. Using human-written risks, we can compare how accurate our risks are compared to actual risks that financial analysts extracted for a company. Due to confidentiality, the research reports used in this study will not be made publicly available upon publication.

### 3.2 Results

Table 3.1 reports the evaluation metrics for the baseline and contrastive approaches across the three OpenAI models used in the final inference layer. Across all metrics, the contrastive methodology consistently outperforms the baseline, demonstrating its ability to retrieve risks more aligned with human-written reports and investment theses. Additionally, comparing models, the O3 reasoning model achieves a higher BERTScore and ROUGE than both GPT-4o and GPT-4.1.

To better illustrate the results, we analyzed five groups of companies, each belonging to a specific sector or subsector, and examined the risk topics extracted by our pipeline. The results are presented in Table 2. As shown, the identified risks align closely with the business characteristics of each sector.

## 4 Literature Review

Traditionally, investment banks spent significant time extracting details from filings, earnings call transcripts, and corporate presentations before synthesizing them into reports. NLP methods dramatically accelerate this process, and various methodologies are proposed to identify risk and summarize financial information using NLP[6, 13, 4]. Beyond supporting investment decisions, the

| Sector/Industry | Identified Risks |
|---|---|
| Oilfield Services | Commodity Price & Customer Spending Cyclicality; Energy Transition & New Technology Execution; Contractual/Operational Execution Risk; Tariff, Trade & Regional Market Risk Sensitivity |
| Railing | Tariff, Trade & Regional Market Risk Sensitivity; Product & Innovation Execution, Raw Material Sourcing & Seasonality; Product Quality & Compliance; Competitive Pricing Pressure |
| Tech | Competitive Pricing Pressure; Technology Platform Execution Risk; Intellectual Property Risk; FX Sensitivity & Macro Dependence |
| Healthcare | Patent/Exclusivity & Product Concentration; Regulatory, Pricing & Market Access; Pipeline, R&D, M&A, and Integration Risk |
| Financial Services | FX Sensitivity & Macro Dependence; Actuarial/ Underwriting/Reserve Risk; Transition Finance & ESG Exposure; Investment Portfolio & Market Volatility |

Table 2: Risk titles identified for each industry/sector.

extracted and summarized information can also serve as input to downstream LLM pipelines, such as those designed to assess stock valuations given risk and other factors [12, 1].

Recent work has combined RAG and reasoning components for general or specialized tasks. C-RAG [7] introduces a contrastive framework in which retrieved documents are used to generate explanatory arguments, and teacher-model explanations serve as demonstrations for student models, blending retrieval with contrastive few-shot reasoning in general QA tasks. RAFT proposes a training scenario that strengthens in-domain RAG by teaching models to identify and cite relevant evidence while ignoring "distractor" documents, coupling retrieval with chain-of-thought reasoning to improve factual grounding in domains such as biomedicine [10]. RankRAG [9] integrates retrieval ranking and answer generation within a single instruction-tuned LLM, showing that incorporating ranking signals improves both context selection and downstream reasoning, often outperforming expert rankers and strong generation baselines across multiple benchmarks. Finally, RAG+ explicitly incorporates application-aware reasoning by retrieving not only factual knowledge but also aligned application examples, enabling structured, goal-directed inference; this design yields consistent improvements across domains such as mathematics, law, and medicine[8].

## 5 Limitations

One limitation of our pipeline is its lack of temporal awareness. The methodology does not account for the timeline of ongoing company events, meaning emerging developments are not captured, even though such events may materially affect risk exposure. As a result, the system may overlook dynamic changes that are crucial for timely and accurate risk assessment.

Another limitation lies in evaluation. Current assessments rely primarily on NLP-based text generation metrics, which measure surface-level similarity but fail to capture whether the extracted risks are truly aligned with a company's actual exposures. A stronger evaluation framework would incorporate expert human judgment as well as reasoning-based methods that go beyond lexical overlap to assess factual accuracy and contextual relevance.

## Acknowledgments

This research was conducted at Surlamer Investments. The work, including all findings and results presented in this paper, is the property of Surlamer Investments.

## 6 References

## References

[1] Ali Elahi and Fatemeh Taghvaei. Combining financial data and news articles for stock price movement prediction using large language models. In *2024 IEEE International Conference on*

*Big Data (BigData)*, pages 4875–4883, 2024.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[3] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Mahmoud Mahfouz, Armineh Nourbakhsh, and Sameena Shah. A framework for institutional risk identification using knowledge graphs and automated news profiling. *arXiv preprint arXiv:2109.09103*, 2021.

[5] MSCI. The global industry classification standard (gics®). `https://www.msci.com/indexes/index-resources/gics`. Accessed: 2025-08-31.

[6] Jiaxin Pei, Soumya Vadlamannati, Liang-Kang Huang, Daniel Preoţiuc-Pietro, and Xinyu Hua. Modeling and detecting company risks from news. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 63–72, 2024.

[7] Leonardo Ranaldi, Marco Valentino, and André Freitas. Eliciting critical reasoning in retrieval-augmented language models via contrastive explanations. *arXiv preprint arXiv:2410.22874*, 2024.

[8] Yu Wang, Shiwan Zhao, Zhihu Wang, Ming Fan, Yubo Zhang, Xicheng Zhang, Zhengfan Wang, Heyuan Huang, and Ting Liu. Rag+: Enhancing retrieval-augmented generation with application-aware reasoning. *arXiv preprint arXiv:2506.11555*, 2025.

[9] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

[10] Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.

[11] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[12] Tianjiao Zhao, Jingrao Lyu, Stokes Jones, Harrison Garber, Stefano Pasquali, and Dhagash Mehta. Alphaagents: Large language model based multi-agents for equity portfolio constructions, 2025.

[13] Tianyu Zhou, Pinqiao Wang, Yilin Wu, and Hongyang Yang. Finrobot: Ai agent for equity research and valuation with large language models, 2024.

# A Queries and Prompts

**Risk Query**

Major Risks of this Company:
Risks are factors related to a company that the performance of the company can be most reliant on; factors that can determine the performance eg: Strategic Supplier Dependence, Tariff and Trade Policy Sensitivity, Customer & Revenue Concentration, Geographical Concentration, Geopolitical/Regional Exposure, Energy Transition & Technology Investment, Supply Chain Fragility, Cybersecurity and Digital Risk, Capital Allocation / Financial Structuring, Regulatory/Legal Complexity, Human Capital & Succession, Macroeconomics/currency correlation, number of the suppliers, regions of activities, Any specific customer, Any specific products, informative actions and events that can cause change in the stock price or future of the company.

**Risk Information Extraction Prompt**

You are an expert-level equity analyst with deep expertise in `industry`. I am a hedge fund portfolio manager retrieving information for an investment committee meeting. You will be given a section of a 10-K/10-Q file, Earning Call Transcripts, or Analyst Reports, and you should retrieve related information about a given query.

Company Info : `name`, Ticker: `ticker`, Industry: `industry`

Data: `cosine similarity extracted text chunk`

Task: List the Major Risks of this Company.
Description of risks: Risks are factors related to a company that the performance of the company can be most reliant on; factors that can determine the performance eg: Strategic Supplier Dependence, Tariff and Trade Policy Sensitivity, Customer & Revenue Concentration, Geographical Concentration, Geopolitical/Regional Exposure, Energy Transition & Technology Investment, Supply Chain Fragility, Cybersecurity and Digital Risk, Capital Allocation / Financial Structuring, Regulatory/Legal Complexity, Human Capital & Succession, Macroeconomics/currency correlation, number of the suppliers, regions of activities, Any specific customer, Any specific products, informative actions and events that can cause change in the stock price or future of the company.

Return a list of key phrases as specific intrinsic risks (not general and market risks) and explain why they can trigger and cause a risk. Or why is it informative?

Do not generate any information that is not included in the given text. Do not use prior knowledge; only extract / retrieve and structure relevant information. Avoid any additional descriptions. State the most relevant knowledge from the text based on the given question. Avoid generic and general answers, and too broad answers. Be specific about the company and the industry. Report any information that can be useful from an investment perspective within the given query scope.

**Aggregation Prompt**

You are an expert-level equity analyst with deep expertise in `industry`. I am a hedge fund portfolio manager retrieving information for an investment committee meeting. Your inputs include summaries of SEC filings, analysts' reports, and earnings call transcripts. You need to answer a question or provide information about the given query based on the given summaries and retrieved information. You should select and aggregate relevant and trustworthy answers and construct a well-rounded analysis on the given query. The length of the answer should depend on what was asked.

Company Info : name, Ticker: `ticker`, Industry: `industry`

Question: `question`

Analysts answers: `data`

The given data is retrieved from different sources of information about the given query. Try to select information from various sources and do not rely only on each of the sources (SEC filings, analysts' reports, and earnings call transcripts). State the sources as much as possible. Take into account the industry of the company and see broadly, and do not give generic and general answers. Try to connect the information together to come up with new observations about the question that can be informative for investment purposes.

**Contrasticve Risk Identification Prompt**

Here is a list of major risks for `sub-sector` companies. Your task is to generate a risk summary for `target_company_name` (`target_company_ticker`), using comparative insights from the other companies in the same industry:
You are receiving information for all companies simultaneously, so you should identify risks specific to `target_ticker` in contrast to the others. Avoid generic or universally applicable risks; instead, highlight how such risks manifest uniquely for `target_ticker`.
Here is the risk information:

`target_company_name` (`target_company_ticker`):
`Risks for target company`

`Peer Company 1` (`peer_company_ticker`):
`Risks for peer company`

`Other Peers Information`

Only give the risks for the company `target_company_name` (`target_company_ticker`). Focus on company-specific, non-generalized/generic insights. Choose the most 3–5 important risks that drive the company's performance. Your tone should be technical but smooth, including valid reasons and arguments. Also include sources of information and numerical backup only if available and necessary.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Mentioned in the Introduction.

   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: Limitation section.

   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: No theorems were mentioned that require proof.

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: The code will not be released with the publication due to corporate restrictions. However, the paper details the prompts and methodology, and the dataset used is publicly available.

   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: The code will not be released with the publication due to corporate restrictions. However, the paper details the prompts and methodology, and the dataset used is publicly available.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The code will not be released with the publication due to corporate restrictions. However, the paper details the prompts and methodology (as well as model selection and model temperatures), and the dataset used is publicly available.

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: [Yes]

Justification: The generations were done three times, and the reported metrics are the mean of the results.

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: We used OpenAI APIs, so there is no information on resources. The time elapsed for text generation was not mentioned.

   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:

   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: [No]

    Justification:

    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:

    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: [NA]

    Justification:

    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:

    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:

    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:

    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: [NA]

    Justification:

    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.