TempPerturb-Eval: On the Joint Effects of Internal Temperature and External Perturbations in RAG Robustness

Yongxin Zhou, Philippe Mulhem, Didier Schwab
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000, Grenoble, France
{yongxin.zhou, philippe.mulhem, didier.schwab}@univ-grenoble-alpes.fr

Abstract

The evaluation of Retrieval-Augmented Generation (RAG) systems typically examines retrieval quality and generation parameters like temperature in isolation, overlooking their interaction. This work presents a systematic investigation of how text perturbations (simulating noisy retrieval) interact with temperature settings across multiple LLM runs. We propose a comprehensive RAG Perturbation-Temperature Analysis Framework that subjects retrieved documents to three distinct perturbation types across varying temperature settings. Through extensive experiments on HotpotQA with both open-source and proprietary LLMs, we demonstrate that performance degradation follows distinct patterns: high-temperature settings consistently amplify vulnerability to perturbations, while certain perturbation types exhibit non-linear sensitivity across the temperature range. Our work yields three key contributions: (1) a diagnostic benchmark for assessing RAG robustness, (2) an analytical framework for quantifying perturbation-temperature interactions, and (3) practical guidelines for model selection and parameter tuning under noisy retrieval conditions.

Keywords: Retrieval-Augmented Generation (RAG), Temperature, Perturbation Analysis

1. Introduction

Retrieval Augmented Generation (RAG) (Lewis et al., 2020) is a prompt engineering strategy that augments the internal capacity of Large Language Models (LLMs) with external knowledge. In a RAG, incorrect retrieved documents can introduce external noise that affects output quality (Fang et al., 2024; Wang et al., 2025; Kang et al., 2025). One RAG output also depends on the hyperparameters of its LLM, e.g., the *temperature*: the generated text is more (resp. less) deterministic for low (resp. large) temperature values (Holtzman et al., 2020).

Furthermore, *Perturbations* serve as adversarial examples in evaluating RAG robustness, simulating small input changes that can deceive models into incorrect predictions. These modifications help quantify how much specific input features must change to alter model outcomes (Anand et al., 2022). Previous research has employed various perturbation strategies for RAG question-answering systems, such as the *leave-one-token-out* approach that systematically removes individual sentences from input texts (Sudhi et al., 2024).

However, existing literature overlooks a critical dimension: the interaction between perturbations and generation hyperparameters, particularly temperature. This gap is significant given that temperature substantially affects output quality across various tasks (Renze, 2024; Du et al., 2025; Li et al., 2025), with low temperature values not always constituting the optimal choice. Consequently, current perturbation-based evaluations may yield misleading robustness assessments by failing to account for temperature variability.

Our work addresses this limitation by systematically investigating how perturbations interact with temperature settings in RAG systems. While traditional evaluations examine retrieval quality and generation parameters in isolation, they overlook their practical interdependence. By integrating both dimensions, our framework provides more reliable faithfulness explanations and accurate robustness measurements under realistic deployment conditions.

We approach the RAG LLM as a black box and experimentally quantify the *temperature effect* using the HotpotQA dataset (Yang et al., 2018), a benchmark for multi-hop question answering that demands complex reasoning and yields explanation-rich answers. We pair this dataset with systematic text perturbations to simulate noisy retrieval. To our knowledge, this represents the first comprehensive study of temperature-perturbation interactions in RAG systems[1]. As shown in Figure 1, our work systematically investigates this complex interplay, with three key contributions:

- A comprehensive taxonomy of perturbations for RAG, synthesizing and categorizing methods from information retrieval literature.
- A publicly released diagnostic benchmark that quantifies RAG robustness across 440 experimental conditions, spanning multiple models, temperatures, perturbation types, and question types.

[1]The source code and experimental framework are available at https://github.com/yongxin2020/TempPerturb-Eval, and the complete dataset of model inputs and generations is released on Hugging Face at https://huggingface.co/datasets/yongxin2020/TempPerturb-Eval-data.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 117–127
11–16 May 2026.

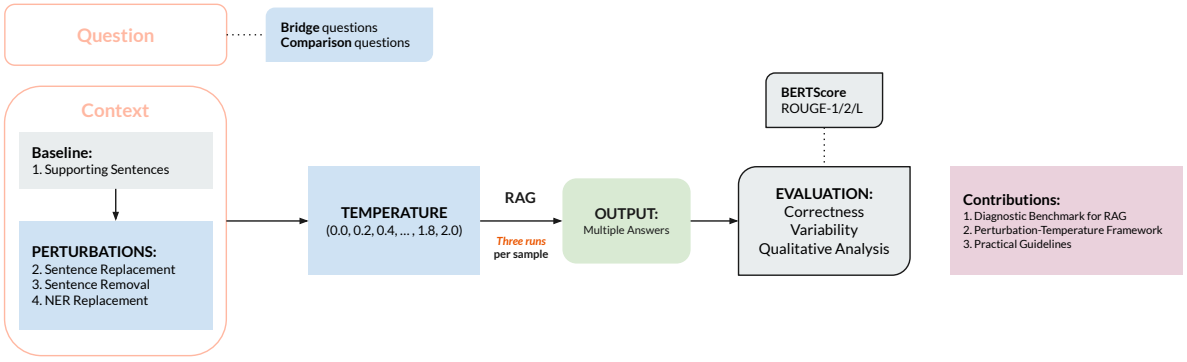

Figure 1: RAG Perturbation-Temperature Analysis Framework. The methodology stresses system robustness along two axes: external context perturbations (replacement, removal, NER substitution) and internal LLM temperature variation. The evaluation measures correctness and output variability across these conditions to establish a benchmark and derive practical guidelines.

- An analytical framework that models the joint impact of temperature and perturbations, accompanied by practical guidelines for robust model deployment.

2. Related Work

2.1. Perturbations in IR and RAG

Previous work have used perturbations as adversarial examples to examine the robustness of Information Retrieval (IR) models (Raval and Verma, 2020; Wu et al., 2023; Liu et al., 2024). Typical perturbations include removing, adding, or replacing words, phrases, sentences, passages, or entire documents. For instance, Raval and Verma (2020) found that even minimal token changes (1-3), an *attacker* can produce semantically similar perturbed documents capable of fooling document rankers.

Perturbations are also used in counterfactual explanations in explainable IR, where the closest samples on which the model makes a different prediction serve as example-based explanations (Poché et al., 2023). Some demonstrations (Rorseth et al., 2023, 2024) provide examples of explanations for RAGs obtained using different perturbation methods, but they do not evaluate their proposals. In the RAG context specifically, only the *leave-one-token-out* strategy has been evaluated on a question-answering task (Sudhi et al., 2024).

We review the relevant literature to provide an overview of perturbation methods from explainable information retrieval that can assess the robustness of IR and RAG systems (Zhou et al., 2025a). Table 1 summarizes these approaches, organized according to the following dimensions:

- **Target**: The IR component being perturbed (document or query).
- **Perturbation Category**: The high-level strategy for modifying content (e.g., subset selection, addition, replacement).
- **Specific Method**: The concrete technique used to implement the perturbation.
- **Granularity**: The textual unit affected by the perturbation (e.g., token, sentence, passage).
- **Description and Application**: An explanation of the method and its use cases.

The retrieval perturbations summarized in Table 1 introduce controlled deviations from an ideal retrieval result. The most direct way to link these perturbations to Information Retrieval (IR) performance is through document- or passage-level modifications. For example, removing a relevant document from the ranked list demonstrably lowers standard evaluation metrics such as precision, recall, and nDCG. A similar effect is observed when swapping the positions of a relevant and a non-relevant passage.

In contrast, perturbations at a finer granularity, such as synonym replacement, present a more complex scenario, as they may not inherently alter a document's underlying relevance. Other token-level perturbations, including random noise injection or entity replacement, are even more challenging to evaluate using classical IR metrics, which assume relevance judgments are based on unmodified text. To address this, the relevance of a perturbed document can be estimated by computing its similarity to the original version and applying a threshold, providing a pseudo-relevance score. Alternatively, LLM-as-a-judge methodologies (Gu et al., 2025) offer a flexible approach. These techniques effectively augment original relevance assessments, analogous to data augmentation strategies used in computer vision (Szegedy et al., 2014), enabling a more comprehensive evaluation of robustness under textual variations.

Target	Category	Specific Method	Granularity	Description and Application
Document	Subset	Removal	Sentence-level	Identifies a minimal subset of sentences whose removal lowers the document's rank beyond a threshold *k*. *Application:* Document ranking (Rorseth et al., 2023), QA (Sudhi et al., 2024)
		Combination	Sentence/Passage-level	Identifies how combinations of elements influence results, often via fixed-size random sampling. *Application:* Open-book QA (Rorseth et al., 2024)
	Permutation	Source Reordering	Passage-level	Identifies the effect of source order by finding permutations that place relevant sources in high-attention positions. *Application:* Open-book QA (Rorseth et al., 2024)
		Word Reordering	Word-level	Alters the sequence of words within each source of the input text. *Application:* Discussed in QA context (Sudhi et al., 2024)
	Replacement	Unit Replacement	Sentence/Passage-level	Replaces one sentence or passage at a time. *Application:* Document ranking (Goren et al., 2020)
		Entity Replacement	Word-level	Identifies entities (nouns, proper nouns) and replaces them with random words. *Application:* Discussed in QA context (Sudhi et al., 2024)
		Antonym Replacement	Word-level	Replaces one or more words with their antonyms. *Application:* Discussed in QA context (Sudhi et al., 2024)
		Synonym Replacement	Word-level	Replaces one or more (important) words with their synonyms. *Application:* QA (Sudhi et al., 2024), Document ranking (Wu et al., 2023)
	Injection	Random Noise	Word-level	Inserts random words into or around the source text. *Application:* Discussed in QA context (Sudhi et al., 2024), Passage ranking (Raval and Verma, 2020)
Query	Addition	Prefix Injection	Token-level	Insertion of a short prefix to the prompt leads to generation of factually incorrect outputs. *Application:* QA (Hu et al., 2024)
		Term Augmentation	Token-level	Minimal perturbations to a search query that raise the rank of a given document. *Application:* Document ranking (Rorseth et al., 2023)

Table 1: Taxonomy of perturbation methods for evaluating Information Retrieval (IR) and Retrieval-Augmented Generation (RAG) systems.

2.2. LLM Temperature Impact

LLMs generate token sequences using token logits l_k for each token v_k. The temperature modifies the output probabilities of the tokens so that the distribution peaks (resp. is flat) for large (resp. low) temperature values. Then it also influences the sampling of these tokens and therefore the whole generation. High-temperature values are supposed to add diversity to generation: several *runs* of the same prompt may generate very different responses. With the notation of Renze (2024), the probability of v_k, using the temperature hyperparameter T, is:

$$p(v_k) = \frac{e^{l_k/T}}{\sum_i e^{l_i/T}} \quad (1)$$

OpenAI and *DeepSeek* API documentations provide temperature recommendations for several tasks without documented support for these values. However, Renze (2024) found that varying sampling temperature from 0.0 to 1.0 does not yield statistically significant differences in problem-solving performance on multiple-choice question answering (MCQA) tasks across several LLMs. Building on this insight, we extend the analysis to RAG systems by investigating how perturbations interact with temperature.

3. Methodology

We seek to estimate the impact, if any, of the LLM temperature hyperparameter when perturbing a RAG LLM input. We cope with the internal variability coming from the LLM by presenting the same prompt (perturbed or non-perturbed) several times. Using this, our methodology assesses the behavior of the perturbations along the temperature evolution. We compare each generated text by the LLM with a processed ground-truth (see Section 4.3) using classical semantic similarity measures and compute the mean, variance and standard deviation for the same prompt. We then build graphics that present these comparisons.

This analysis allows us to determine: (i) whether certain perturbation types consistently degrade performance across all temperature values; (ii) whether the effect of specific perturbations at certain temperatures is statistically indistinguishable from the non-perturbed baseline; and (iii) whether the relative impact of different perturba-

tions changes with temperature (e.g., if *Perturbation A* has more impact than *Perturbation B* at a low temperature, but less impact at a high temperature).

4. Experiments

4.1. Dataset and Perturbations

We selected for our experiments the HotpotQA (Yang et al., 2018) dataset, dedicated to question-answering (QA) systems that perform complex reasoning and provide explanations for their answers. It contains 113k Wikipedia-based QA pairs[2]. This dataset was selected for three key characteristics: (i) sentence-level supporting facts that facilitate clean baseline establishment; (ii) multi-hop QA structure requiring reasoning across multiple documents, making it well-suited for perturbation testing; and (iii) availability of ground-truth answers for each query.

Furthermore, the classification of queries into "bridge" and "comparison" types enables the investigation of system behavior across distinct reasoning categories. *Bridge* questions are those where, to arrive at the answer, one must first identify a bridge entity and then find the answer in relation to it. The other type of multi-hop questions consists of *Comparison* questions, which require comparing two entities from the same category. A subset of these comparison questions are yes/no questions.

More precisely, we utilized the training set of the *fullwiki* version of the HotpotQA dataset. After analyzing the statistics of the data set, we randomly selected 100 samples for each category of facts (2, 3 and 4 facts) and for each type of question ("bridge" and "comparison"), resulting in a total of 600 samples for experimentation[3].

In our experiments, we establish a baseline using all original supporting sentences. Building upon this baseline, we systematically introduce three types of perturbations, selected for their relevance to real-world retrieval errors and alignment with established evaluation frameworks such as RAG-Ex (Sudhi et al., 2024). Our perturbation strategy includes[4]: (1) **Sentence Replacement:** replacing the latter portion of supporting sentences with irrelevant sentences from the same title, which simulates retrieval of correct entities with incorrect evidence, a common and realistic failure mode in QA systems; (2) **Sentence Removal:** deleting the latter half of supporting sentences; and (3) **NER Replacement:** masking named entities in the last supporting sentence(s) by replacing them with *[MASK]* tokens, focusing particularly on title-related entities to probe model sensitivity.

This procedure generated three perturbed input conditions in addition to the original baseline. The resulting setup allows for a controlled investigation of core perturbation effects against a stable reference point, establishing a reproducible framework for future robustness studies.

4.2. Models and RAG Configuration

We conducted experiments with five LLMs, categorized as follows:

- **Proprietary GPT Models[5]:** `gpt-3.5-turbo`, `gpt-4o`;
- **Open-Source LLaMA Models:** `Llama-3.1-8B-Instruct`[6] and `Llama-3.2-1B-Instruct`[7].
- **DeepSeek reasoning model**: `deepseek-reasoner`[8].

The chosen models (GPT-family, Llama-family, and deepseek-reasoner) offer a strategically diverse mix of architectural families, parameter scales, and commercial vs. open-weight availability, allowing us to evaluate robustness across different model types.

For each condition (model, temperature, perturbation type), we executed the same query three times to account for intrinsic stochasticity and to help distinguish the effect of the model's internal noise (due to temperature) from that of external perturbations. All other LLM hyperparameters were set to their default values, except for max_tokens, which is set to 1000.

4.3. Evaluation Methodology

Evaluation Metrics. While Exact Match (EM) and F1 are widely adopted metrics, their limitations in evaluating long-form generative outputs are well-documented. In RAG settings, models frequently

[2] https://huggingface.co/datasets/hotpotqa/hotpot_qa

[3] This sample size balances statistical reliability against computational constraints, given our fine-grained experimental design of 3 runs per (model, temperature, perturbation, query) condition.

[4] For each sample, the number of altered sentences was scaled by fact count: one sentence for 2-fact samples, one (33%) for 3-fact, and two for 4-fact samples. All other supporting sentences remained unperturbed.

[5] https://platform.openai.com/docs/models

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, pretrained and fine-tuned text models in 8B sizes.

[7] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, pretrained and instruction-tuned generative models in 1B sizes.

[8] https://api-docs.deepseek.com/guides/reasoning_model. Before delivering the final answer, the model first generates a Chain of Thought (CoT) to enhance the accuracy of its responses.

produce elaborated answers containing correct core information alongside supplementary explanations. Consequently, EM scores may be artificially low despite semantic correctness, and these standard metrics often fail to capture subtle differences in perturbed answers. The F1 metric offers greater robustness by rewarding token-level overlap, but unlike binary or multiple-choice QA, real-world RAG systems generate free-form answers requiring more nuanced evaluation.

Therefore, we report similarity metrics, which better reflect nuanced changes than exact matching. For instance, BERTScore (Zhang et al., 2020) can detect minor perturbations, such as passive/active voice shifts or small rewrites that retain the same meaning, while being more sensitive to semantic alterations than token-based metrics.

Reference Answer Processing. The reference answers of HotpotQA are short, for example, "flew in space" for the *bridge* question type, and "Yes" or "No" for the *comparison* question type. Since our framework uses a similarity measure to assess the influence of temperature and perturbations on the output, we transform the reference information into sentence form by combining the original question and the short answer. This is achieved using `GPT-4o` (with default hyperparameters) as the backbone model. The model is prompted with a combination of the question and the candidate answer, using the following template:

Prompt for Complete Answer Generator

Question: `{question}`
Answer: `{answer}`
Generate a complete and coherent answer based on the given question and answer, being as brief as possible:

Because the generated output are lengthy, we extract only the first sentence — or the first two if they began with "Yes" or "No" — as reference answers for comparison.

Given that `GPT-4o`'s role was limited to the low-complexity task of formatting answers (e.g., converting "flew in space" to "Both X and Y are astronauts who flew in space"), we observed high formatting accuracy. To verify this, we manually inspected a subset of the generated answers, including the final reference answers used for evaluation. A spot-check of 20 samples revealed no hallucinations.

Metric Selection and Reporting. We evaluated semantic similarity using both BERTScore (Zhang et al., 2020) and ROUGE-1/2/L (Lin, 2004), two standard metrics for natural language generation (Zhou et al., 2025b). While all metrics exhibited consistent trends across experimental conditions, we selected BERTScore as our primary measure due to its stronger alignment with human judgment of semantic equivalence. We report BERTScore F1 scores computed using the default `RoBERTa-large` model (Liu et al., 2019) as the backbone.

5. Experimental Results and Analysis

5.1. Correctness Analysis

We analyze BERTScore trends across temperature settings for different models in Figure 2. For each experimental condition, we compute the mean and standard deviation of scores across three runs, then aggregate these values across all samples per condition.

Our results reveal distinct model-specific temperature sensitivity patterns. While `deepseek-reasoner` maintains nearly invariant performance across the temperature range, GPT models exhibit degradation beginning at $T = 1.4$. In contrast, Llama models demonstrate earlier performance deterioration at $T = 0.6$, though with a more gradual decline slope compared to GPT models' sharper descent.

Taking `gpt-4o` for example (third column graphics from Figure 2), its results reveal that different perturbation types exhibit varying sensitivity to temperature increases. *NER Replacement* induces minimal degradation at $T = 2.0$, whereas *Sentence Replacement* and *Sentence Removal* lead to more substantial performance loss. Notably, all perturbation types demonstrate amplified sensitivity compared to baseline conditions as temperature rises, suggesting that temperature acts as a performance degradation amplifier.

As temperature varies, we observe a shifting performance hierarchy. At lower temperatures ($T < 1.4$), GPT models achieve the highest BERTScore, followed by Llama models and then `deepseek-reasoner`. However, this ranking reverses at higher temperatures ($T = 2.0$), where `deepseek-reasoner` maintains consistent performance while GPT models degrade below the levels held by Llama models.

Question type (comparison vs. bridge) has minimal impact on temperature sensitivity, with both types exhibiting nearly identical degradation curves across all models and perturbations. This suggests that temperature effects are largely orthogonal to question complexity. However, absolute performance is consistently higher for bridge questions than for comparison questions across all models.

For deployment scenarios requiring temperature tuning, we recommend: (1) `deepseek-reasoner` for applications requiring consistent performance across diverse temperature settings; (2) GPT mod-

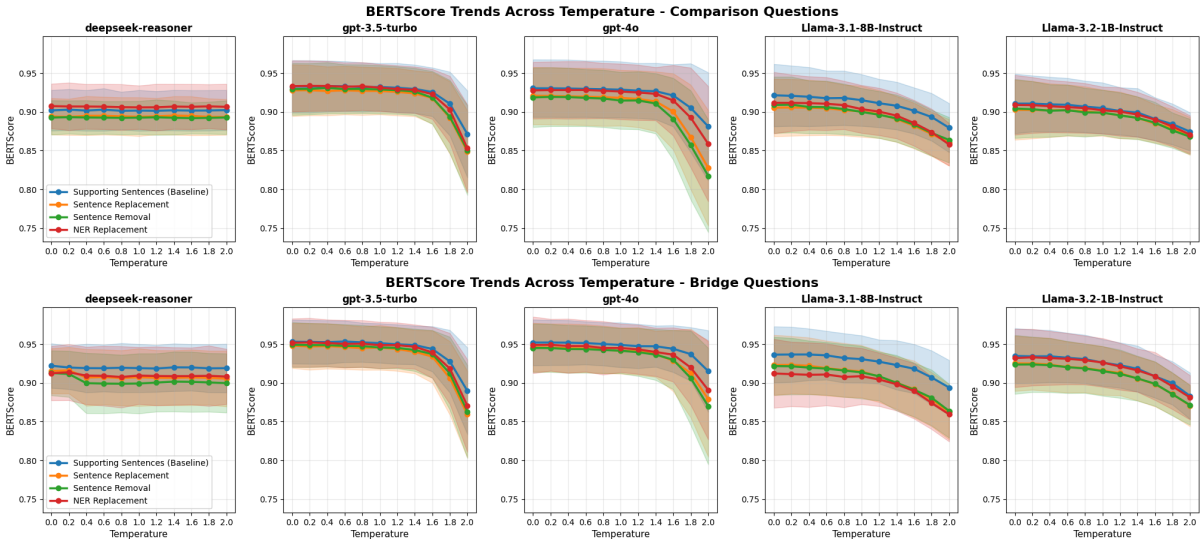

Figure 2: BERTScore trends across temperature variations for different models, comparing response types under perturbation. Solid lines represent mean scores across samples, while shaded areas denote $\pm$ standard deviation. The top row presents results for comparison questions; the bottom row presents results for bridge questions.

els with temperature ceilings of $T \leq 1.4$ to avoid sharp performance cliffs; and (3) Llama models with conservative temperature limits of $T \leq 0.6$ to maintain acceptable correctness levels.

5.2. Output Variability Analysis

To quantify performance sensitivity, we employ the Coefficient of Variation (CV), which measures relative variability by normalizing the standard deviation against the mean performance. Figure 3 visualizes these results, with gray dotted lines indicating each model's stability baseline, calculated as the average CV for the original (unperturbed) context across all temperatures.

Our analysis reveals that temperature exerts a stronger influence on output variability than perturbation types in most models. However, Llama models exhibit distinct behavior: `Llama-3.2-1B-Instruct` shows no noticeable variations for comparison questions and bridge questions, whereas `Llama-3.1-8B-Instruct` exhibits variation that depends on both the perturbation type and the temperature for bridge questions. GPT models demonstrate particularly high temperature sensitivity, with significant variability emerging at $T \geq 1.4$. In contrast, `deepseek-reasoner` and Llama models maintain more consistent performance across the temperature range. For the `deepseek-reasoner` model, *NER Replacement* perturbations have the greatest impact on comparison questions, while all three perturbation types impact bridge questions, with sensitivity emerging from $T \geq 0.2$.

The impact of question type on output variability is model- and perturbation-dependent. While bridge questions consistently show lowest variability in the unperturbed baseline across most models, comparison questions exhibit no consistent pattern: for instance, *Sentence Removal* yields the lowest CV for `deepseek-reasoner`, but not for other models. This heterogeneity highlights the complex interplay between question type, perturbation, and model architecture.

6. Qualitative Analysis of Model Sensitivity

To complement our quantitative findings, we conducted a qualitative analysis of model behavior under varying temperatures and input perturbations. We selected `gpt-4o` and `deepseek-reasoner` for this analysis based on their contrasting sensitivity profiles observed in previous experiments: with `gpt-4o` demonstrating higher temperature sensitivity and `deepseek-reasoner` showing greater stability. We examined model outputs at two temperature extremes: $T = 0.6$ (representing more deterministic generation) and $T = 2.0$ (producing more stochastic outputs).

6.1. BERTScore distributions

Figures 4 and 5 illustrate the BERTScore distributions for bridge-type questions under different perturbations. Temperature significantly impacts output quality, particularly for `gpt-4o`. At $T = 2.0$, performance degrades across all perturbations, with BERTScore values frequently falling between 0.70 and 0.80 and occasionally dropping below 0.70,

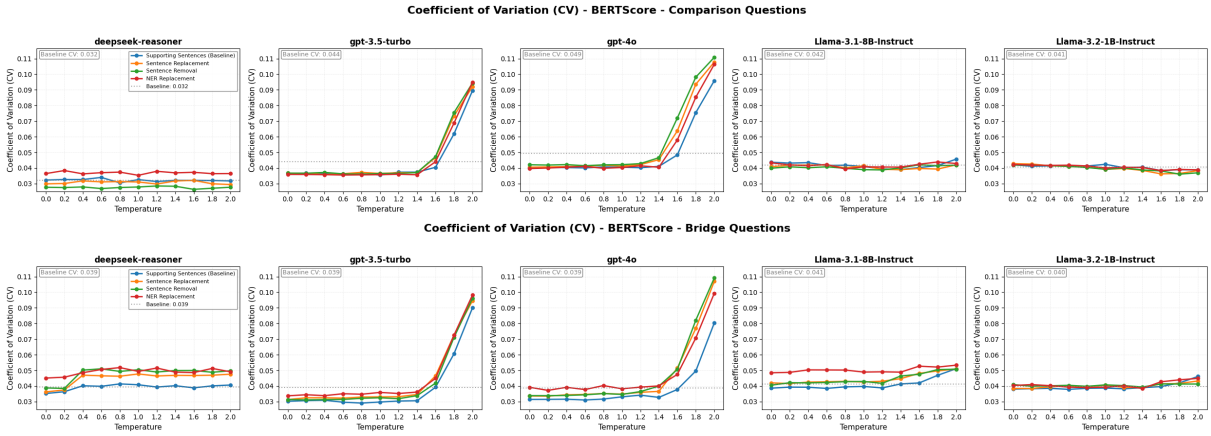

Figure 3: Coefficient of Variation (CV) for BERTScore across models, temperatures, and perturbation types. Each subplot displays CV trends for a model. The baseline CV value (average CV for the original, unperturbed context across all temperatures) is indicated in the top left of each subplot. The top row presents results for comparison questions; the bottom row presents results for bridge questions.

indicating increased output variability and reduced semantic faithfulness at higher temperatures.

In contrast, `deepseek-reasoner` exhibits stability across temperature settings. While $T = 2.0$ introduces slightly higher score variance, the median BERTScore remains consistent across temperatures for each perturbation type, suggesting more robust generation under temperature variation.

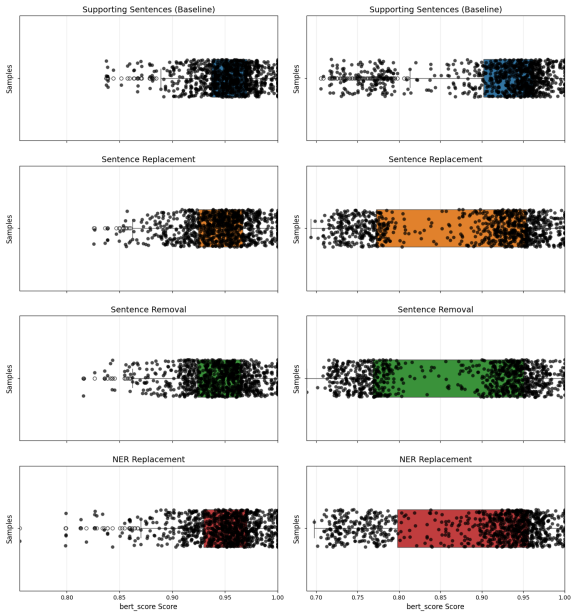

Figure 4: BERTScore distribution for `gpt-4o` on bridge questions across perturbation types at two temperatures (Left: $T = 0.6$, Right: $T = 2.0$). Each subplot shows a boxplot representing median, interquartile range, and whiskers, with individual sample scores (black dots) and outliers (white dots).

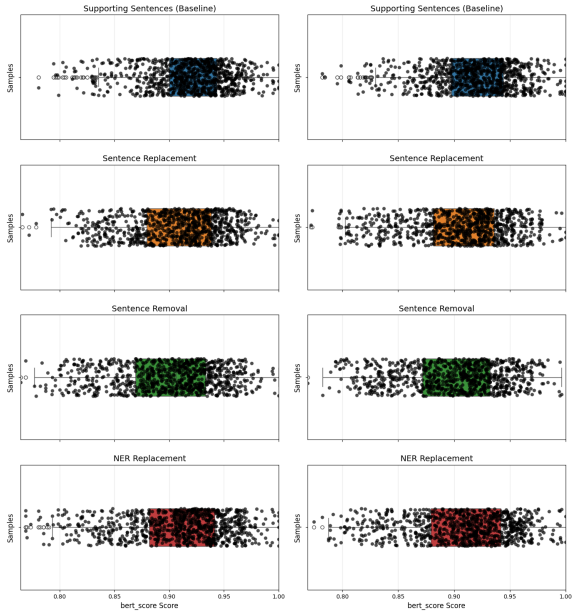

Figure 5: BERTScore distribution for `deepseek-reasoner` on bridge questions across perturbation types at two temperatures (Left: $T = 0.6$, Right: $T = 2.0$). Each subplot shows a boxplot representing the same elements as in Fig. 4.

6.2. Sample Analysis

To identify representative cases of model sensitivity, we selected, for each model studied in Section 6.1, temperature, question type, and perturbation type, the sample with the largest BERTScore gap between original and perturbed conditions. This method highlights key fragility patterns.

Perturbation-Type Analysis. Our examination of these cases reveals distinct failure modes across perturbation types. *Sentence Replacement*

and *Sentence Removal* perturbations frequently trigger model refusal behaviors, with responses such as "The retrieved document does not provide..." becoming common[9]. At higher temperatures ($T = 2.0$), these perturbations often result in garbled or nonsensical outputs containing code snippets, random tokens, and mixed languages. *NER Replacement* perturbations prove effective at disrupting model performance, causing failures in entity recognition and relationship inference that lead to incomplete or incorrect answers.

Temperature Effects on Output Quality. Temperature settings influence how models degrade under perturbation. At $T = 2.0$, we observe severe output degradation characterized by nonsense and complete failure to address the query. In contrast, at $T = 0.6$, models demonstrate greater robustness, though they still exhibit cautious response patterns (e.g., "I cannot determine..."), partial answers, and occasional factual errors. This suggests that while lower temperatures improve stability, they do not eliminate sensitivity to perturbations.

Question-Type Sensitivity. Bridge questions show particular sensitivity to entity removal or replacement, likely due to their reliance on connecting information across multiple facts. Comparison questions, while still affected, occasionally maintain correctness through external knowledge utilization, suggesting different reasoning pathways may exhibit varying robustness.

Model-Specific Degradation Patterns. The two models exhibit distinct failure characteristics. `gpt-4o` typically produces fluent but incorrect responses under perturbation, maintaining coherence while sacrificing accuracy. `deepseek-reasoner`, conversely, often fails more gracefully with concise but incomplete answers (e.g., responding with single words like "Brewery" rather than generating nonsensical text). This difference likely stems from their distinct training objectives; as a reasoning model, `deepseek-reasoner` may prioritize logical coherence and conciseness over the discursive fluency characteristic of a general-purpose model like `gpt-4o`, a hypothesis that merits further investigation.

Robustness Insights. Despite overall sensitivity patterns, we observe instances where models maintain correctness under perturbation, indicating some degree of inherent robustness or effective internal knowledge utilization. The significant performance variability across samples suggests that certain question structures or knowledge domains are inherently more fragile than others.

[9]For example, `gpt-4o` with *Sentence Replacement* at $T = 0.6$ output: "The retrieved document does not provide specific information about the campus sizes of Indiana University or Ohio State University to determine which has the third-largest university campus in the United States. To accurately answer the query, more detailed data on the campus sizes or student populations of both universities is required."

7. Discussion and Conclusion

This study investigated the relative impact of internal temperature versus external perturbations on RAG system performance. Our analysis reveals that temperature introduces a more pronounced influence on model correctness than specific perturbation types, with performance degrading significantly above certain temperature thresholds across most tested models and perturbation conditions.

The interaction between temperature and perturbations proves particularly critical: while models demonstrate relative robustness to perturbations at lower temperatures ($T \leq 0.6$), they exhibit severe performance degradation under the same perturbations at higher temperatures ($T \geq 1.4$). This joint effect creates a fragility landscape where systems that appear stable under standard evaluation conditions can fail dramatically when facing real-world noise combined with typical sampling strategies.

Notably, we observed instances where models maintained correctness despite substantial perturbations, suggesting utilization of internal knowledge rather than strict reliance on retrieved documents. However, the unpredictable nature of this phenomenon, where models sometimes bypass corrupted context entirely but other times produce confidently wrong responses, highlights the challenge of determining when and how internal knowledge mechanisms activate in RAG settings.

Our findings carry implications for RAG deployment. From a temperature perspective, we demonstrate that this hyperparameter must be carefully calibrated alongside perturbation robustness considerations. From a retrieval perspective, our results reinforce the importance of filtering uncertain or irrelevant content, aligning with principles in active retrieval methods like FLARE (Jiang et al., 2023). Based on our comprehensive evaluation, we propose the following deployment strategies: `deepseek-reasoner` for applications requiring consistent performance across diverse temperature settings; configure GPT models with a temperature ceiling of $T \leq 1.4$ to avoid sharp performance degradation; and employ Llama models with a conservative temperature limit of $T \leq 0.6$ to maintain acceptable correctness levels.

The main contribution of this work is to reveal the critical yet overlooked interaction between internal and external noise sources in RAG systems. A system that performs well on conventional benchmarks may prove surprisingly fragile when facing the combined effects of sampling stochasticity and real-world document perturbations. To address this

gap, we introduce a dedicated benchmark and analytical framework designed to quantify this joint effect. We note that the present study isolates the LLM generator's sensitivity by perturbing gold contexts, thereby controlling for retrieval noise. A future direction is to incorporate actual retrieval systems to examine how retrieval inaccuracies and generation sensitivity compound in end-to-end pipelines.

Limitations

Our empirical findings are currently based on the HotpotQA dataset, which is a multi-hop factoid QA benchmark. While this allows for controlled perturbation analysis, the generalizability of the observed temperature-perturbation interaction patterns to other task types, such as summarization or open-domain dialogue, remains to be validated and is an important direction for future work.

This study focused specifically on temperature as the key stochasticity parameter, holding other generation parameters (e.g., top-p, frequency penalty) at default values. We acknowledge that these parameters could interact with retrieval noise, and their joint effects constitute a promising avenue for extending this framework.

Acknowledgments

This work was partially funded by the "Intelligent Systems for Bridging Data, Knowledge and Humans" axis of the Grenoble Computer Science Laboratory (LIG). It was also conducted within the framework of the AugmentIA Chair, led by Didier Schwab and hosted by the Grenoble INP Foundation, thanks to the patronage of the Artelia Group. The chair also receives support from the French government, managed by the National Research Agency (ANR) under the France 2030 program with reference number ANR-23-IACL-0006 (MIAI Cluster). We also thank anonymous reviewers for their insightful comments.

8. Bibliographical References

Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable information retrieval: A survey.

Weihua Du, Yiming Yang, and Sean Welleck. 2025. Optimizing temperature for language models with multi-sample inference. In *Forty-second International Conference on Machine Learning*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. Association for Computational Linguistics.

Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2020. Ranking-incentivized quality preserving content modification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 259–268, New York, NY, USA. Association for Computing Machinery.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 1119–1130, New York, NY, USA. Association for Computing Machinery.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jeongwoo Kang, Markarit Vartampetian, Felix Herron, Yongxin Zhou, Diandra Fabre, and Gabriela Gonzalez-Saez. 2025. Getalp@automin 2025: Leveraging rag to answer questions based on meeting transcripts.

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. 2020. Retrieval-augmented generation

for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251. International Neural Network Society Workshop on Deep Learning Innovations and Applications 2025.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Robust neural information retrieval: An adversarial and out-of-distribution perspective.

Antonin Poché, Lucas Hervier, and Mohamed-Chafik Bakkay. 2023. Natural example-based explainability: A survey. In *Explainable Artificial Intelligence*, pages 24–47, Cham. Springer Nature Switzerland.

Nisarg Raval and Manisha Verma. 2020. One word at a time: adversarial attacks on retrieval models.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Joel Rorseth, Parke Godfrey, Lukasz Golab, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. 2023. Credence: Counterfactual explanations for document ranking. *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3631–3634.

Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. Rage against the machine: Retrieval-augmented llm explanations.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. Rag-ex: A generic framework for explaining retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2776–2780, New York, NY, USA. Association for Computing Machinery.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten De Rijke, Yixing Fan, and Xueqi Cheng. 2023. Prada: Practical black-box adversarial attacks against neural ranking models. *ACM Trans. Inf. Syst.*, 41(4).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yongxin Zhou, Philippe Mulhem, and Didier Schwab. 2025a. Explicabilité par Perturbations pour les Systèmes RAG. In *Actes de l'atelier Accès à l'information basé sur le dialogue et grands modèles de langage 2025 (DIAG-LLM)*, pages 1–6, Marseille, France. ATALA and ARIA.

Yongxin Zhou, Fabien Ringeval, and François Portet. 2025b. Can GPT models follow human summarization guidelines? a study for targeted communication goals. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 249–273, Hanoi, Vietnam. Association for Computational Linguistics.

9. Language Resource References

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Field	Content
Question	New Faces of 1952 is a musical revue with songs and comedy skits, it helped jump start the career of which young performer, and American actress?
Answer	Carol Lawrence
Supporting Facts	
New Faces of 1952 (sent 2)	It helped jump start the careers of several young performers including Paul Lynde, Alice Ghostley, Eartha Kitt, Robert Clary, Carol Lawrence, Ronny Graham, performer/writer Mel Brooks (as Melvin Brooks), and lyricist Sheldon Harnick.
Carol Lawrence (sent 0)	Carol Lawrence (born September 5, 1932) is an American actress, most often associated with musical theatre, but who has also appeared extensively on television.
Example Distractor Context	
Guess Who I Saw Today (sent 0)	"Guess Who I Saw Today" is a popular jazz song written by Murray Grand with lyrics by Elisse Boyd.
Guess Who I Saw Today (sent 1)	The song was originally composed for Leonard Sillman's Broadway musical revue "New Faces of 1952" in which it was sung by June Carroll.
Monotonous (sent 0)	"Monotonous" is a popular song written by June Carroll and Arthur Siegel for Leonard Sillman's Broadway revue "New Faces of 1952".
Monotonous (sent 1)	The song was written based on the experiences of its singer Eartha Kitt.

Table 2: A HotpotQA bridge question example (ID: 5a76a401554299373536012b) from our evaluation set, showing supporting facts and sample distractor context.

A. Dataset and Example

HotpotQA is distributed under a CC BY-SA 4.0 License. The dataset can be downloaded from: https://hotpotqa.github.io/. Table 2 presents an example from the HotpotQA dataset.

Our generated model outputs and experimental data are available on Hugging Face at: https://huggingface.co/datasets/yongxin2020/TempPerturb-Eval-data.