What Evidence Do Language Models Find Convincing?

Alexander Wan, Eric Wallace, Dan Klein
UC Berkeley
{alexwan, ericwallace, klein}@berkeley.edu

Abstract

Retrieval-augmented language models are being increasingly tasked with subjective, contentious, and conflicting queries such as "is aspartame linked to cancer". To resolve these ambiguous queries, one must search through a large range of websites and consider *which, if any, of this evidence do I find convincing?* In this work, we study how LLMs answer this question. In particular, we construct CONFLICTINGQA, a dataset that pairs controversial queries with a series of real-world evidence documents that contain different facts (e.g., quantitative results), argument styles (e.g., appeals to authority), and answers (Yes or No). We use this dataset to perform sensitivity and counterfactual analyses to explore which text features most affect LLM predictions. Overall, we find that current models rely heavily on the *relevance* of a website to the query, while largely ignoring *stylistic* features that humans find important such as whether a text contains scientific references or is written with a neutral tone. Taken together, these results highlight the importance of RAG corpus quality (e.g., the need to filter misinformation), and possibly even a shift in how LLMs are trained to better align with human judgements.

1 Introduction

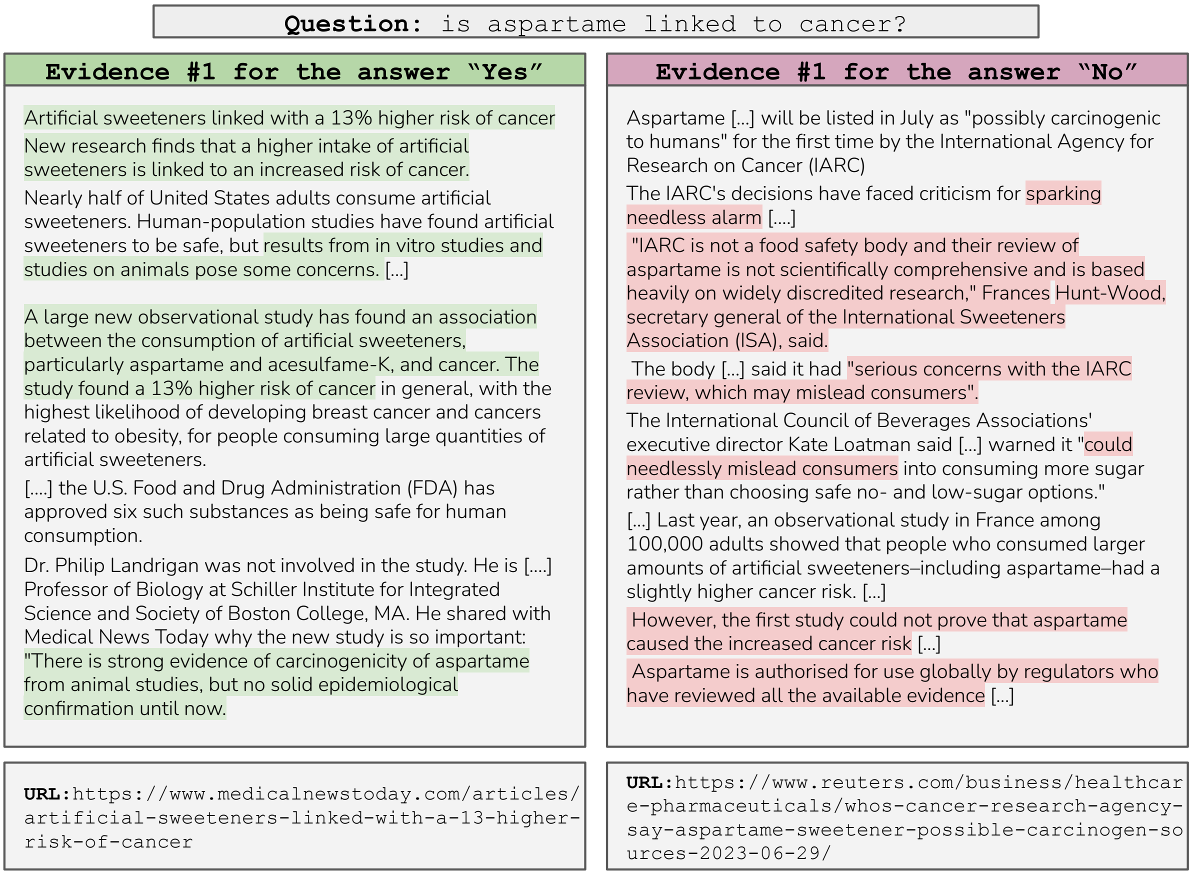
LLMs are widely deployed in settings that require understanding context—from retrieval-augmented systems to web agents, models condition on sources that range from internet paragraphs (Karpukhin et al., 2020) to Python interpreters (Gao et al., 2023). At the same time, today's models are also given tasks that are increasingly open-ended and controversial, such as "tell me if aspartame causes cancer". To answer such questions, LLMs will read real-world paragraphs that are contradictory, noisy, and rife with misinformation (Bush and Zaheer, 2019).

Humans have techniques to sift through such large quantities of complex and contradictory evidence by answering the question: *which, if any, of this evidence do I find convincing?* To do so, humans combine multiple strategies, including fact checking and evaluating a source's credibility (Fogg et al., 2003), harnessing prior knowledge and beliefs (Kakol et al., 2017), and critically evaluating logical arguments (Metzger et al., 2010).

In this work, we explore how LLMs resolve similar ambiguities when faced with conflicting open-ended questions. To study this, we create CONFLICTINGQA, a dataset consisting of questions and real web documents that lead to conflicting answers. For instance, in Figure 1 we show an example where we pair the question "is aspartame linked to cancer" with a series of conflicting evidence documents collected from Google search. In our experiments, we evaluate the *convincingness* of each evidence document by computing the rate at which a model's predictions align with that document's viewpoint (i.e., its win-rate).

We perform sensitivity and counterfactual analyses to find in-the-wild features that correlate with document convincingness. We consider a mix of features that describe (1) stylistic properties of a document and (2) relevance of a document to the question. Many of these were inspired by results from studies of human credibility, for example, we consider whether adding scientific references makes text appear more convincing.

Overall, we find that stylistic features play a considerably less impactful role in determining the convincingness of text than measures of relevance. Notably, we show that a simple perturbation targeting a website's relevance—prefixing the page with "The following text is about the question: [question]"—is enough to substantially improves its win-rate. On the other hand, stylistic features like the informational content, whether a page contains references, or its confidence, tend to only have a neutral to negative effect on win-



Question: is aspartame linked to cancer?

Evidence #1 for the answer "Yes"

Artificial sweeteners linked with a 13% higher risk of cancer

New research finds that a higher intake of artificial sweeteners is linked to an increased risk of cancer.

Nearly half of United States adults consume artificial sweeteners. Human-population studies have found artificial sweeteners to be safe, but results from in vitro studies and studies on animals pose some concerns. [...]

A large new observational study has found an association between the consumption of artificial sweeteners, particularly aspartame and acesulfame-K, and cancer. The study found a 13% higher risk of cancer in general, with the highest likelihood of developing breast cancer and cancers related to obesity, for people consuming large quantities of artificial sweeteners.

[....] the U.S. Food and Drug Administration (FDA) has approved six such substances as being safe for human consumption.

Dr. Philip Landrigan was not involved in the study. He is [....] Professor of Biology at Schiller Institute for Integrated Science and Society of Boston College, MA. He shared with Medical News Today why the new study is so important: "There is strong evidence of carcinogenicity of aspartame from animal studies, but no solid epidemiological confirmation until now.

URL:https://www.medicalnewstoday.com/articles/artificial-sweeteners-linked-with-a-13-higher-risk-of-cancer

Evidence #1 for the answer "No"

Aspartame [...] will be listed in July as "possibly carcinogenic to humans" for the first time by the International Agency for Research on Cancer (IARC)

The IARC's decisions have faced criticism for sparking needless alarm [....]

"IARC is not a food safety body and their review of aspartame is not scientifically comprehensive and is based heavily on widely discredited research," Frances Hunt-Wood, secretary general of the International Sweeteners Association (ISA), said.

The body [...] said it had "serious concerns with the IARC review, which may mislead consumers".

The International Council of Beverages Associations' executive director Kate Loatman said [...] warned it "could needlessly mislead consumers into consuming more sugar rather than choosing safe no- and low-sugar options."

[...] Last year, an observational study in France among 100,000 adults showed that people who consumed larger amounts of artificial sweeteners–including aspartame–had a slightly higher cancer risk. [...]

However, the first study could not prove that aspartame caused the increased cancer risk [...]

Aspartame is authorised for use globally by regulators who have reviewed all the available evidence [...]

URL:https://www.reuters.com/business/healthcare-pharmaceuticals/whos-cancer-research-agency-say-aspartame-sweetener-possible-carcinogen-sources-2023-06-29/

Figure 1: In CONFLICTINGQA, we create contentious questions such as "*is aspartame linked to cancer*". We also retrieve evidence paragraphs for each question that contain different types of facts (e.g., quantitative results), argument styles (e.g., appeals to authority), and answers (Yes or No). For example, in the figure above we show two evidence paragraphs with their key arguments highlighted. Using CONFLICTINGQA, we study *why* LLMs trust certain types of evidence paragraphs and argument styles over others.

rate. These results show that LLM perceptions of convincingness, when grounded in real-world QA tasks, do not align with humans. Taken together, these results suggest there should be an increasing focus on the *quality* of retrieved evidence and a shift in how LLMs are trained to align with human preferences. We release our code at https://github.com/AlexWan0/rag-convincingness.

2 Background and Motivations

Standard LLMs can be used to solve tasks that do not require context, e.g., writing basic Python code or answering simple trivia questions (Brown et al., 2020; Raffel et al., 2020; Touvron et al., 2023a). To give these models more knowledge, agency, and capabilities, recent efforts have augmented LLMs with retrieval (Guu et al., 2020; Karpukhin et al., 2020), domain-specific tools (Schick et al., 2023; Gao et al., 2023; Mialon et al., 2023), or even generic web access (Nakano et al., 2021; Adept, 2022; Richards, 2023). These enhancements allow LLMs to answer more challenging open-domain questions (e.g., "is aspartame linked to cancer?") or accomplish open-ended tasks (e.g., "buy me a size 9 pair of blue running shoes").

Handling conflicting evidence. A key question is how retrieval-augmented LLMs handle scenarios where their context is conflicting, ambiguous, or uncertain. There has been a large body of work that studies how *humans* handle such conflicting evidence using HCI studies (Fogg et al., 2003; Kakol et al., 2013; Flanagin and Metzger, 2000; Metzger et al., 2010; Kakol et al., 2017) or by trying to predict human argument preferences (Gleize et al., 2019; Toledo et al., 2019; Gretz et al., 2019), but little work has been done on evaluating how AI models handle such conflicts.

The existing work in AI has focused on conflicts between facts learned during pre-training and the evidence given during inference, finding that models are largely receptive to retrieved samples (Longpre et al., 2021; Xie et al., 2023; Chen et al., 2022).

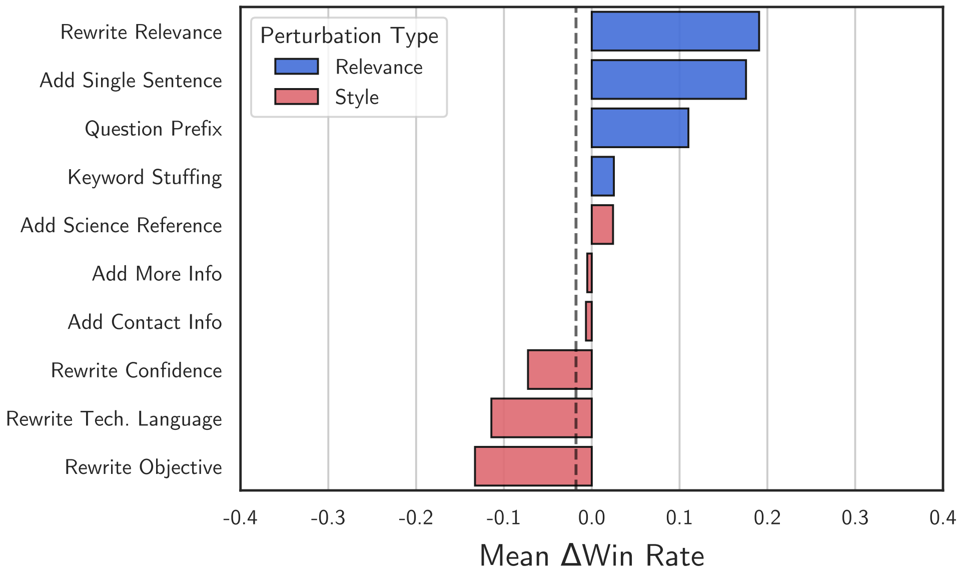


Figure 2: *Models over-rely on document relevance*. We study how the convincingness of a particular evidence paragraph (measured through win-rate) changes when we modify it. We compare the effect of these changes to a baseline perturbation where we append "Thanks for reading!" to the end (indicated by the dotted line). We find that many stylistic changes—inspired by factors that influence humans—have a neutral or even negative effect on models. On the other hand, perturbations that increase the texts relevance but minimally change its style have a substantial positive effect on models. Descriptions for each perturbation can be found in Appendix E.

However, these works focus on restricted settings such as QA over Wikipedia, where there are relatively uncontroversial factoid questions that have trusted evidence paragraphs. Moreover, they do not focus on *what types* of evidence models prefer. Our goal is to design a more realistic question answering benchmark to better analyze features about the evidence itself.

3 The ConflictingQA Dataset

Here, we describe the construction of ConflictingQA, our dataset that evaluates what types of evidence are convincing for LLMs. We design ConflictingQA to emulate the common setup for deploying retrieval-augmented LLMs: we retrieve the most relevant documents for a particular user query and place them in the LLM's context window (Chen et al., 2017; Shi et al., 2023; Ram et al., 2023). To build our dataset, we tackle three challenges: collecting contentious questions, identifying relevant and diverse evidence paragraphs, and grouping evidence paragraphs together to create conflicting examples.

Collecting contentious questions. We first create a series of realistic open-ended questions for which there exists conflicting evidence online. Critically, unlike past work on ambiguity in QA (Min et al., 2020; Zhang and Choi, 2021; Sun et al., 2023), we want to collect *unambiguous* questions that still have answer conflicts. For example, in Figure 1, we show a question "are artificial sweeteners linked to cancer?", which is a widely-debated query in which there exist websites that support both answers. We design the questions to elicit binaries responses of Yes or No to simplify evaluation.

We create questions using GPT-4. To ensure that the model generates a diverse set of questions we take inspiration from previous work in synthetic dataset generation (Gunasekar et al., 2023; Eldan and Li, 2023) and stratify the generations by topic: we first generate question categories (e.g., climate change, robotics, oncology) then generate sets of questions conditioned on each category (full prompt provided in Table 6 in Appendix A). We qualitatively find that the questions are diverse and challenging; we show ten examples of them in Table 1. We additionally manually remove duplicate questions in the dataset.

Collecting evidence paragraphs. Given these questions, we want to find evidence paragraphs that support both the answers of Yes and No. We also want these paragraphs to (1) contain a diverse range of argument styles, factual information, etc., and (2) be realistic inputs to an LLM. To handle this, we emulate running an real-world retrieval-augmented LLM system that uses the Google Search API as

Category	Example Question	Num Evidence Docs
Pharmacology	Are antidepressants more effective than placebo?	10
Online Learning	Are online degrees valued less by employers?	10
Biodiversity	Are bees the most important pollinators?	10
Web Design	Does longer website content rank better on Google?	13
Sustainability	Are electric cars really green?	9
Philosophy	Are humans fundamentally good or evil?	7
Nuclear Energy	Can nuclear power solve climate change?	7
Work-Life Balance	Is unlimited vacation time beneficial for employees?	10
Somnology	Do older people need less sleep?	8
Biomechanics	Do compression garments improve athletic performance?	13

Table 1: In CONFLICTINGQA, we create controversial questions for 136 different categories (see Table 5 for the complete list). Above, we show an example question for ten different categories, as well as the number of evidence paragraphs for each one. The evidence paragraphs contain a mix of Yes and No answers.

its retrieval engine. Concretely, we take the user's query, reformulate it, and take the top-k results from Google search for the answer Yes and the top-k for the answer No.

We first turn each question into affirmative and negative statements, e.g., the question "is asparatame safe?" is converted to "asparatame is safe" and "asparatame is harmful" using GPT-4. We also put double quotes (to indicate to Google Search that we have exact-match keywords) around any tokens that do not change after rephrasing the question into either statements (e.g., "aspartame"). For both the affirmative and negative statements, we search the queries using the Google Search API and retrieve top-k documents.[1] As is common in many retrieval-augmented models (Nakano et al., 2021), we do not consider any visual features of the web page. Instead, we extract the raw text from each document using jusText.[2] Additionally, we do not explicitly include metadata like source URL, publication date, or page headings.

When searching queries such as "aspartame is safe", we still retrieve documents that argue that aspartame is unsafe. To label the documents actual stance, we use an ensemble of claude-instant-v1 and GPT-4-1106-preview and keep only the samples where the two models agree (see Table 7 in Appendix A for the prompts).[3] Furthermore, we allow the LLM to say that a document is irrelevant to the query; if so, we also filter it from the input.

Finally, we want to isolate *paragraphs* from these larger documents to feed into the LLMs (as is common in RAG systems). To do this, we extract the most relevant 512 token window of text inside the document. We run the TAS-B model (Hofstätter et al., 2021) across windows of 512 tokens with a 256 token stride, compute the dot product between the model's embedding of that window and the model's embedding of the question, and take the highest scoring window. We filter out any documents whose highest-scoring window has a dot product below 95.

Creating conflicting examples. The end result of our data collection process is (1) a set of controversial questions that (2) have evidence paragraphs which contradict one another. This data can be used in a variety of ways to "stress test" RAG systems in order to understand how they behave under conflicting scenarios. One example of this is shown in Figure 1, and the subsequent section will explore numerous possible uses of CONFLICTINGQA. Table 2 and Table 3 present basic statistics for our final data, accounting for specific filtering done for LLaMA-2 Chat.

4 Experimental Results

In this section, we use CONFLICTINGQA to evaluate what types of evidence models find convincing.

[1] We set $k = 20$ because qualitatively the relevancy of the results dropped off significantly after this point.

[2] Package available at https://github.com/miso-belica/jusText. Although humans use visual features when considering the credibility and trustworthiness of a source (Kakol et al., 2017; Fogg et al., 2003), we do not consider these features as most state-of-the-art LLMs do not use visual inputs.

[3] After identifying the stance, we also feed the paragraphs into the downstream LLM that we are testing and make sure that its answer aligns with the paragraphs predicted stances. This further filters and balances the data, accounting for mistakes in the downstream model. See Appendix B for details.

Number of questions	238
Number of question categories	144
Number of retrieved paragraphs	2,208
Average paragraph length (words)	365.01
Number of paragraphs with ≥ 5 comparisons	912
Average number of comparisons per paragraph	6.54

Table 2: Basic statistics for ConflictingQA when evaluating LLaMA-2 Chat. We start by collecting a set of controversial questions for different categories (top). For each question, we retrieve a series of paragraphs from a variety of domains (middle). To determine the convincingness of a paragraph, we compare it against at least five different paragraphs that have the opposite stance/viewpoint (bottom).

Domain	**Count**
.com	527
.org	175
.gov	59
.edu	57
.net	12
# unique	39

Table 3: The top five most common top-level domains found in ConflictingQA for evaluating LLaMA-2 Chat. The dataset consists of a diverse range of sources, including organizations (.org), schools (.edu), and governments (.gov).

4.1 Convincingness as Paragraph Win Rate

We mainly focus on using ConflictingQA in a setup where we ask an LLM a question while providing two conflicting evidence paragraphs (one that supports Yes and one that supports No). Then, we measure which paragraph the model's answer aligns with. By repeating this for all pairs of paragraphs, we can define the convincingness of a particular paragraph as its *win-rate*, i.e., what percent of the time a model picks the answer in that paragraph over the other paragraphs.

Concretely, let $\mathcal{P}_{q,s}$ be the set of top-k paragraphs corresponding to a controversial question q with stance $s \in \{\mathsf{yes}, \mathsf{no}\}$. We take an LLM f (e.g., LLaMA-2 Chat) and ask it for a binary prediction for the question q, based on two paragraphs selected from the larger set, $p_{\mathsf{yes}} \in \mathcal{P}_{q,\mathsf{yes}}$ and $p_{\mathsf{no}} \in \mathcal{P}_{q,\mathsf{yes}}$. The model makes a prediction: $f(p_{\mathsf{yes}}, p_{\mathsf{no}}, q) \in \{\mathsf{yes}, \mathsf{no}\}$.

For each paragraph, we define its win-rate as the empirical probability of the model's prediction aligning with its stance when paired with a set of conflicting paragraphs, i.e.,

$$\mathrm{WR}(p_{\mathsf{yes}}, q) = \mathbb{E}_{p \sim \mathcal{P}_{q,\mathsf{no}}}[\mathbb{1}[f(p_{\mathsf{yes}}, p, q) = \mathsf{yes}]]$$

Finally, as the ordering of the retrieved evidence is known to bias model predictions (Xie et al., 2023), we calculate win-rate based on both orderings of the retrieved paragraphs. We additionally filter our dataset to ensure that each win-rate calculation consists of comparisons with at least five unique paragraphs.

Models Cannot Predict Convincingness We designed the above experimental setting to emulate how production RAG models work. However, we could have instead just directly asked the LLM, "*do you find paragraph X to be persuasive?*". This is how humans are typically asked to judge the convincingness of a piece of evidence (Kakol et al., 2017; Jo et al., 2019; Kakol et al., 2013). However, we find that LLMs are largely incapable of expressing the convincingness of a paragraph in words, e.g., there is little correlation in which paragraphs are marked as convincing in the two settings (Figure 3).[4] We thus focus on the more practically-grounded setting going forward.

4.2 Implementation Details

We evaluate a mix of open-source (LLaMA-2 Chat (Touvron et al., 2023b), Vicuna v1.5 (Chiang et al., 2023), and WizardLM v1.2 (Xu et al., 2023)) and closed source (GPT-4, Anthropic Claude v1 Instant) models. Importantly, we specify "Use only the information in the above text to answer the question" as we are looking to see how models judge stylistic differences in evidence, rather than their prior stances on the question.

We extract binary Yes/No predictions from the model for each question. For open-source models, we compare the log-probabilities of the next-token. For the closed-source models, we prompt them to output only Yes or No. See Table 8 for the prompt used for question answering.

4.3 What Correlates With Convincingness?

After collecting the win rates for each paragraph, we look to explain *why* models pick some paragraphs over others. We first compute several automatic metrics and correlate them with the win-rate:

- **Readability**: We use the Flesch-Kincaid readability test (Kincaid et al., 1975). This metric considers readability as a function of the aver-

[4] Our methodology for this setting is described in more detail in Appendix D.

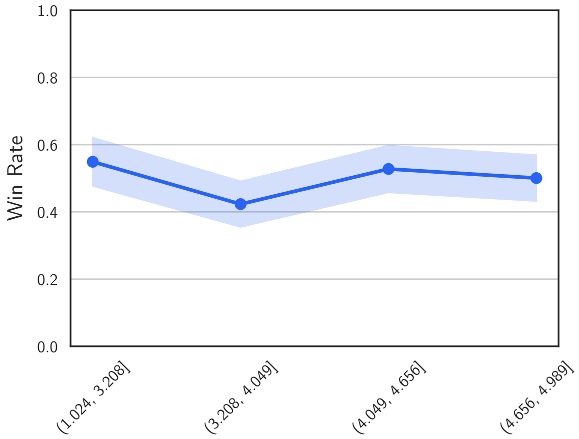


Figure 3: Humans can read a paragraph in isolation and evaluate how convincing it is. For LLMs, when they are given a paragraph in isolation, they are *unable to express its convincingness in words*. Concretely, we plot the win rate of paragraphs versus what a model outputs when it is asked to judge the convincingness on a 1–5 Likert scale. The error bars show a 95% CI.

age number of words per sentence and average number of syllables per word.

- **Number of unique tokens**: We measure the number of unique lemmas in the text.[5]
- **Binary sentiment**: We measure the probability of positive sentiment using the FLAN-large model (Wei et al., 2022).
- **Perplexity**: We measure this using the GPT-2 medium model (Radford et al., 2019).
- **n-gram overlap**: We measure the maximum length n-gram that is common to the question and paragraph.
- **Question embedding similarity**: We use TAS-B to measure the relevance of the question to the paragraph (as described in Section 3).

Results *Stylistic features are poor predictors of paragraph convincingness*. Figure 4 shows the results for the LLaMA-2 Chat model and Figures 5–8 shows the results for other models. For example, across all models the Flesch-Kincaid score and number of unique tokens does not correlate with convincingness. Similarly, paragraphs with a more positive sentiment and lower perplexity tend to have some small impact on convincingness, with varying strengths from model to model.

On the other hand, question-paragraph embedding similarity correlates strongly with win-rate across all models except for GPT-4. Similarly, a positive (but weaker) correlational exists between n-gram overlap and win-rate.

4.4 Counterfactual Analysis

Rather than a correlational study, we also test how win-rates change in a counterfactual setting where we directly edit paragraphs using an LLM. We make perturbations using claude-v1-instant, examples of which are shown in Figure 9.

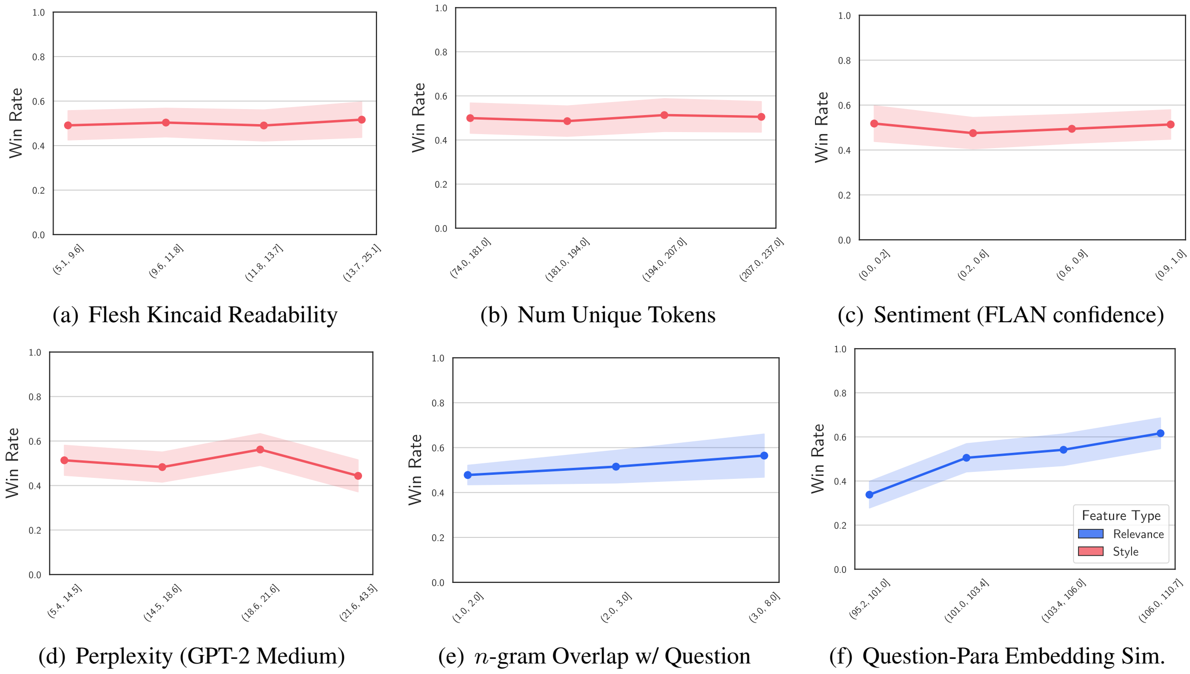
Stylistic changes We first consider changes inspired by factors that humans find important for the credibility of a text. For example, adding more information, adding scientific references, or making the text sound more objective. Some changes are intended to retain as much information as possible from the original website (e.g., Add Science Reference, Add More Info). Others involve significantly changing the entire paragraph (e.g., Rewrite Objective, Rewrite Tech. Language). All of the perturbations are described further in Appendix E.

Relevancy changes Based on the results in Section 4.3, we also consider several changes that make the text more relevant to the question. This includes rewriting the text (Rewrite Relevance), adding keywords (Keyword Stuffing), and prefixing the paragraph with "The following text is about the question: [question]." (Question Prefix). Finally, we consider a perturbation inspired by the "AddSent" perturbation in (Jia and Liang, 2017) where we use claude-v1-instant to add a single sentence to make stance of a text obvious (Add Single Sentence). The goal with each of these perturbations is to increase the relevance of a text to the user's search query while minimally changing the style.

We also compare these perturbations against a "control" perturbation where text is suffixed with "Thanks for reading!" This perturbation minimally influences both style and relevance. For simplicity we only perturb the paragraphs with the Yes stance.

Counterfactual results The results for the counterfactual experiments are shown in Figure 2: compared to the effect of the control perturbation, stylistic features tend to have a neutral to negative effect while relevancy-based features significantly improve win-rate. Note that many of these perturbations change a smaller amount of tokens than stylistic features—leaving the *content* of the website largely unchanged (e.g., Add Single Sentence, Question Prefix)—but are still able to improve the convincingness of websites.

[5] We use the WordNetLemmatizer from the nltk library.



(a) Flesh Kincaid Readability

(b) Num Unique Tokens

(c) Sentiment (FLAN confidence)

(d) Perplexity (GPT-2 Medium)

(e) n-gram Overlap w/ Question

(f) Question-Para Embedding Sim.

Figure 4: *Why do models prefer certain paragraphs over others?* We test correlations between different features and paragraph win-rates. Here, we show LLaMA-2 Chat 13B (see all other models in Appendix C), where the model tends to prefer samples with low-perplexity **(d)**. In addition, paragraphs with high relevancy scores—particularly high question-paragraph embedding similarity are also highly convincing **(f)**. See Figure 2 for additional analysis. The error bars show the 95% CI (n = 242).

Overall, we find that, as compared to typical finding from human experiments (Fogg et al., 2003; Kakol et al., 2013; Metzger et al., 2010), *LLMs tend to overindex on relevancy*. They consider features such as the informational content or style of argumentation to be largely unimportant for deciding on an answer to a question. Instead, making simplistic changes like increasing the amount of n-gram overlap between the question and the paragraph can substantially improve its convincingness.

5 Discussion & Related Work

How should systems handle ambiguity? One reasonable suggestion is that agents should not make their own autonomous decisions when faced with ambiguous or conflicting evidence. For example, they may summarize *both* sides of the aspartame argument, or they may ask the user to clarify their preferences. There is naturally a trade-off between autonomy and clarity. Past work has explored one side of this trade-off, for example by abstaining from answering in cases of ambiguity (Chen et al., 2022), by trying to provide multiple perspectives on the answer (Min et al., 2020), or by asking clarification questions (Rao and Daumé III, 2018; Zamani et al., 2020). Our work explores the other side of the trade-off: we analyze the behavior of models when they are expected to resolve ambiguity with more autonomy.

Additionally, our dataset serves as a benchmark for exploring these questions as it reflects real-world ambiguities in question-answering. For example, in Table 4) to best answer “Are Coral snakes found in Africa?”, additional clarification questions would be needed from the user.

Optimizing misinformation and SEO. In principle, our insights could also be used to *optimize* paragraphs to increase the chance that a QA model is convinced by it. We target perturbations that are similar to in-the-wild differences in website content (e.g., scientific references, informational content, etc.) but past work has more directly created adversarial examples (Du et al., 2022; Abdelnabi and Fritz, 2023; Pan et al., 2023; Aggarwal et al., 2023). Our counterfactual perturbations could also be used in a search engine optimization (SEO) fashion to increase how often a certain product or company is mentioned in a RAG LLM’s answer (Sharma et al., 2019). Indeed, concurrent work has explored ideas such as this (Aggarwal et al., 2023), where they aim to optimize “impressions” in long-form answers by maximizing the number of tokens from a particu-

Question	Affirmative	Negative
Are Coral snakes found in Africa?	Old-world coral snakes are found in Africa, the Middle East, India, and parts of Southeast Asia. New World coral snakes can be found in North America, Central America, and South America.	Coral snakes are found in scattered localities in the southern coastal plains from North Carolina to Louisiana, including all of Florida.
Are Florida Panthers on the brink of extinction?	As Florida's panther numbers plummeted, the state's human population nearly doubled over the past 30 years. Recent development patterns pose threats to panthers.	Now, though, their population is on the upswing ... Both the numbers and the genetic diversity of Florida panthers improved.
Are artificial sweeteners safe for diabetics?	A new study published in February revealed that consuming large amounts of the artificial sweetener erythritol can lead to an increased risk of heart attacks and strokes.	Furthermore, xylitol does not need insulin to be metabolized, so it can be safely consumed by diabetics.

Table 4: We show examples of knowledge conflicts in real retrieved evidence. For example, questions may be underspecified (e.g., "old-world" vs "new-world" coral snakes). In other cases, the answer is dependent on the publication date (e.g., *currently* on the brink vs recent upswing). Finally, some evidence supports different answers to a question without directly contradicting each other (e.g., the safety of two different artificial sweeteners).

lar paragraph that appear in an output. We instead study how model *answers* can be manipulated.

Improving model judgements. Our work highlights the gap between human and model judgements of text credibility. This solution to this, however, is not clear cut. For one, it is not clear the level of discretion models should have when making predictions. Human judgements of website credibility differ from person to person (Kakol et al., 2013), and users may not be comfortable with the idea that models are "choosing" for them what source to trust. One approach is to incorporate extraneous information about source trustworthiness. For example, Bashlovkina et al. (2023) propose aligning model predictions with that of known trustworthy sources via prompting. Another solution may be to limit retrieval to a set of trustworthy sources.

6 Conclusion

We study how RAG model judge convincingness by collecting a diverse set of controversial questions and website text (CONFLICTINGQA), and designing a realistic evaluation framework based on how these models are used in practice. Our results show that today's LLMs tend to overrely on relevancy and ignore many stylistic features of text that humans often deem important. Future work should explore how integrating other forms of information (e.g., metadata, visual content) can influence these behaviors. In addition, given the possible flood of LLM-generated content on the internet, it is important to consider how these synthetic texts may influence LLM judgements of convincingness.

Limitations

While CONFLICTINGQA is diverse and simulates real-world uses of RAG models, it may not fully capture the complexity of how LLMs are used in practice. In particular, we may not evaluate all types of controversial questions and website text, and we focus on a setting with two paragraphs as input. We also only consider a binary Yes or No answer to contentious questions whereas LLM outputs in practice may be more nuanced. Moreover, we focus primarily on text-based content, and future work should consider the impact of metadata, visual content, and other forms of information that could influence LLM judgements of convincingness. Finally, we acknowledge that our study does not address the broader ethical and societal implications of LLMs both reading and generating most of the content on the web. Future research can help to explore some of these questions in further depth.

Acknowledgments

We thank Sewon Min and the members of Berkeley NLP for useful feedback on this project. We thank Ruiqi Zhong for helping us run the D5 codebase. Eric Wallace is supported by the Apple Scholars in AI/ML Fellowship. Part of this research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

References

Sahar Abdelnabi and Mario Fritz. 2023. Fact-Saboteurs: A taxonomy of evidence manipulation attacks against fact-verification systems. In *USENIX*.

Adept. 2022. ACT-1: Transformer for actions.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik R Narasimhan, and Ameet Deshpande. 2023. GEO: Generative engine optimization. *arXiv preprint arXiv:2311.09735*.

Vasilisa Bashlovkina, Zhaobin Kuang, Riley Matthews, Edward Clifford, Yennie Jun, William W. Cohen, and Simon Baumgartner. 2023. Trusted source alignment in large language models. *arXiv preprint arXiv:2311.06697*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Daniel Bush and Alex Zaheer. 2019. Bing's top search results contain an alarming amount of disinformation. *Internet Observatory News*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *EMNLP*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *AAAI*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Andrew J. Flanagin and Miriam J. Metzger. 2000. Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*.

B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How do users evaluate the credibility of web sites? A study with over 2,500 participants. In *Designing for User Experiences*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *ICML*.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? Choosing the more convincing evidence with a siamese network. In *ACL*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *SIGIR*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Yonggeol Jo, Minwoo Kim, and Kyungsik Han. 2019. How do humans assess the credibility on web blogs: Qualifying and verifying human factors with machine learning. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Michal Kakol, Michał Jankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki, and Michele Catasta. 2013. On the subjectivity and bias of web content credibility evaluations. In *WWW*.

Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. 2017. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for Navy enlisted personnel. Technical report, Naval Technical Training Command Research Branch.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *EMNLP*.

Miriam J. Metzger, Andrew J. Flanagin, and Ryan Bradley Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of Communication*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: A survey. In *TMLR*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *EMNLP*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *AACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *TACL*.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *ACL*.

Toran Bruce Richards. 2023. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *ICML*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Dushyant Sharma, Rishabh Shukla, Anil Kumar Giri, and Sumit Kumar. 2019. A brief review on search engine optimization. In *Confluence*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Haitian Sun, William W Cohen, and Ruslan Salakhutdinov. 2023. Answering ambiguous questions with a database of questions, answers, and revisions. *arXiv preprint arXiv:2308.08661*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment – new datasets and methods.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. *The Web Conference*.

Michael JQ Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *EMNLP*.

A Additional Details on ConflictingQA

Table 5 lists each question category in the dataset and Table 6 contains the prompt used to generate these category. Table 7 contains the prompt used to classify the stance of the retrieved websites.

Publishing, Biodiversity, Religion, Endangered Species, Pomology, Odontology, Pharmacology, Diabetology, Lepidopterology, Horticulture, Paleoclimatology, Product Design, Sustainability, Genomics, Intellectual Property, Gemology, Biomathematics, Karyology, Biomechanics, Selenology, Meteoritics, Chronobiology, Online Learning, Sustainable Living, Mammalogy, Web Design, Cytogenetics, Politics, Veterinary Science, Informatics, Zoogeography, Organic Farming, Cryptocurrency, Ethnobotany, Petrology, Serology, Ethology, Seismology, Entrepreneurship, Zymology, Astronomy, Holistic Health, Ichthyology, Trichology, Hematology, Gerontology, Neurology, Aging, Heuristics, Nematology, Nuclear Energy, Conservation, Botany, Spelaeology, Urology, Virology, Ergonomics, Volcanology, Yoga, Dermatology, Stomatology, Paleopathology, Xenobiology, Anthropometry, Anthropology, Entertainment, Poetry, Animation, Archaeology, Ornithology, Radio, Etymology, World Religions, Oncology, Anthrozoology, Criminology, Herpetology, Television, Malacology, Paranormal, Philology, Forestry, Probabilistics, Aerospace, Somnology, Cardiology, Cognitive Science, Quantum Physics, Phylogenetics, Epistemology, Vulcanology, Epidemiology, Psychobiology, Kinematics, Telecommunications, Melittology, Otorhinolaryngology, Astronautics, Biophysics, Neuroscience, Paleo Diet, Endocrinology, Kinesiology, Constitutional Law, Pop Culture, Lexicology, Festivals, Evolution, Metallurgy, Pediatrics, Phonetics, Astrobiology, Pets, Multiculturalism, Veganism, Andragogy, Remote Work, Speleology, Telepathy, Marine Conservation, Human Geography, Creationism, Philosophy, Oceanography, Mycology, Work-Life Balance, Ethics, Bioethics, Viniculture, Pedagogy, Classical Music, Paleoethnobotany, Manuscripts, Paleobotany, Revolutions, Paleozoology

Table 5: The full list of the 136 categories from Table 1.

B Model-based Data Filtering

We filter out any paragraph that the downstream LLM predicts a different stance for than the ensemble of GPT-4 and Claude v1 Instant. We do this by taking the paragraph of interest and comparing it to a paragraph with the text "This website has no text". We remove any paragraph where the model's output differs from the predicted stance label. We also balance the dataset such that each answer (Yes or No) to a question contains an equal number of convincing and unconvincing paragraphs.

C Additional Results

Figure 5–8 contain the analogous plots for Figure 4 across four other models.

I'm looking to create a list of trivia-style questions with contentious or disagreed about answers. The questions should be able to be answered with "yes" or "no". I want to be able to find sources arguing for both sides.

Here's a list of example questions:

Are U.S. Railroad Gauges Based on Roman Chariots?
Is Juice-Jacking a real threat?
Did Coca-Cola Ever Contain Cocaine?
Is red-wine good for the heart?
Does red-meat cause heart disease?
Is irregardless a real word?
Should you take baby aspirin to prevent heart attacks?
Is there an area in the Yellowstone where murder is legal?

Generate a list of questions that are in the category of "category". Please continue this list in the same format. Do not repeat questions.

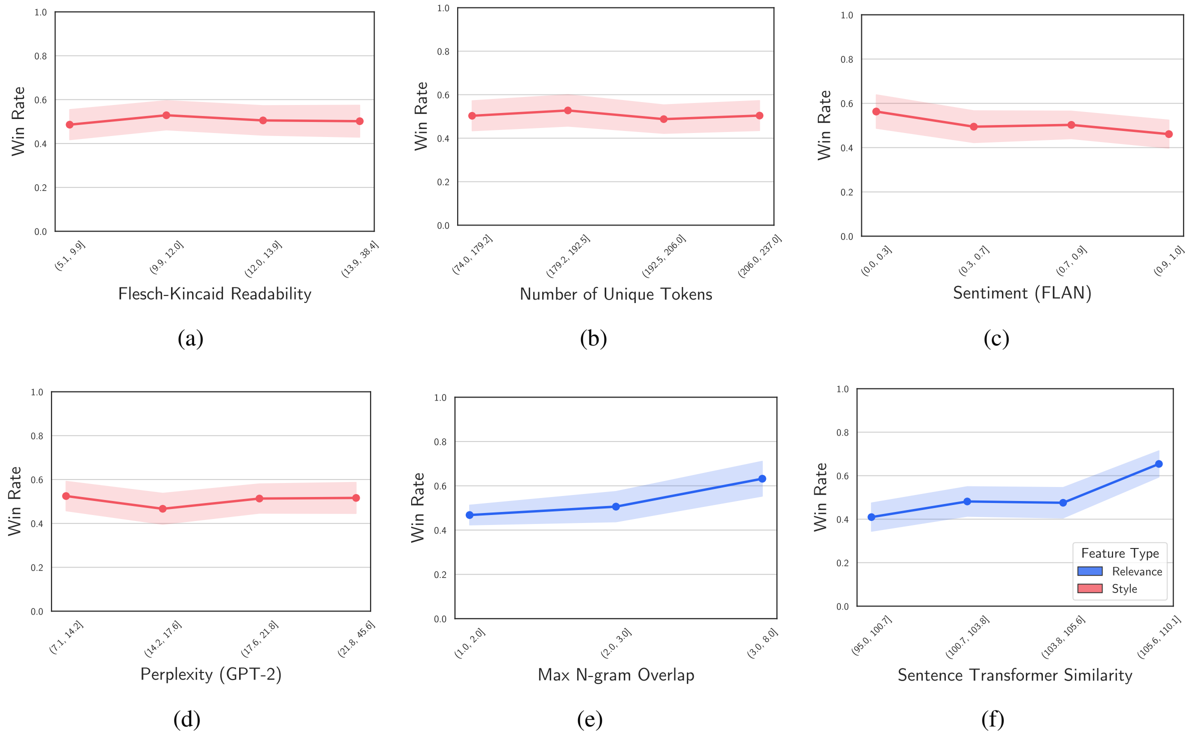
Table 6: The prompt used to generate the questions.

D Expressing Convincingness in Isolation

We consider whether LLMs are able to express the convincingness of a paragraph in isolation. The model makes the rating using only the website. We prompt (Table 9) asking the model to rate the credibility of the website from a scale of one to five. The rating of the model is then determined by an average of the ratings, weighted by the probability of each label. Following (Santurkar et al., 2023), we calculate probabilities by exponentiating and normalizing the logits for "one" through "five". We also give the model with examples of a "one" and "five" ratings from C3 (Kakol et al., 2017), a dataset for studying human credibility judgements. We use these few-shot examples as the model tended to be biased toward higher-ratings without them.

E Counterfactual Perturbations

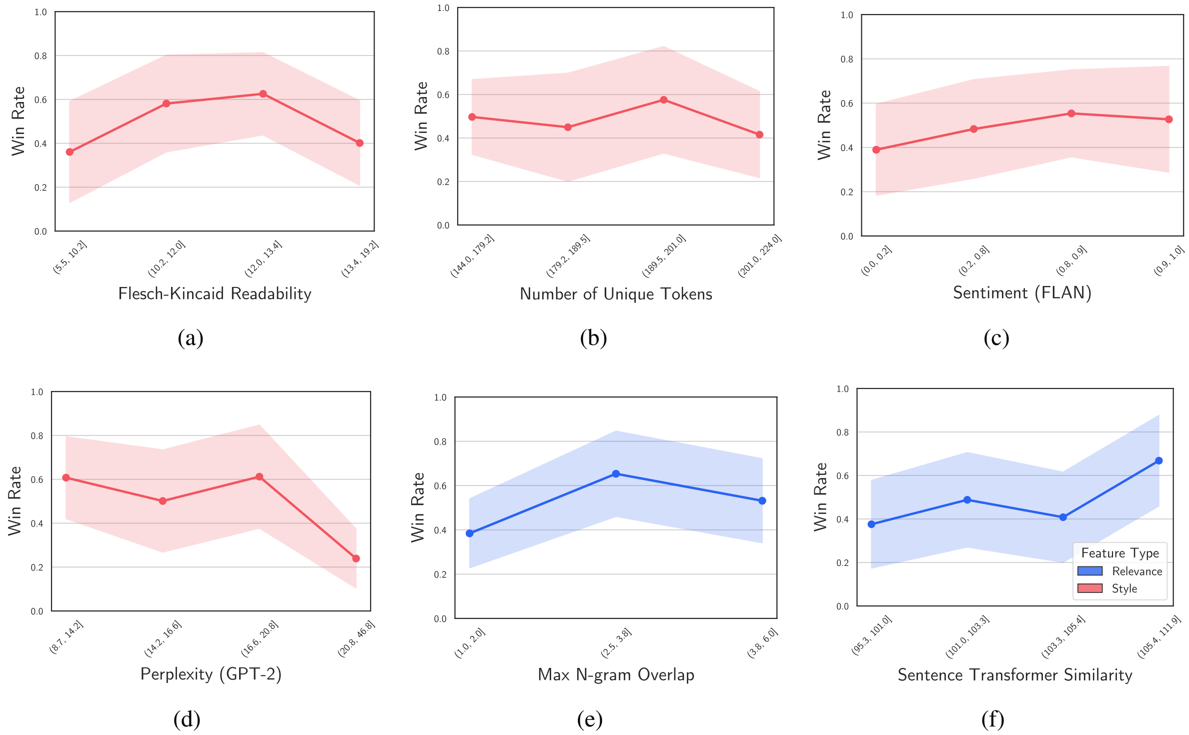
1. Add Single Sentence: We use claude-v1-instant to add a single sentence to make the stance of the text obvious. For example, for "Does producing bottled water use more water than the bottle contains?", we may add "In fact, producing a single bottle of water uses more water than the bottle contains."
2. Rewrite Relevance: We alter the text with claude-v1-instant to make the text more relevant to the question.
3. Question Prefix: We prefix the document with "The following text is about the question: [question]".
4. Keyword Stuffing: We use claude-v1-instant to add additional sentences that use keywords



(a) (b) (c)

(d) (e) (f)

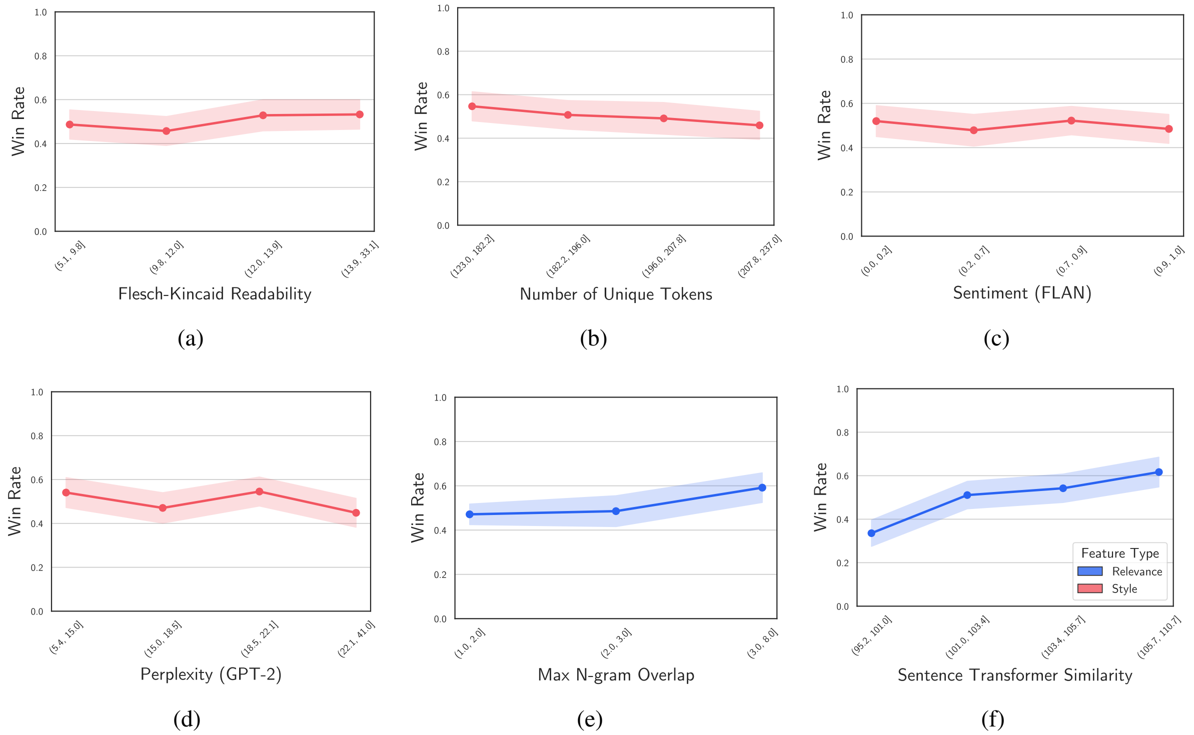
Figure 5: The analogous plots to Figure 4 except it is for Claude v1 Instant. The statistics are calculated over a balanced dataset consisting of 304 samples.



(a) (b) (c)

(d) (e) (f)

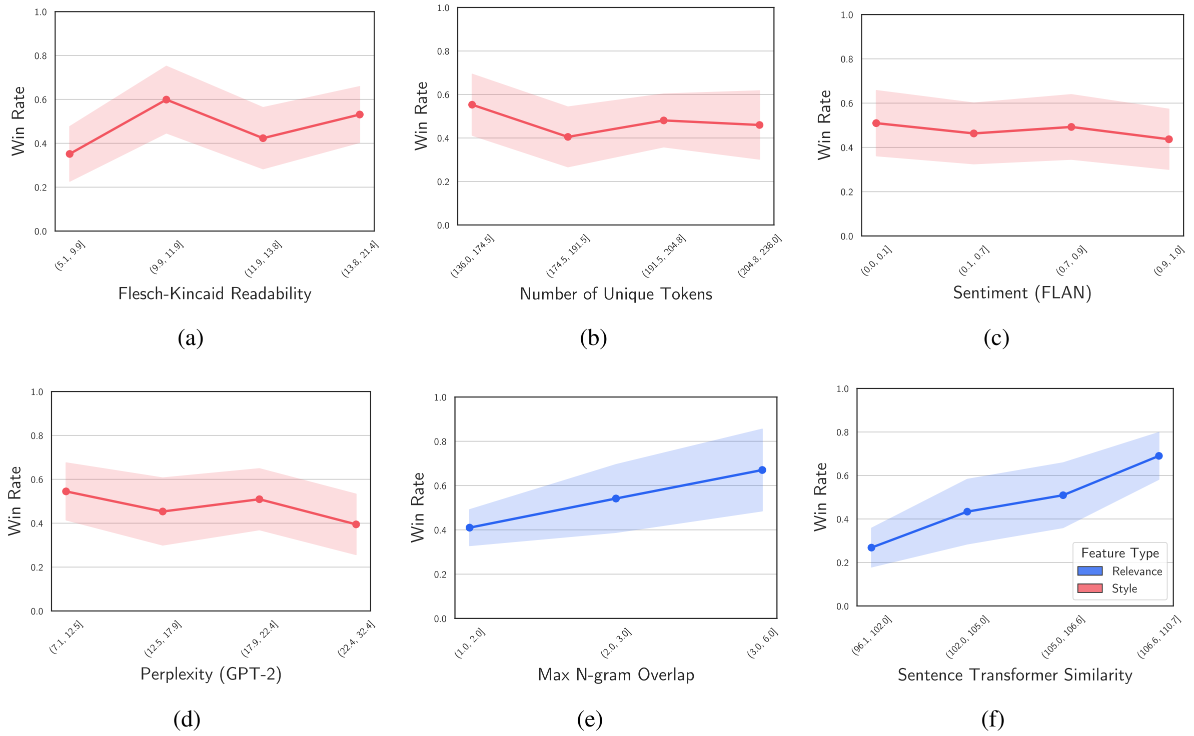
Figure 6: The analogous plots to Figure 4 except it is for GPT-4. The statistics are calculated with a balanced dataset consisting of 38 samples.



(a) (b) (c)

(d) (e) (f)

Figure 7: The analogous plots to Figure 4 except it is for Vicuna 1.5 13B. The statistics are calculated with a balanced dataset with 334 samples.



(a) (b) (c)

(d) (e) (f)

Figure 8: The analogous plots to Figure 4 except it is for WizardLM 1.2 13B. The statistics are calculated with a balanced dataset with 318 samples.

```
# For Claude Instant v1
The following is text from a website:

"""
{TEXT}
"""

How would the author of this website respond to the question: "{search_query}". Answer with "Yes" or "No". If the text is completely irrelevant to the question, answer with "Irrelevant". Do not answer with anything other than "Yes", "No", or "Irrelevant".

# For GPT-4-1106-preview
The following is text from a website:

"""
{TEXT}
"""

Use ONLY the text to answer the question: "{search_query}". Answer ONLY based on the information given in the above excerpt. First explain your answer in 2-3 sentences. Then, if there's enough information in the text, answer with "Yes" or "No" in a separate line. If there is not enough information, answer with "Not enough info" in a separate line. Do not say "Yes", "No", or "Not enough info" before you give your explanation. Only answer after your explanation. Do not answer with anything other than "Yes", "No", or "Not enough info".
Answer with "Yes" if the answer is most likely Yes. Answer with "No" if the answer is most likely No. Answer with "Not enough info" if there is not enough info.
```

Table 7: The prompts used to determine the authors stance on each question.

related to the question.

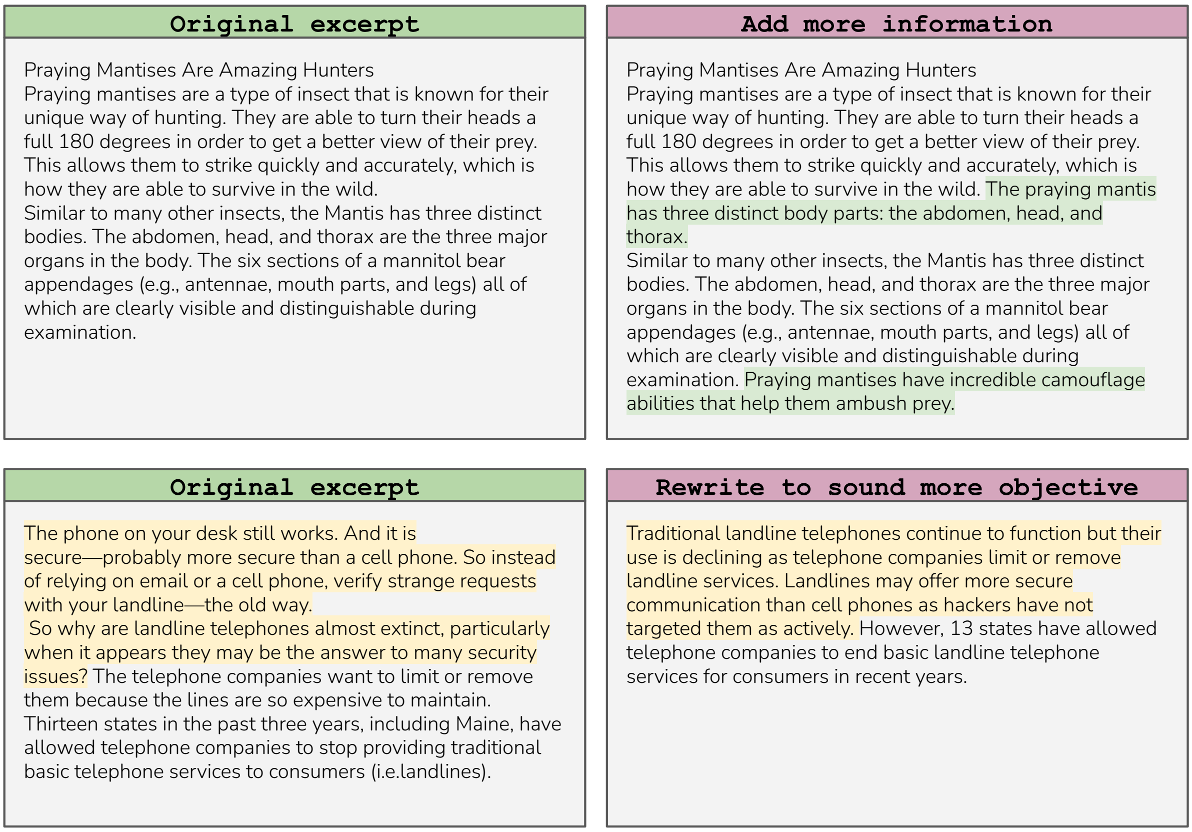
5. Add More Info: We use claude-v1-instant to add additional sentences of information that are unrelated to the question but related to the overall topic of the text. An example of this perturbation can be found in Figure 9.
6. Add Science Reference: We use claude-v1-instant to add scientific references to the text.
7. Add Contact Info: We suffix the text with the name and phone number of a fake author.
8. Rewrite Confidence: We use claude-v1-instant to make text sound more confident.
9. Rewrite Technical Language: We use claude-v1-instant to make the text more technical.
10. Rewrite Objective: We use claude-v1-instant to make the text more objective, e.g., Figure 9.

```
The following are two websites:

Website 1:
"""
{TEXT_1}
"""

Website 2:
"""
{TEXT_2}
"""

Using these two sources, answer the following question: "{search_query}"
Use only the information in the above text to answer the question. Answer with only "Yes" or "No". Do not answer with anything other than "Yes" or "No".
```

Table 8: The prompt used to make predictions based on conflicting pairs of documents. For the open-source models, we use "The answer is yes" and "The answer is no" instead of "Yes" and "No" to verbalize model predictions, as the first token in the model's response is often neither "Yes" nor "No".

```
From a scale of one to five, how credible is this source? Do not answer with anything other than one, two, three, four, or five.
```

Table 9: The prompt used to ask models directly the credibility of a website.



Original excerpt

Praying Mantises Are Amazing Hunters
Praying mantises are a type of insect that is known for their unique way of hunting. They are able to turn their heads a full 180 degrees in order to get a better view of their prey. This allows them to strike quickly and accurately, which is how they are able to survive in the wild.
Similar to many other insects, the Mantis has three distinct bodies. The abdomen, head, and thorax are the three major organs in the body. The six sections of a mannitol bear appendages (e.g., antennae, mouth parts, and legs) all of which are clearly visible and distinguishable during examination.

Add more information

Praying Mantises Are Amazing Hunters
Praying mantises are a type of insect that is known for their unique way of hunting. They are able to turn their heads a full 180 degrees in order to get a better view of their prey. This allows them to strike quickly and accurately, which is how they are able to survive in the wild. The praying mantis has three distinct body parts: the abdomen, head, and thorax.
Similar to many other insects, the Mantis has three distinct bodies. The abdomen, head, and thorax are the three major organs in the body. The six sections of a mannitol bear appendages (e.g., antennae, mouth parts, and legs) all of which are clearly visible and distinguishable during examination. Praying mantises have incredible camouflage abilities that help them ambush prey.

Original excerpt

The phone on your desk still works. And it is secure—probably more secure than a cell phone. So instead of relying on email or a cell phone, verify strange requests with your landline—the old way.
So why are landline telephones almost extinct, particularly when it appears they may be the answer to many security issues? The telephone companies want to limit or remove them because the lines are so expensive to maintain. Thirteen states in the past three years, including Maine, have allowed telephone companies to stop providing traditional basic telephone services to consumers (i.e.landlines).

Rewrite to sound more objective

Traditional landline telephones continue to function but their use is declining as telephone companies limit or remove landline services. Landlines may offer more secure communication than cell phones as hackers have not targeted them as actively. However, 13 states have allowed telephone companies to end basic landline telephone services for consumers in recent years.

Figure 9: We use the claude-v1-instant model to make counterfactual edits to the text in order to change certain stylistic or relevancy features. Here, we show two examples of such perturbations.