Could You Say That Again?

Hallucinatory Variance in Untuned Large Language Models

Dean Cahill
deancahill@brandeis.edu

Advisor: James Pustejovsky

Abstract

We propose Hallucinatory Variance, a novel metric for measuring the effect of repeated prompt contexts on pre-trained language models. By repetitively prompting untuned models to emulate a multi-turn encounter with an uninformed end user, we hope to measure the effect of extending a chat context on model hallucination rates and give insight into how seemingly imperceptible changes in the discursive grounding of the model context - asking the same question twice, contradicting the model, etc. - modulate the responses of Large Language Models to produce hallucinations. In order to measure this phenomenon, we apply BERTscore (Zhang* et al., 2020) across time and compare this value to an expected BERTscore derived from the overall model output. While our results are inconclusive, we hope future research can build on the mistakes and shortcomings of this project and work toward a more nuanced approach for evaluating hallucinations.

1. Introduction

Pre-trained Large Language Models (LLMs) have become a cultural force in recent years, owing to . Lending to their admittedly impressive performance across a variety of tasks, the effective mimicry of human speech, as well as a memetic proliferation of a variety of both LLM and other generative models in online spaces, LLMs have developed a status that precedes them; however, this cultural status should be measured against a careful understanding of the mechanics of these models and their limitations.

The fundamental problem of these generative architectures is that of the hallucination (Rawte et al., 2023). The vastness of these models' parameters, and the resulting size & complexity of the attention-space these parameters sit within, often produce output which is misleading, incorrect, or which produces entirely foreign information (Ji et al., 2023). The LLM capacity for hallucinations is treated with due caution within ML communities, and the concept has broken into the mainstream, but the nature of the problem - the why and how of hallucinations - is lost, meaning that the common use cases for popular interfaces fail to account for the fragility of knowledge, truth, and grounding in model output. As such, a deeper understanding of the hallucinatory capacity within the LLM's social context should be explored.

This paper aims to do just that: by repeatedly prompting LLMs, we hope to create experimental conditions that reflect a kind of "interactive search engine" usage that is common in end-user deployment of LLM chat interfaces. In doing so, we hope to show whether the social function of the language model - to speak and to speak well - contradicts the nature of its output in a way that affects the ability for said model to be consistently grounded within what is being asked of it.

2. Prior Work

A variety of metrics and benchmarks exist to measure hallucinations (Li et al., 2023, Dziri et al., 2022, Wang et al., 2023). Additionally, distance metrics such as EMD (Zhao et al., 2019) are commonly used in the mitigation of hallucination in image generation. BERTscore as a metric has also been explored thoroughly (Sun et al., 2022) across a breadth of domains including machine translation (Jauregi Unanue et al., 2021, Leiter, 2021)). The shortcomings of BERTscore (Hanna & Bojar, 2021, Sun et al., 2022) are significant, but functional enough for our purposes. Modern approaches for mitigating hallucination focus on retrieval augmentation as a way to leverage existing knowledge bases within LLMs. We intend to evaluate BERTscore on such a graph-retrieval model, but this is slightly beyond the scope of this paper.

3. Methodology

3.1 Dataset

We use Google's Natural Questions dataset (Kwiatkowski et al., 2019). This dataset consists of 300k+ Open-Domain questions, framed as Google queries, and Wikipedia pages which contain their answers. This dataset was chosen because the questions are reflective of the kinds of questions that would be asked of the LLM by an untrained end-user. Our initial study consists of 260 questions from this dataset, chosen sequentially from the load order of the NQ training set (to ensure ease of replication).

3.2 Models

3.2.1 Model Selection

To properly evaluate the use-case of interactive chat QA, we focus our attention on two "Designer" LLMs - GPT and Mixtral. These models not only represent the current state of the art regarding language modeling, but are also the primary public engagement with language models as a tool.

3.3 Prompting

By exploring repetition, we attempt to schematize particular contexts which reflect common use cases in our target domain (that domain being the meta-task of interactive end-user QA). As such, we've developed 4 fundamental types of prompt repetition: (1) Rote (basic) repetition, (2) Chain-of-Thought repetition, (3) Instructive repetition, and (4) Contradictive repetition.

In addition to these prompt types, we attempt to overlay a series of contexts. Each model is prompted (a) with simply the question, (b) with an additional QA context prepended, and (c) with a meta-context prepended, which describes the role of hallucinations in the QA task. These additional initial contexts are understood as the initial system prompt.

Rote/Basic

The simplest form of repetition, this consists of simply repeating the prompt over and over again. In order to capture this, we first send a prompt to the model, concatenate the decoder output to this prompt, concatenate the prompt to the end of this output, and repeat the process. This creates a sort of "folding over" in the prompting, which should reflect a growing context in a chat interface over multiple conversational turns:

```
# Repetition Zero =====================================|
=> SYSTEM: You are a helpful assistant
=> USER  : What is the color of the sky?
=> AGENT : The color of the sky is blue.
# Repetition One  =====================================|
=> USER  : What is the color of the sky?
=> AGENT : The color of the sky is blue.
# =====================================================|
```

Chain-of-Thought

Essentially the same as Rote repetition, but with the added constraint of chain-of-thought. This output is noticeably more verbose, which is factored into the evaluation metrics.

```
# Repetition Zero =====================================|
=> SYSTEM: You are a helpful assistant
=> USER  : What is the color of the sky? Explain each step of your reasoning.
=> AGENT : Sunlight is scattered by the particles in the air.
Since blue light is a shorter wavelength, more blue light is scattered.
This means that the color of the sky is blue.
# Repetition One  =====================================|
=> USER  : What is the color of the sky? Explain each step of your reasoning
=> AGENT : Blue. Water in the air scatters sunlight,
   leading to shorter blue wavelengths being spread through the air.
# =====================================================|
```

Instructive

Instructive prompting can be conceptualized as a form of "indirect conflict" with the model output. Instead of highlighting a specific error or correcting the model, we simply **ask** it to "repeat itself". This repetition is modulated by certain requests ("say that again, but {use synonyms, use more verbs, etc.}). This should reflect a misalignment in user understanding of a prompt, but trust in the model itself.

```
# Repetition Zero =====================================|
=> SYSTEM: You are a helpful assistant
=> USER  : What is the color of the sky?
=> AGENT : The color of the sky is blue.
# Repetition One  =====================================|
=> USER  : Could you say that again?
=> AGENT : Certainly! The color of the sky is blue.
# Repetition Two  =====================================|
=> USER  : Could you say that again, but with different nouns?
=> AGENT : Certainly! The hue of the air is blue.
# =====================================================|
```

Contradictive

As opposed to Instructive prompting, Contradictive prompting is based on direct conflict with the output, without directly articulating the location of the error. Essentially, this category consists of directly saying "This is incorrect". This highlights the converse of Instructive prompting - a distrust in the output reflected in a single point where the user understands a mistake has been made.

```
# Repetition Zero =====================================|
=> SYSTEM: You are a helpful assistant
=> USER  : What is the color of the sky?
=> AGENT : The color of the sky is blue.
# Repetition One  =====================================|
=> USER  : That is incorrect.
=> AGENT : I'm sorry, I was mistaken. The color of the sky is blue.
# Repetition Two  =====================================|
=> USER  : You're wrong.
=> AGENT : The color of the sky is blue.
# =====================================================|
```

3.4 Evaluation

3.4.1 Hallucinatory Variance

The core of these experiments lies in the aggregation of our evaluation metrics across prompt repetition. To measure this repetition, we propose Hallucinatory Variance (HV).

Hallucinatory variance can best be understood in relation to typical statistical variance, but with key differences. We define HV as the deviation from expectation of some other metric, *f*, as the size of the context increases. Hallucinatory variance is a measure across "turns", reflecting the impact of contextual extension on the model's tendency to hallucinate. HV, for a given model **M**, a conversational turn *t*, and a question *q*, can be defined as:

$$HV_f(t) = \frac{\mu(f(t))}{f(\boldsymbol{M})} - \mu(\frac{f(q)}{f(\boldsymbol{M})})$$

This definition is a function of the current turn position. As the number of turns in the conversational context increases, we expect the hallucination rate to increase. Thus, the hallucinatory variance of a given **model** for a given *metric* can be represented as the rate of the change in deviation from expected hallucination. Within our experiments, f is BERTscore, defined as a greedy matching of maximum pairwise cosine similarity of BERT embeddings between a reference (in our case, the "long answer" to the question) and the model output. We get the mean of this BERTscore empirically by averaging the BERTscores of each question's repetitions, and normalizing this by the total average BERTscore of all repetitions of all questions.

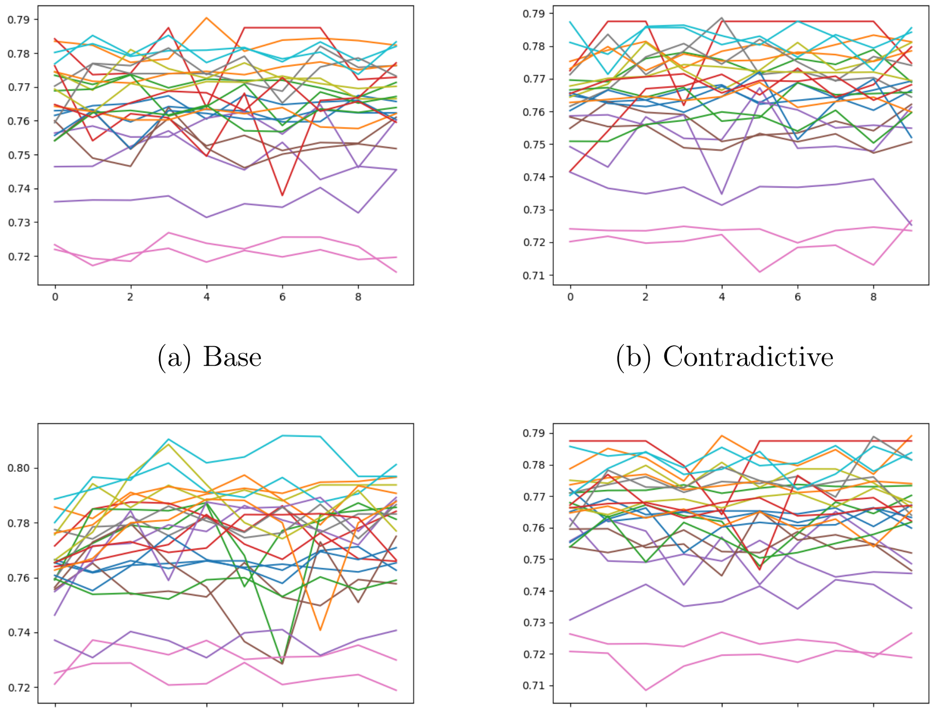
4. Results

We evaluate BERTscore and hallucinatory variance over 10 repetitions of each question in the dataset for both GPT3 and Mixtral. We refrain from increasing the number of repetitions too much, as this begins to drift from the phenomenon we're discussing (a person wouldn't ask the same thing **50** times). Limitations on runtime meant we couldn't perform very much parameter searching, so we elected to prompt the models with a relatively high temperature (0.8 for GPT, 0.6 for mixtral). We provide our results for the F1 BERTscore, as well as HV calculated over that F1 score. Within the BERTscore graphs, each line represents the trend of a different question. Figures for precision and recall, as well as for statistical variance (calculated on the same parameters as HV) are available in the repository housing the code for this paper.

4.1 GPT-3

Due to collection issues, the GPT model was evaluated over a collection of 60 questions x 10 repetitions. This is less statistically significant than our Mixtral results, but hopefully a sample of 600 reiterations is enough for these charts to be substantive.

4.1.1 BERTscore



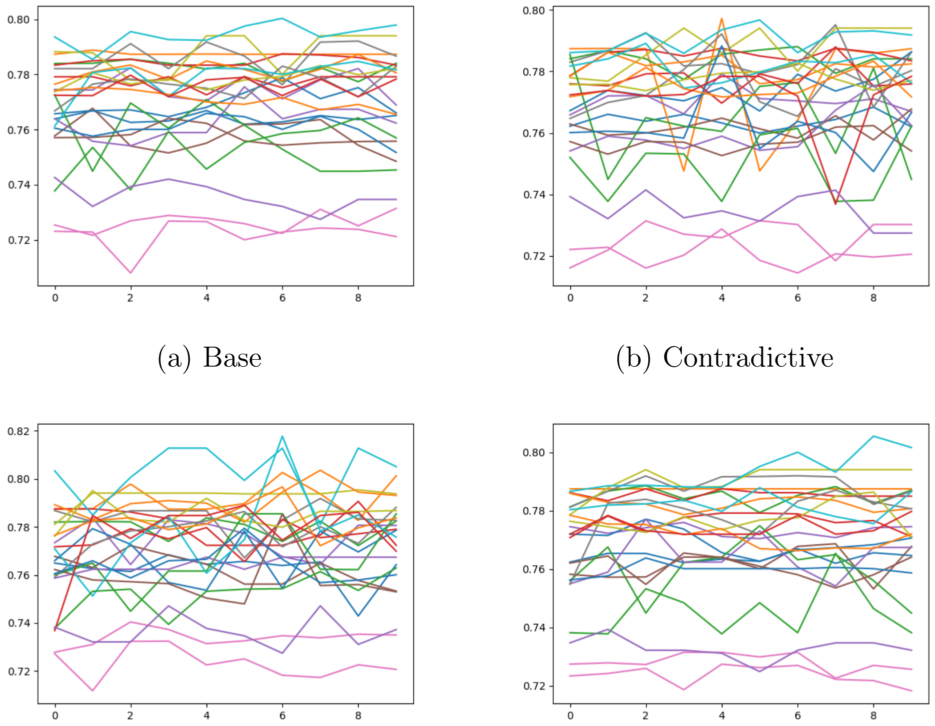
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 1: BERTscore - No initial context



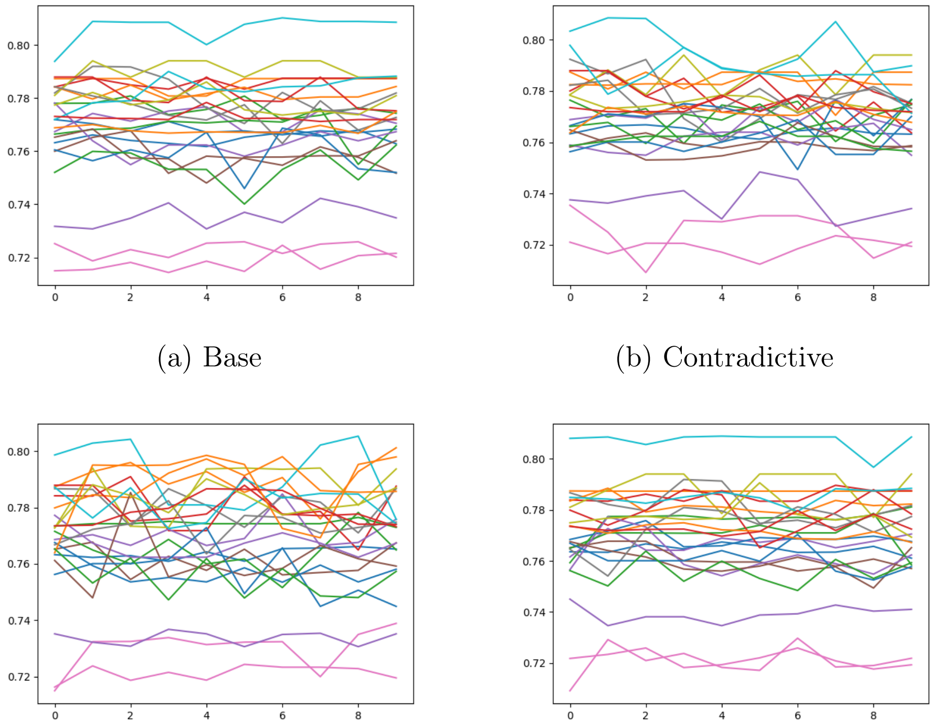
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 2: BERTscore - QA context



(a) Base

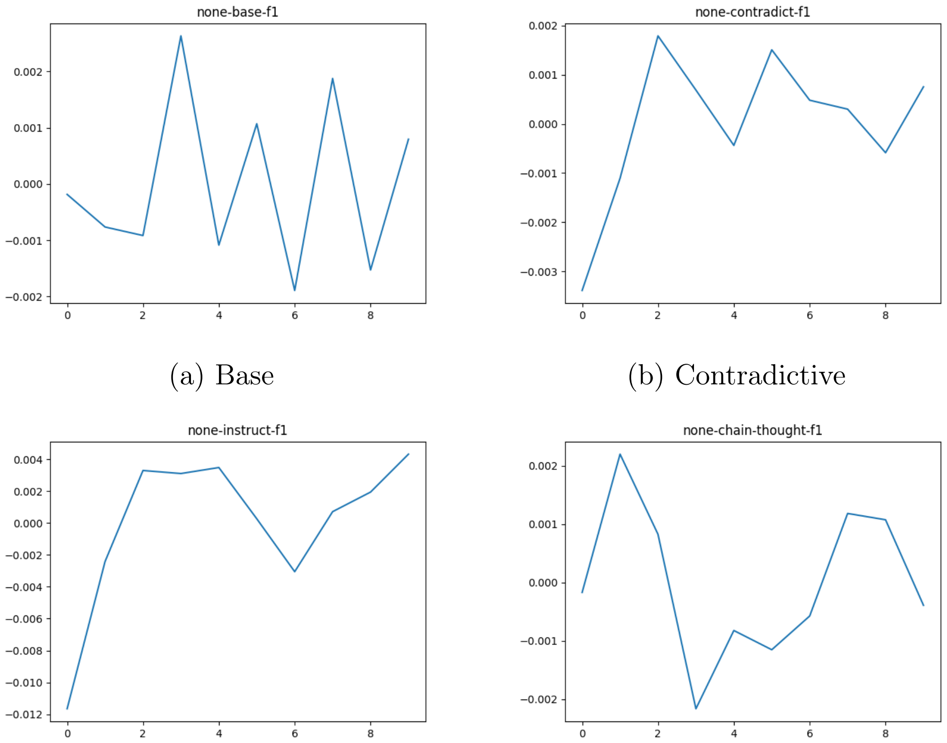
(b) Contradictive

(c) Instructive

(d) CoT

Figure 3: BERTscore - Hallucination initial context

4.1.2 HV



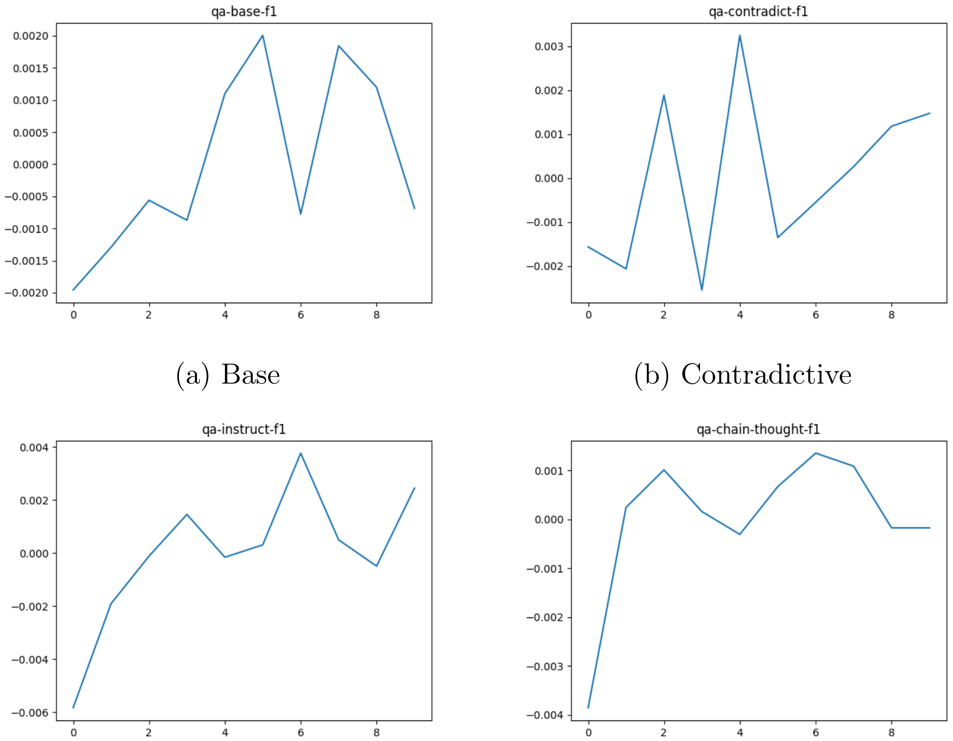
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 4: Hallucinatory Variance - No initial context



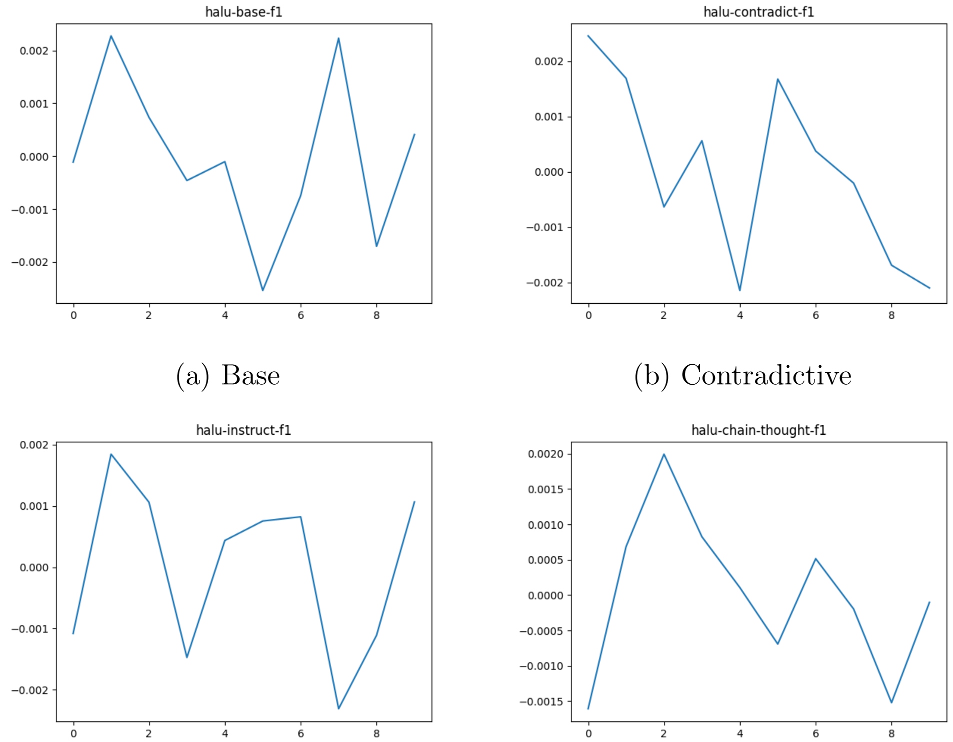
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 5: Hallucinatory Variance - QA context



(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

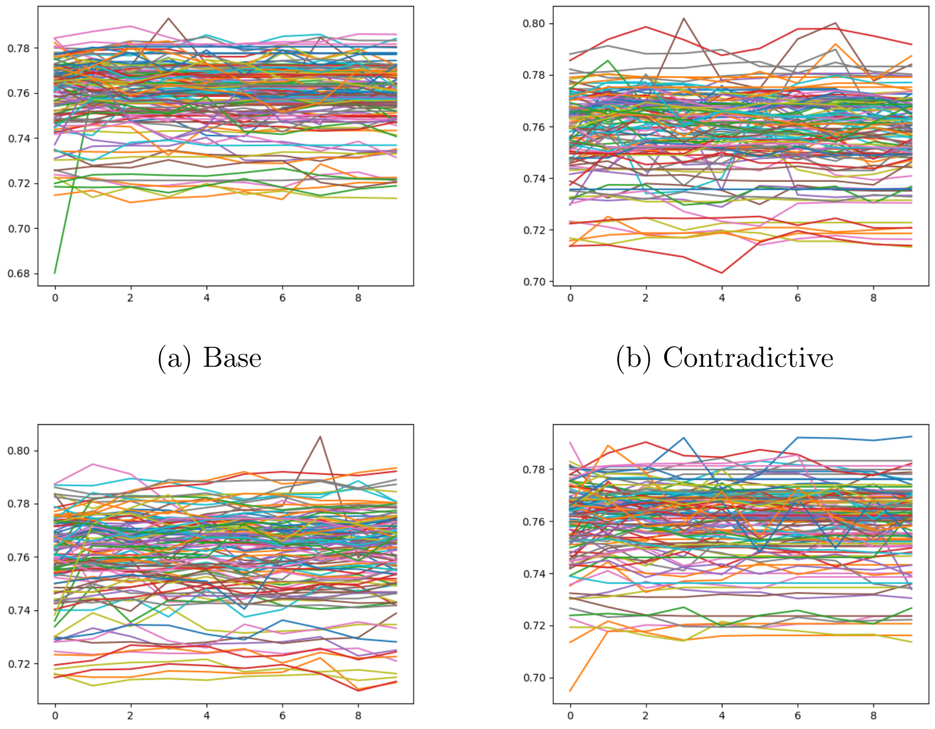
Figure 6: Hallucinatory Variance - Hallucination initial context

What immediately jumps out here is that most of the scores don't actually change that much - large drops in BERTscore across all 3 contexts are largely instance-oriented: the model get a low score on one of the repetitions, but immediately course-corrects. Moreover, the spread among the scores is pretty slim. The model gets between 0.70-0.80 average BERTscore along all experimental runs, implying that the extended context does not actually factor in very much.

4.2 Mixtral

The Mixtral dataset consists of 217 questions, again repeated 10 times. The BERTscore graphs get a little busy, this is to reflect how closely grouped the data is, and how little it changes over time.

4.2.1 BERTscore



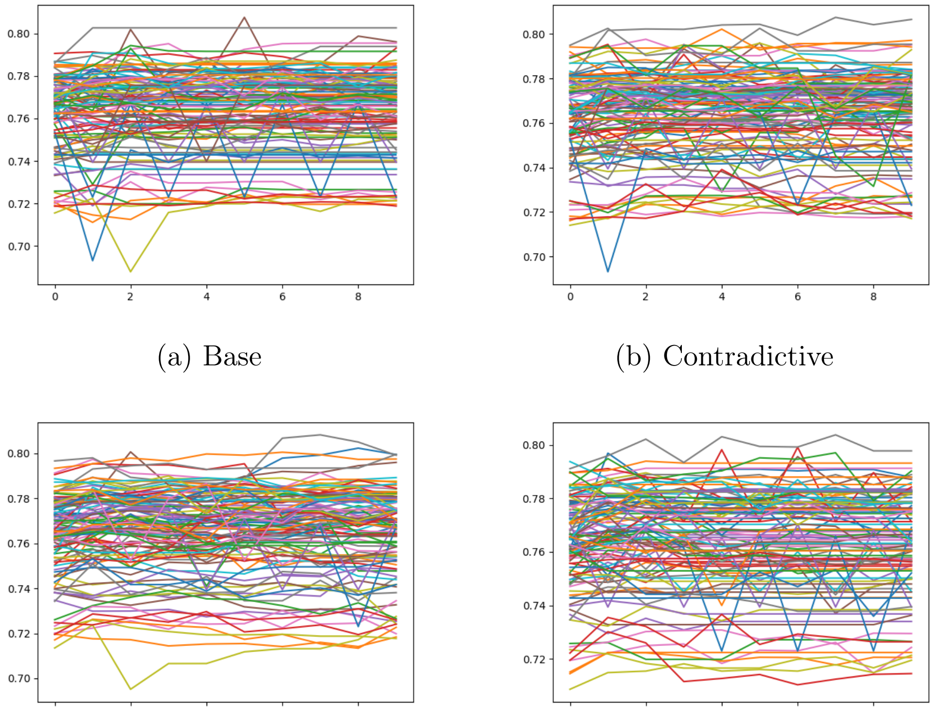
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 7: BERTscore - No initial context



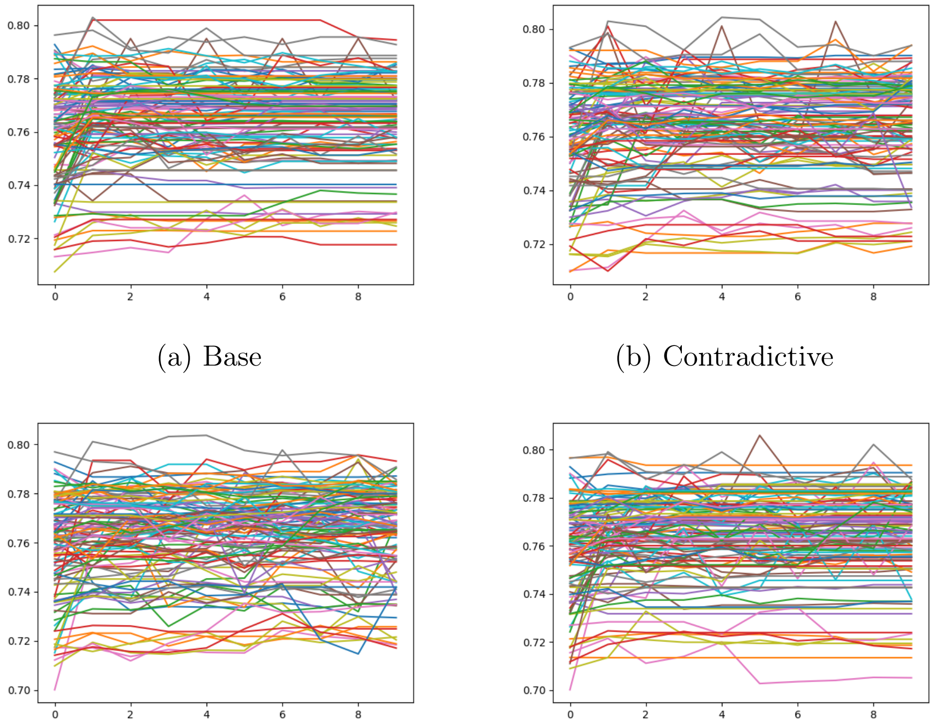
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 8: BERTscore - QA context



(a) Base

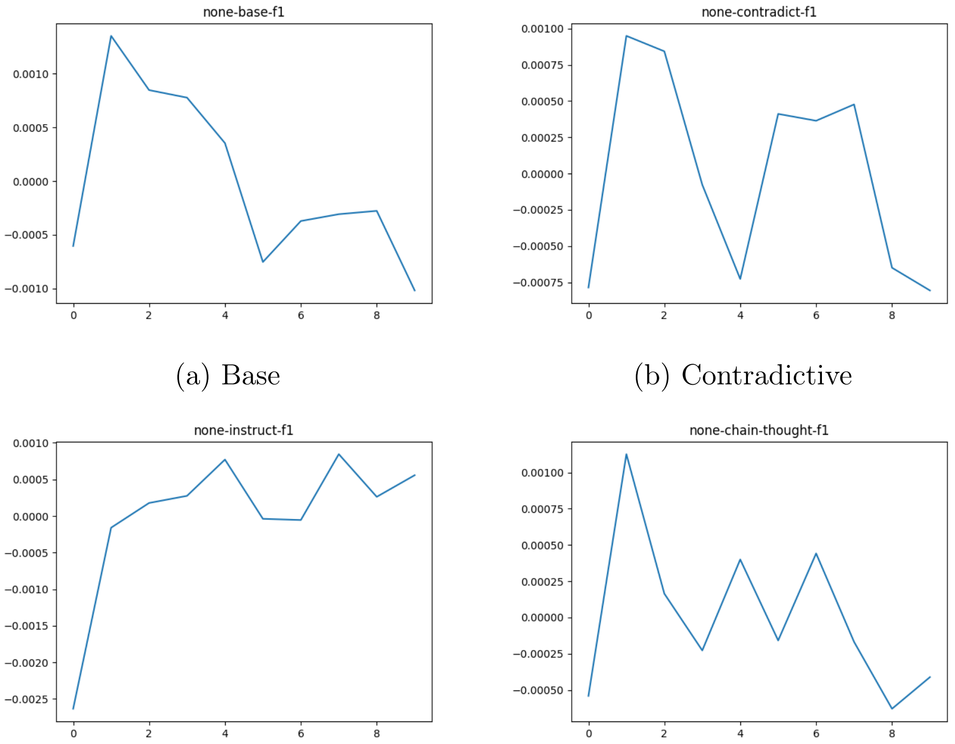
(b) Contradictive

(c) Instructive

(d) CoT

Figure 9: BERTscore - Hallucination initial context

4.2.2 HV



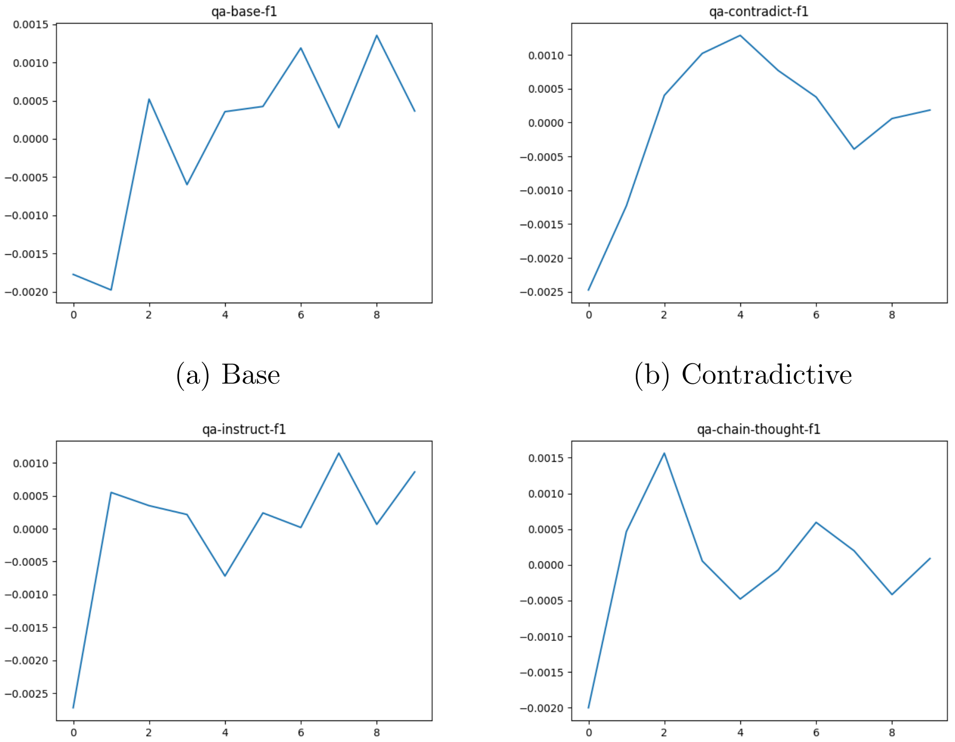
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 10: Hallucinatory Variance - No initial context



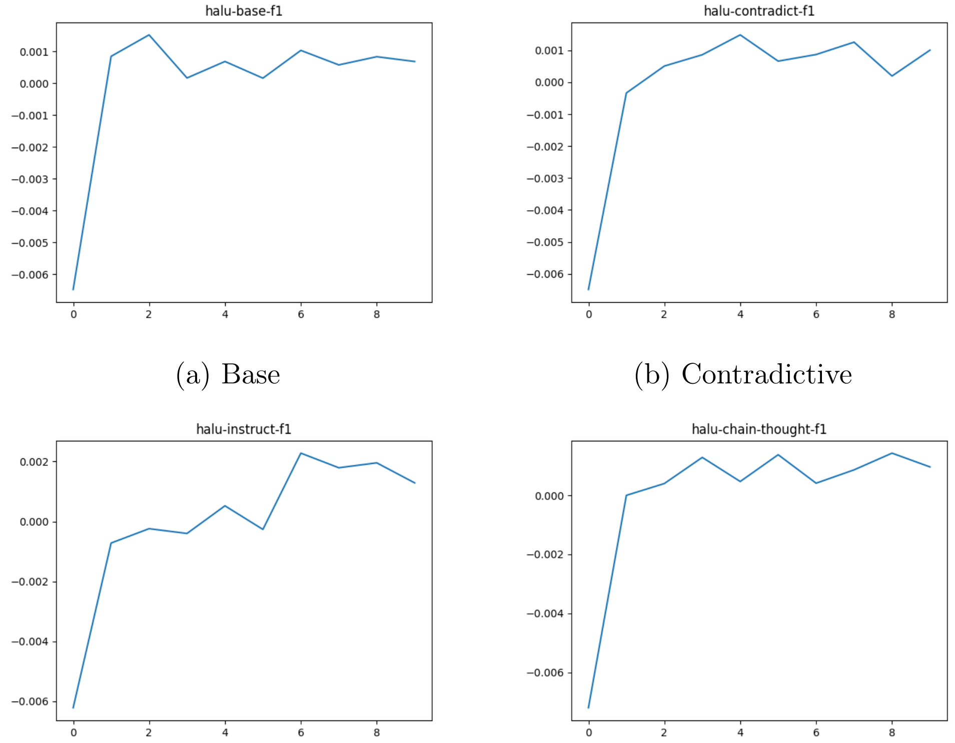
(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 11: Hallucinatory Variance - QA context



(a) Base

(b) Contradictive

(c) Instructive

(d) CoT

Figure 12: Hallucinatory Variance - Hallucination initial context

5. Discussion

These results seem to imply a pattern. Let's observe the "none" initial context first. We also mostly focus on the mixtral results as they are more substantive.

We can see patterns that seem to reflect trends in hallucination rate. In these graphs, a score below 0 reflects an observed value that is less than expected - a model that hallucinates more than it should given the prompt. The effect of the growing context length can thus be seen most clearly in figure 10. We see with the base chat-context that there is a steady decline in the difference between the score that rapidly becomes negative, meaning that the model quickly begins to falter as it is bombarded with the same information in a growing context. Chain of thought functions the same way, but it is more clear as well as more verbose, so the graph has more extremity in the peaks and valleys. Similarly, the Contradiction graph reflects a kind of wavering, bouncing above and below the zero line. This reflects the neutral/slightly affirmative way in which the model reacts to negative reinforcement, being more likely to hallucinate if told that a correct answer is wrong but then bouncing back when this second assumption is challenged. Notably, the Instructive chat context much more closely reflects that of the other Initial contexts, which implies that this method is possibly the best; this could be for a variety of reasons, however it is most interesting to consider that this version does not repeat the question, merely pointing to the model response, which may keep the attention closer to the initial generations.

The other initial contexts seem to buck this trend, as the patterns are all largely similar. It does seem here that repetition in fact legitimizes the model output, as there is a slight trend upward before leveling out *above 0*. Once again, in the context of our results, this implies that the model falters slightly at the beginning more frequently, but quickly adapts.

5.1 Challenges / Error Analysis

Essentially every decision made in this project had knock-on effects that made generating proper results much more difficult. For instance, the choice of Natural Questions is well motivated, and does not immediately raise any questions; however, by the time HV was more clearly conceptualized (i.e., when I decided to implement it via BERTscore), a problem had surfaced - not every question in NQ is annotated with gold spans. This is because the dataset is designed to include wikipedia articles which do not have the answer located within them - which is a strength of the dataset in contexts where a model is being fine-tuned and needs to be able to generalize broadly, but for a system designed to leverage BERTscore - semantic distance between the predicted answer and a gold reference - it simply gets in the way. The particular flaws of BERTscore are also relevant here, but ultimately the consistency of the BERTscore values implies that the source of error was elsewhere.

There are also flaws in the conceptualization of context in terms of "turns" for this metric. Given the auto-regressive nature of generation, it would be more rigorous to explore the variance as a function of the context length as a sequence. The mistake was in treating each generative output as a single instance, rather than looking at the relative effect of context length as the overall context increases.

6. Conclusion

We have shown that context length has an effect, though marginal, on the deviation in hallucination rates. We also show that the nature of user repetition can affect this deviation in ways that map to intuitions about the model response. Additionally, initial context seems quite important to the shape of this variation, as both QA and Hallucination contexts made the graphs follow a similar pattern that levels out very close to 0.

Future work should certainly consider the problem of hallucinations in broader terms than the discrete framing of "correct / incorrect". The hold that language models have on the people in control of cultural production means we need to thoroughly situate our notions of "the model" in an understanding of how we construct truth, as well as how we interface with these machines as a means/*source* of truth. While model adjudication is a quick and simple solution, we must not lose sight of the deeper questions asked by the LLM as a cultural signifier and how these interface with the ways that they are deployed.

References

Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E. M., & Reddy, S. (2022). FaithDial: A faithful benchmark for information-seeking dialogue (B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, 10, 1473–1490. https://doi.org/10.1162/tacl_a_00529

Hanna, M., & Bojar, O. (2021, November). A fine-grained analysis of BERTScore. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita & C. Monz (Eds.), *Proceedings of the sixth conference on machine translation* (pp. 507–517). Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.59

Jauregi Unanue, I., Parnell, J., & Piccardi, M. (2021, August). BERTTune: Fine-tuning neural machine translation with BERTScore. In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 915–924). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.115

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research (L. Lee, M. Johnson, B. Roark & A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, 7, 452–466. https://doi.org/10.1162/tacl_a_00276

Leiter, C. W. (2021, November). Reference-free word- and sentence-level translation evaluation with token-matching metrics. In Y. Gao, S. Eger, W. Zhao, P. Lertvittayakumjorn & M. Fomicheva (Eds.), *Proceedings of the 2nd workshop on evaluation and comparison of nlp systems* (pp. 157–164). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eval4nlp-1.16

Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023, December). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6449–6464). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.397

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. T. I., Chadha, A., Sheth, A., & Das, A. (2023, December). The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 2541–2573). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.155

Sun, T., He, J., Qiu, X., & Huang, X. (2022, December). BERTScore is unfair: On social bias in language model-based metrics for text generation. In Y. Goldberg, Z. Kozareva & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 3726–3739). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.245

Wang, X., Yan, Y., Huang, L., Zheng, X., & Huang, X. (2023, December). Hallucination detection for generative large language models by Bayesian sequential estimation. In H. Bouamor, J. Pino & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 15361–15371). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.949

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019, November). MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In K. Inui, J. Jiang, V. Ng & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 563–578). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1053