

# A Reality Check on Context Utilisation for Retrieval-Augmented Generation

Lovisa Hagström<sup>✉</sup> Sara Vera Marjanović<sup>✉</sup>

Haeun Yu<sup>✉</sup> Arnav Arora<sup>✉</sup> Christina Lioma<sup>✉</sup>

Maria Maistro<sup>✉</sup> Pepa Atanasova<sup>✉</sup> Isabelle Augenstein<sup>✉</sup>

<sup>✉</sup>Chalmers University of Technology <sup>✉</sup>University of Copenhagen  
lovhag@chalmers.se

## Abstract

Retrieval-augmented generation (RAG) helps address the limitations of parametric knowledge embedded within a language model (LM). In real world settings, retrieved information can vary in complexity, yet most investigations of LM utilisation of context has been limited to synthetic text. We introduce DRUID (Dataset of Retrieved Unreliable, Insufficient and Difficult-to-understand contexts) with real-world queries and contexts manually annotated for stance. The dataset is based on the prototypical task of automated claim verification, for which automated retrieval of real-world evidence is crucial. We compare DRUID to synthetic datasets (CounterFact, ConflictQA) and find that artificial datasets often fail to represent the complexity and diversity of realistically retrieved context. We show that synthetic datasets exaggerate context characteristics rare in real retrieved data, which leads to inflated context utilisation results, as measured by our novel ACU score. Moreover, while previous work has mainly focused on singleton context characteristics to explain context utilisation, correlations between singleton context properties and ACU on DRUID are surprisingly small compared to other properties related to context source. Overall, our work underscores the need for real-world aligned context utilisation studies to represent and improve performance in real-world RAG settings.

## 1 Introduction

Retrieval-augmented generation (RAG) can be used to alleviate problems arising from imperfect parametric knowledge of language models (LMs), which may encode limited and potentially outdated information (Gao et al., 2024; Vu et al., 2024). However, the benefits of RAG are only realised if 1) the retrieval module retrieves helpful information and 2) the generative model successfully leverages the retrieved information. As a consequence, there have been many studies looking at the performance

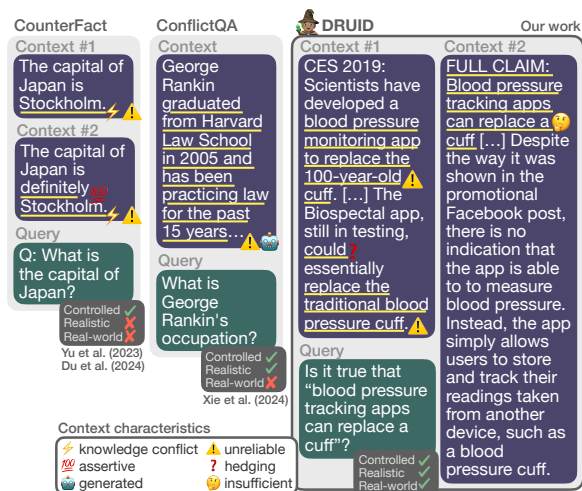


Figure 1: Datasets for context usage investigations.

and interaction of these two components and how to improve them (Gao et al., 2024).

However, existing research has mainly studied RAG in a disjoint manner, where studies of the quality and relevance of the retrieved information are detached from studies of LM context usage (Shi et al., 2023; Xie et al., 2023; Tan et al., 2024; Du et al., 2024). Hence, little is understood about 1) the characteristics of retrieved contexts and 2) their impact on LM context usage. Most notably, studies of LM context usage have leveraged controlled datasets using synthesised context to emulate a limited set of context characteristics (see Figure 1, left). For example, CounterFact and its variants are template-based, lending a *controlled* albeit very artificial and simplistic setup (Yu et al., 2023; Du et al., 2024). ConflictQA, on the other hand, is based on a mix of generated and retrieved contexts to study context usage in a more *realistic* setup with coherent and convincing contexts (Xie et al., 2024). Nevertheless, the scenarios described by these datasets are not representative of *real-world* RAG scenarios, as the context types do not reflect the diversity and complexity of the ones returned

by an actual retriever present in RAG (Longpre et al., 2021; Ravaut et al., 2024; Ortu et al., 2024).

This work studies context usage for RAG in real-world scenarios with real-world queries and context, as opposed to artificial samples. To this end, we focus on the prototypical information-seeking task of fact verification, where retrieving and utilising real-world information is vital. For the task, an agent is provided with a statement about the world – a *claim* – and needs to decide whether it is true or false using context retrieved from an external source – *evidence* (Guo et al., 2022). We take real fact-checked claims as ‘queries’ and the retrieved evidence as ‘context’ to evaluate RAG in this real-world setting, which naturally facilitates our goal of studying real-world context properties in RAG (Samarinas et al., 2021; Atanasova et al., 2022; Chrysidis et al., 2024; Glockner et al., 2024).

In particular, this work makes three main contributions. First, we introduce DRUID (Dataset of Retrieved Unreliable, Insufficient and Difficult-to-understand context) with real-world (query, context) pairs to facilitate studies of context usage and failures in real-world scenarios (§3). Second, we introduce a novel context-usage measure, *ACU*, which rectifies issues in previous measures. Thirdly, we highlight major differences between popular synthetic datasets and real-world data (DRUID): both in over-arching characteristics (§4), as well as how the provided context is used across different popular LMs (§5).

We show that synthetic datasets oversell the impact of certain context characteristics (e.g. knowledge conflicts), which are rare in retrieved data. Furthermore, synthetic data exaggerates the ‘context-repulsion’ seen for LMs, as we rarely see this behaviour in realistic data. Finally, we show that there is no singleton context characteristic (e.g. context length or perplexity) indicating RAG failure in real-world settings. Altogether, our work provides a reality check on LM context usage and points to the need for real-world aligned studies to fully understand and improve context utilisation for RAG. We also provide tools and resources to facilitate such studies.<sup>1</sup>

## 2 Related Work

**Claim Verification Datasets** Claim verification datasets typically measure an LM’s ability to assess

the veracity of a claim based on retrieved context (*evidence*). Importantly, the information requirements of this task can be challenging, resulting in retrieved evidence that is noisy (Samarinas et al., 2021; Atanasova et al., 2022; Fajcik et al., 2023; Chrysidis et al., 2024; Glockner et al., 2022; Warren et al., 2025). Furthermore, as LMs fine-tuned for claim verification have been shown to ignore evidence (Schuster et al., 2019, 2021), it is important to understand the causes of this behaviour. Therefore, with DRUID, we are the first to collect annotations for a range of ‘noisy’ characteristics of retrieved real-world contexts to assess how they affect LMs. Furthermore, unlike concurrent claim verification datasets, which either present artificial samples or a limited or less realistic scenario for context retrieval (Thorne et al., 2018; Augenstein et al., 2019; Diggelmann et al., 2020; Schlichtkrull et al., 2023), DRUID includes contexts *automatically retrieved* from the web to assess their impact on RAG, leading to a wide diversity of context properties including *insufficient* and *leaked* evidence. No other existing fact-checking datasets fulfil all of these properties (see Table 4 in the Appendix).

**Datasets for Context Usage Investigations** Two popular datasets used for context usage investigations are CounterFact and ConflictQA (Meng et al., 2022; Xie et al., 2024). These datasets contain synthesised queries based on fact triplets from LAMA (Petroni et al., 2019) (e.g. Thomas Ong-citizen of-Singapore) for which some contexts have been synthesised to induce *knowledge conflicts* by promoting answers in conflict with the parametric memory of the studied LM (e.g. ‘Pakistan’ as opposed to ‘Singapore’). The datasets have found widespread use for work on mechanistic interpretability and the evaluation of context utilisation (Meng et al., 2022; Geva et al., 2023; Ortu et al., 2024).

Similarly to CounterFact and ConflictQA, DRUID contains queries and corresponding contexts together with gold labels to facilitate evaluations of context utilisation. The main difference between DRUID and the two other datasets lies in how the queries and contexts were produced. For CounterFact and ConflictQA, the queries have been automatically synthesised based on WikiData subject-relation-object knowledge triplets (e.g. ⟨George Larkin, occupation, lawyer⟩ to produce “George Larkin is a lawyer.”) and the contexts have either been synthesised based on an edited knowledge triplet or generated by an LLM prompted to pro-

<sup>1</sup><https://github.com/copenlu/context-utilisation-for-RAG>

duce alternative context supporting some edited knowledge triplet. This makes it easy to infer the gold labels for the synthesised contexts, while it is not representative of a real-world context usage scenario. DRUID, on the other hand, is based on queries sampled from naturally occurring claims and contexts from the web, retrieved by automated retrieval methods representative of a real-world RAG setup.

### Impact of Context Characteristics for RAG

Work in information retrieval and RAG has identified several qualities in retrieved or synthesised contexts that impact context utilisation by humans and/or LMs. Retrievers typically provide overly long or corrupted text, which are *difficult to understand*, and impact LM output (Gao et al., 2024; Vladika and Matthes, 2023). Similarly, typos (Cho et al., 2024) and high perplexity (Gonen et al., 2023) have been identified as potential disruptors for RAG systems. Furthermore, *implicit* contexts, lacking an explicit connection to the query, have been identified as a prevalent failure cause in RAG (Li et al., 2024b). For automated retrieval situations, the rate of implicit contexts can be high due to chunking of text (Wang et al., 2024a). Instead, LMs have been shown to prefer context with high *query-context similarities* (Wan et al., 2024).

Most studies on RAG have focused on open-domain question answering (Kasai et al., 2023; Wu et al., 2024). Yoran et al. (2023); Shi et al. (2023) found that LMs are fragile to *irrelevant information* in the context, harming performance. Furthermore, in the case of *knowledge conflicts*, when context conflicts with parametric knowledge, LMs have been shown to ignore the conflicting context (Longpre et al., 2021), while other studies show that models prefer contextual information, as long as it is coherent and convincing (Xie et al., 2023). Sun et al. (2025) also connect knowledge conflicts to prediction uncertainty in fact-checking settings. Recently, Xu et al. (2024) have proposed more granular categories for knowledge conflicts, using *context-memory conflict* to denote the aforementioned phenomenon, and *inter-context conflict* to refer to different contexts contradicting each other. Marjanovic et al. (2024) further study *real-world knowledge conflicts* caused by dynamic facts, finding that RAG struggles the most with these.

*Unreliable* contexts have been studied by Chrysidis et al. (2024) in a fact-checking setup, for which misinformation is prevalent. This type of

Source	#claims	#samples	IAA
checkyourfact	220	890	0.77
science.feedback	220	913	0.64
factcheckni	109	429	0.50
factly	180	739	0.80
politifact	220	931	0.74
srilanka.factcrescendo	156	598	0.75
borderlines	224	990	0.53
Total	1,329	5,490	0.71

Table 1: Statistics for the DRUID dataset. IAA denotes inter-annotator agreement measured by Krippendorff’s alpha. science.feedback also includes claims from climate.feedback and health.feedback.

information is typically overlooked in more generic RAG QA setups, potentially because the retrieval corpora usually are based on Wikipedia or pre-curated datasets. *References to external sources* may convince a human reader of the credibility of some context, yet LMs seem to be impervious (Wan et al., 2024). However, expressed *certainty/uncertainty* in text and its impact on LM context usage has recently been studied by Du et al. (2024), where assertive contexts are found to be more convincing.

In our creation of DRUID we combine all these aforementioned insights to annotate naturally occurring context characteristics of interest.

## 3 DRUID

Previous studies of context utilisation leverage synthetic datasets with synthesised claims and contexts, ignoring the retrieval part in RAG (Yu et al., 2023; Xie et al., 2024). We develop the datasets DRUID (5,490 samples) and DRUID+ (48,517 samples) to enable studies of context utilisation for real-world scenarios. To this end, we collect *real-world claims from fact-checking sites* and use *automated retrieval to fetch corresponding evidence from the web*. DRUID is a high-quality subset of DRUID+ manually annotated for evidence relevance and stance. A DRUID sample consists of a ⟨claim, evidence, labels⟩ triple. More details on the dataset can be found in Table 1 and Appendix C.

### 3.1 Claim Collection

We sample claims verified by fact-checkers using Google’s Factcheck API.<sup>2</sup> We only sample claims in English. The claims are collected from 7 diverse

<sup>2</sup><https://developers.google.com/fact-check/tools/api/reference/rest>.

fact-checking sources, representing science, politics, Northern Ireland, Sri Lanka, the US, India, France, etc. All claims have been assessed by human fact-checkers. Further details on the claim collection can be found in Appendix D.

### 3.2 Evidence Collection

For each claim in DRUID and DRUID+, we retrieve up to 5 and 40 snippets of evidence, respectively. First, a gold-standard evidence document is retrieved from the original fact-checking site, which is the ‘summary’ of the fact-checking article written by the author of the article. For the remaining snippets of evidence, we use an automated retrieval method (Appendix D). We collect the top 20 search results for each of the Google and Bing search engines. The found webpages are then chunked into paragraphs and reranked by the Cohere rerank model.<sup>3</sup> Evidence corresponding to the top-ranked chunks is included in DRUID.

### 3.3 Relevance and Stance Annotation

Since the evidence is collected using automated retrieval, as opposed to controlled synthesis, we need to assess the relevance of the retrieved information to the claim, and, if it is relevant, what stance it represents (Wang et al., 2024c). For this, we crowdsource evidence-level annotations using Prolific<sup>4</sup> and Potato (Pei et al., 2022). Each evidence piece in DRUID is double annotated for *relevance* (*relevant* or *not relevant*) and *stance* to the claim (*supports*, *insufficient-supports*, *insufficient-neutral*, *insufficient-contradictory*, *insufficient-refutes* or *refutes*). More details on the annotation, guidelines and examples from the annotation interface can be found in Appendix M.

The annotator compensation was approximately 9 GBP/hour (the compensation was fixed for each task while the annotator completion time varied).

## 4 Context Characteristics

To understand the gap between the context provided in current diagnostic datasets for context usage and real RAG scenarios, we compare the characteristics present within our real-world dataset DRUID to the synthetic datasets CounterFact (Ortu et al., 2024) and ConflictQA (Xie et al., 2024). By virtue of their controlled setup, these and similar

datasets have seen much use for the study of context utilisation and mechanisms thereof (Jin et al., 2024; Du et al., 2024; Tan et al., 2024; Kortukov et al., 2024).

To ensure adequate comparison, we recast all samples in CounterFact and ConflictQA to a claim-evidence format (see Appendix E). This can be done without loss of information as all datasets represent a binary task for the LM (answer in alignment with the evidence or not). Furthermore, we show in Appendix F that the analysis of context utilisation and mechanisms thereof are unaffected by the format of the task being either answer completion or claim verification – the reformatting leads to no change in the mechanism employed by the model and its manipulation results.

In addition to the aforementioned datasets, we also present the characteristics of DRUID+ to better understand the impact of only collecting the top-ranked evidence for DRUID.

### 4.1 Detection of Context Characteristics

Several context characteristics impacting context utilisation by humans and/or LMs have been identified by previous work (Section 2). As opposed to synthesising contexts with certain properties, we *detect* those in existing datasets. Along with manual annotation of relevance and stance, we leverage automated methods. We experiment with two types of automated detection methods to assess context characteristics: 1) rule-based methods and 2) prompt-an-LLM methods. For the latter we zero-shot prompt the Cohere Command R+ model.<sup>5</sup>

Initial trials leveraging human annotations of context characteristics showed high annotator disagreements, potentially due to the subjective nature of some of the characteristics, and were consequently abandoned. Instead, we opted to operationalise the properties, as we further describe below. This allows us to explore more model-based measures of context characteristics, which can be expected to have a greater impact on model context usage vis-a-vis subjective human perception of the same characteristics.

**Relevance and stance** For DRUID, we use the manual relevance and stance annotations. For CounterFact and ConflictQA we infer those as follows. CounterFact contains counterfactual claims, for which the evidence is either the claim repeated (supports) or the claim but with the correct object

<sup>3</sup>rerank-english-v3.0 from <https://docs.cohere.com/v2/docs/rerank-2>.

<sup>4</sup><https://www.prolific.com/>

<sup>5</sup>command-r-plus (chat-only mode)



restored (refutes). For each ConflictQA entry, we have a model-generated claim and two types of evidence – *parametric memory aligned* (supports) or *counter memory aligned* (refutes).

**Claim-evidence similarity** This is measured using Jaccard similarity (see Appendix G), which outputs values between  $[0, 1]$ , where 1 signifies maximum similarity. The overlap of claim words with evidence words, scaled by the number of claim words (‘Claim-evidence overlap’) is also measured. In addition, we detect if the evidence repeats the claim verbatim (‘Repeats claim’).

**Difficult to understand** We measure the Flesch reading ease score, claim length (number of characters), evidence length (number of characters) and model context perplexities for our studied models (Llama 3.1 8B and Pythia 6.9B) to proxy how ‘difficult to understand’ is the context. Generally, we may consider samples that correspond to high model perplexities to be confusing to the model.

**Implicit** We detect named entities (NEs) in the claim and measure the overlap with entities found in the evidence (‘Claim entity overlap’). spaCy `en_core_web_trf` (based on RoBERTa-base) is used for the NE detection. Values are  $\in [0, 1]$  where 0 means that no NEs reappear in the evidence (maximum implicitness) and 1 means that all NEs were found in the evidence.

**Refers to external source** Command R+ is prompted to tell whether some evidence contains a reference to an external source or not (‘Detection by LLM’). Initial evaluation results show this detection method to align well with human annotations of the characteristic.

**Uncertain** We use a lexicon-based approach proposed by Islam et al. (2020) to detect hedge words and hedging discourse markers in the evidence to proxy ‘uncertain’ (‘Contains hedging’ and ‘Contains hedging discourse’). If a hedge word or hedging discourse marker is detected in the evidence, it is marked as ‘uncertain’ according to that method.

**Unreliable** We use manually curated lists by Media Bias/Fact Check<sup>6</sup> (MBFC) to automatically detect whether the evidence piece originates from a web page marked as using questionable sources, promoting conspiracy/pseudoscience or being a satire site (‘Unreliable source’). However, due to

<sup>6</sup><https://mediabiasfactcheck.com/>

the sparsity of the MBFC lists, we are unable to detect unreliability for all evidence in DRUID and DRUID+, lacking results for 26% and 34% of the samples, respectively. For CounterFact and ConflictQA there are no evidence sources to analyse.

**Additional characteristics** We check whether the evidence can be seen as directly pointing out a verdict by measuring whether the evidence contains the word ‘True’ or ‘False’. For the DRUID and DRUID+ datasets, we also record whether the evidence was published after the claim was made, as this allows to measure the occurrence of and effects of leaked information (‘Pub after claim’) (Schlichtkrull et al., 2023). Similarly, we measure whether the evidence comes from a fact-check webpage as this can be expected to contain additional leaked information (‘Fact-check source’) and whether it comes from the original fact-checking site summary (‘Gold source’).

## 4.2 Analysis of Context Characteristics

**Relevance and stance** Relevance and stance annotations for all datasets are shown in Tables 8 and 9 in the Appendix. Most contexts are annotated as relevant; however, given the more ambiguous nature of real-world queries, especially in claim verification, there is more variety in the kinds of stances presented by the context provided in DRUID: the majority of the automatically retrieved contexts (50%) do not have a clear stance or are not sufficient for addressing the query. This is the consequence of using automated retrieval, for which not even state-of-the-art methods based on commercial search engines and Cohere modules are capable of consistently retrieving ‘gold standard context’. Admittedly, the retrieval setup is used in a zero-shot fashion and performance may improve somewhat with additional fine-tuning, while it would not solve all insufficiency issues stemming from automated retrieval. Conversely, synthesised samples always assume sufficient context. Our results show a clear discrepancy between synthesised and real-world datasets, proving the need for real-world aligned datasets for studies of context usage.

**DRUID in comparison with other RAG datasets** The detected context characteristics for the synthetic datasets and DRUID are shown in Figure 2. More detailed results can be found in Appendix I. The synthetic CounterFact dataset has the highest Flesch reading ease scores, significantly shorter evidence lengths, frequent repetitions of the claim in

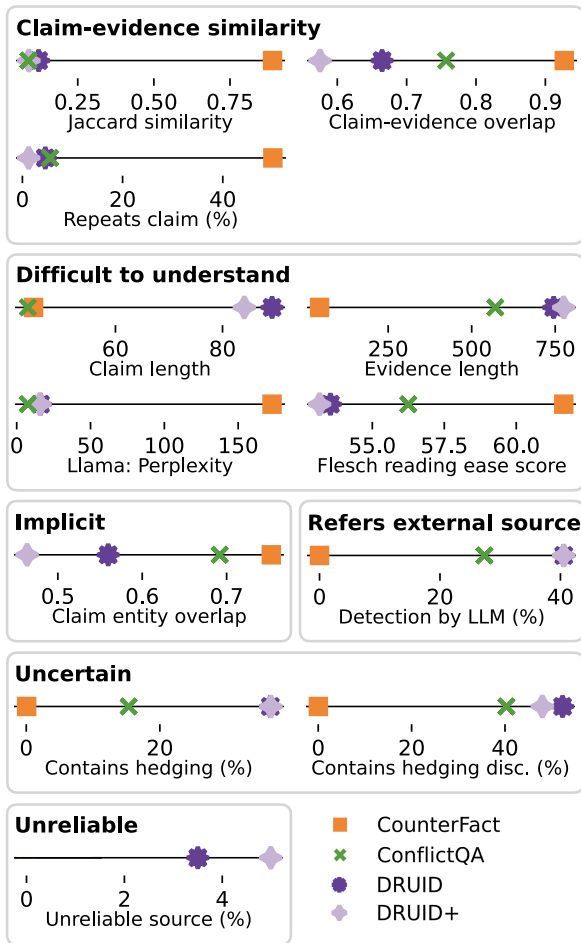


Figure 2: Average values for the context characteristics in CounterFact (Yu et al., 2023), ConflictQA (Xie et al., 2024) and DRUID datasets. The characteristics and their detection are described in Sections 2 and 4.1, respectively.

the evidence, significantly higher Jaccard similarity values and very few uncertainty markers relative to the other investigated datasets. CounterFact was designed to showcase simple knowledge conflict scenarios, causing much greater perplexity for both investigated LMs. In general, CounterFact is tailored to a specific type of context usage that is not indicative of real, retrieved context, as shown by the comparison to DRUID. The generated evidence for ConflictQA has characteristics more similar to those of DRUID. However, we can see that DRUID has much longer claims and evidence than seen in either of the other datasets; furthermore, there are more uncertainty markers and a greater degree of implicitness in the naturally occurring context in the DRUID dataset.

**Memory conflicts are less prevalent in real-world scenarios.** We measure context-memory

conflicts by comparing the parametric model prediction (no context or evidence provided) with the stance of the provided evidence. We have a conflict if the prediction and stance are either ‘Refutes’ or ‘Supports’ and do not align. For Llama 3.1 8B we record memory conflicts on 97.41% of the supporting evidence from CounterFact and on 71.16% of the refuting evidence from ConflictQA. For DRUID, we identify memory conflicts for 58.09% of the supporting evidence. Evidently, the rate of memory conflict is lower in real-world scenarios compared to artificial scenarios. More detailed results and the results for Pythia can be found in Table 21.

## 5 Context Utilisation

We aim to assess the transferability of insights based on synthesised scenarios to real-world scenarios. To this end, we evaluate and compare LM context utilisation results on synthetic datasets to results on DRUID.

### 5.1 Method

We measure the context utilisation of Pythia 6.9B and Llama 3.1 8B, two models from two model families widely used in RAG-evaluation studies (Biderman et al., 2023; Grattafiori et al., 2024; Ortu et al., 2024; Xie et al., 2024; Jin et al., 2024), on the CounterFact, ConflictQA and DRUID datasets. To measure context utilisation, the models are evaluated in two modes: 1) without evidence and 2) with evidence. In both modes, the models are prompted to assess the veracity of a given claim (True, False, or None), without and with evidence respectively. More details on the prompting can be found in Appendix J. We evaluate context utilisation using the softmaxed model logits, which we describe further in the next section. In the main paper, we only show results for supporting and refuting evidence; behaviour for all forms of ‘insufficient’ evidence (where ‘None’ is the expected model output) can be found in Appendices K and L.

### 5.2 Evaluation

There is no consistent measure for context usage across similar work; many studies look simply at changes in overall output distributions (Du et al., 2024; Marjanovic et al., 2024), which does not guarantee that the change is relevant to the provided context. Works in mechanistic interpretability often rely on logit differences for a specific token given evidence (Ortu et al., 2024; Yu et al., 2023), which are not normalised, do not factor in

desired change, and limit comparisons. Due to these issues, we introduce a novel measure (*ACU*), which 1) uses softmax-normalised probabilities, to ensure meaningful comparison, 2) focuses on probabilities of specific tokens, to ensure relevant change, and 3) scales these values by the amount of possible increase in probability. To measure context usage for a model  $M$ , we consider the re-scaled difference in salient token probability  $t \in T = \{\text{True}, \text{None}, \text{False}\}$  for a claim  $C$  between settings with and without evidence  $E$ , as follows.

$$\Delta P_M(t|C, E) = \begin{cases} \frac{P_M(t|C, E) - P_M(t|C)}{1 - P_M(t|C)} & \text{if } P_M(t|C, E) \geq P_M(t|C), \\ \frac{P_M(t|C, E) - P_M(t|C)}{P_M(t|C)} & \text{otherwise.} \end{cases} \quad (1)$$

Here,  $P_M(t|C)$  and  $P_M(t|C, E)$  denote the output probabilities for token  $t \in T$  by model  $M$  given a claim  $C$  and evidence  $E$ , respectively. The rescaling ensures that our metric is less sensitive to the original  $P(t|C)$  value. We expect high positive values of  $\Delta P_M(t|C, E)$  for  $t$  that align with the stance of  $E$  and the opposite for  $t$  that conflict with the stance. For example, given an evidence piece with the stance *refutes* we should ideally measure a high value for  $\Delta P_M(\text{False}|C, E)$  and low values for  $\Delta P_M(\text{True}|C, E)$  and  $\Delta P_M(\text{None}|C, E)$ .

We define a score of accumulated context usage (*ACU*) per sample  $\{C, E\}$  with stance  $S_E$  for a model  $M$  as follows.

$$\begin{aligned} \text{ACU}(C, E, S_E, M) &= \\ &= \frac{1}{|T|} \sum_{t \in T} D(t, S_E) \Delta P_M(t|C, E) \end{aligned} \quad (2)$$

$D(t, S_E)$  denotes the desirable change in  $\Delta P_M$  for maximum context usage, which is either  $\{-1, 1\}$ , depending on the annotated stance of the evidence. For example,  $D(\text{False}, \text{refutes}) = 1$ , whereas  $D(\text{True}, \text{refutes}) = D(\text{None}, \text{refutes}) = -1$ . This limits the range of *ACU* between  $[-1, 1]$ .

### 5.3 How do LMs utilise real-world retrieved context compared to synthesised context?

We inspect the context usage behavior of Pythia and Llama on CounterFact, ConflictQA and DRUID to understand how LMs utilise real-world context compared to synthetic contexts. Accumulated context usage scores (Equation (2)) can be found in

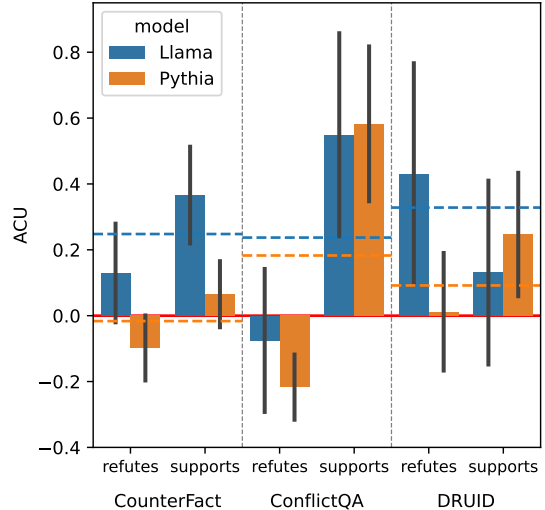


Figure 3: *ACU* (Equation (2)) for each model and dataset. The error bars indicate the standard deviation. Negative *ACU* values indicate ‘context-repulsion’: changes in probability away from the annotated evidence stance. The dashed horizontal lines indicate average *ACU* scores for each model and dataset.

Figure 3. See Appendix K for more granular context usage results. We structure the analysis around a set of main findings, listed below.

#### Synthetic datasets suggest an over-preference of supporting evidence.

While we see variations in model behaviour for our two synthetic datasets, ConflictQA and CounterFact, there are also some over-arching similarities: context utilisation is much greater for supporting evidence. In the case of refuting evidence, we often see negative *ACU* scores, indicating ‘context-repulsion’, changes in probability *away* from the stance of the provided context, which indicates low robustness; this is strongest for ConflictQA, which also has the greatest *ACU* scores for supporting context. This may be the consequence of the ConflictQA claims having been generated by Llama 2 7B and some of the supporting contexts having been generated by ChatGPT (they are *aligned with parametric memory*, which has been shown to increase context utilisation (Xie et al., 2024; Tan et al., 2024)). However, this preference for supporting evidence is also seen in CounterFact, which is surprising as the refuting evidence should align with LM parametric memory. This may be explained by the synthetic and confusing nature of CounterFact samples, leading to high model perplexities (See Figure 2.)

We see a different behaviour with our real-world dataset DRUID: we rarely see context repulsion, and

we see reduced ACU scores for supporting evidence. These lower ACU scores for supporting evidence may stem from the lack of generated context (vis-a-vis ConflictQA), and the increased ACU for refuting evidence may be due to the lower perplexities of the context (vis-a-vis CounterFact). This highlights the need for real-world contexts for studies of context utilisation: automatically generated contexts, by automated template-filling or LLM generation, inevitably induce properties that interfere with studies of context utilisation.

### Different models show different context usage.

Notably, Llama and Pythia behave very differently on all datasets studied. Potentially, this owes to CounterFact having been designed to elicit knowledge conflicts in Pythia and ConflictQA having been based on claims generated by Llama 2 7B. However, DRUID has not been customised to a specific model and results on this dataset clearly show how context usage varies across models. Moreover, we find that Llama on average is more faithful to the contexts of all datasets (and demonstrates less context-repulsion), yet remains understudied in context-utilisation studies (Ortu et al., 2024; Du et al., 2024). These results are further corroborated in Hagström et al. (2025), for which multiple LMs of different model family and size are benchmarked for their context utilisation.

### 5.4 Does LM context usage depend on characteristics of the evidence/context?

We evaluate the influence of different context characteristics (see §4.1) on model context usage. For this, we calculate Spearman correlations between each context property and our context usage metric, ACU (Equation (2)), stratified by the evidence stance for each dataset. The results for Llama are shown in Figure 4. Results for Pythia, insufficient evidence from DRUID and additional fine-grained correlation results can be found in Appendix L. While we see a limited effect of any one characteristic, we highlight overarching findings below.

**Context from fact-check sources have greater ACU scores.** Llama and Pythia are more likely to be faithful to refuting context from a fact-checking source. Most likely, this general property captures a characteristic not fully captured by the more fine-grained detection methods. Previous manual inspection of fact-check articles indicates a higher rate of assertive and to-the-point language, which may explain these observations. Moreover, the fact-

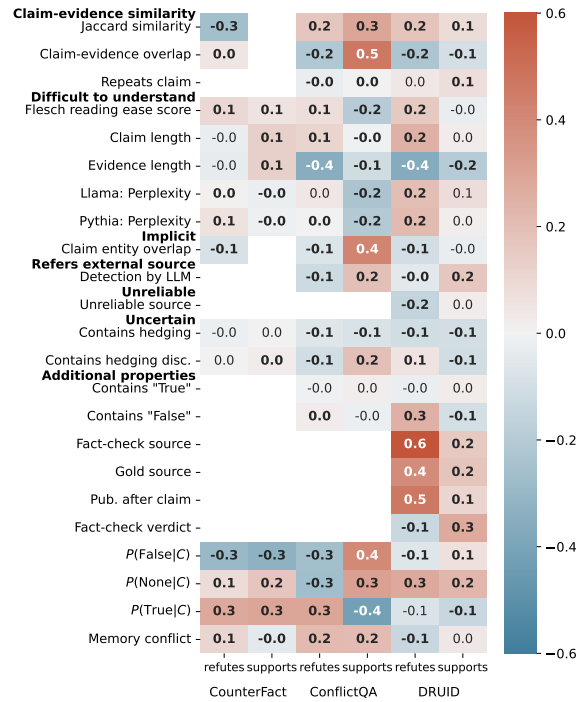


Figure 4: Spearman correlations between context usage measured by ACU (Equation (2)) and different context characteristics for Llama. Significant correlation values (p-value < 0.05) are marked in bold.

check articles are more likely to directly discuss the claims in a one-hop manner with multiple arguments, possibly making them more convincing to the LM. We hypothesise that a similar case holds for ‘Pub. after claim’ and ‘Gold source’. Similar results are observed for Pythia. The restricted generation process of the synthetic datasets makes similar investigations impossible.

Previous studies of context utilisation have focused on the effects of singular context characteristics in isolation. Our work hints at the relevance of including aggregates of features in the analysis, as these may better explain context utilisation in real-world scenarios. For example, contexts from a certain source, like fact-checking articles, may express aggregates of features.

**References to external sources show low correlations with ACU.** We measure low correlations with whether the evidence refers to an external source on both ConflictQA and DRUID. Our results on ConflictQA and DRUID show slightly greater importance of references for supporting contexts, while the values measured are fairly low. Using a synthetic dataset, Wan et al. (2024) found LM context usage to be insensitive to references to external sources. Our results on the real-world DRUID



dataset support this conclusion.

**Correlations with claim-evidence similarity properties are low for DRUID.** For ConflictQA we measure the highest correlation with context usage for claim-evidence overlap for supporting contexts. The same does not hold for DRUID. Previous work by [Tan et al. \(2024\)](#) on controlled context-conflicting datasets found LMs to prioritise contexts with high similarities between query and context. Our results indicate that real-world queries and contexts come with a greater complexity for which context usage cannot be predicted solely based on query-context similarity.

**LMs are less faithful to long contexts.** Llama is less likely to be faithful to long refuting contexts on both DRUID and ConflictQA. We note how the results do not generalise to CounterFact, which might be explained by the more synthetic nature of CounterFact compared to ConflictQA.

## 6 Conclusion

In this work, we ground studies of context utilisation to real-world RAG scenarios. We develop DRUID and compare it to synthesised datasets previously used to study context-utilisation. DRUID is a claim-verification dataset which contains naturally occurring claims and manually annotated evidence automatically retrieved from the web. We find fundamental differences in dataset characteristics between DRUID and synthetic datasets (CounterFact and ConflictQA). We also introduce a novel ACU score to consistently measure context utilisation across LMs and datasets. On DRUID, correlations between singleton context properties and ACU are surprisingly small compared to other properties related to context source (e.g. contexts coming from specific types of websites). We hypothesise that, rather than singleton features, this owes to an aggregation of several characteristics contributing to context usage. This suggests the common factors impacting RAG success are broader than previously expected, and further work needs to be done to identify fine-grained causes of RAG failure. Furthermore, given the use of synthetic datasets to identify mechanistic components of context usage ([Ortu et al., 2024](#); [Yu et al., 2023](#)), our results call into question the generalisability of the findings. With DRUID, we provide resources that better facilitate mechanistic and behavioural studies of context usage in real-world scenarios.

## 7 Limitations

Our work leverages claim verification as a vehicle for studies of realistic context utilisation. It is not fully clear whether insights related to context utilisation on this task will transfer to other RAG tasks, such as question-answering, which is overly represented in RAG evaluations. However, claim verification is a complex information-seeking task and we expect other tasks to have a large overlap or subset of properties with it. For example, as seen in this work, the QA format for CounterFact and ConflictQA is easily recast as a claim verification task. Furthermore, we show that intervention methods developed for QA tasks easily transfer to the same datasets when recast to a claim-verification setting in Appendix F. This suggests that some findings can be generalisable across tasks. Future work could expand on this work to *incorporate other RAG-specific tasks* to better understand the generalisability of context utilisation behaviours.

In our creation of DRUID we leverage an automated retrieval method based on commercial search engines and the Cohere Rerank model. While this method builds on state-of-the-art developments within the field of information retrieval, there are many other methods and tools we could have chosen, which could impact the context characteristics and model behaviour ([Wang et al., 2024b](#); [Katsimpras and Paliouras, 2024](#); [Chen et al., 2024](#)). *A comprehensive comparison of different retrieval methods* and their impact on context utilisation would be an interesting direction for future work.

In our creation of DRUID, we ensure to source claims from many different fact-checking sites to increase the representation of our dataset to the entire English-speaking world. However, it is not a uniform distribution, and the amount of context gathered per claim as well as the inter-annotator agreement for the context stances differs across claim sources. This could be due to unintentional cultural biases within our retrieval system or our annotators. Future work could investigate the impact of these *cultural biases in the retrieval process on model output*. DRUID, given its wide distribution of claim and evidence sources, would be an excellent dataset for such an investigation.

DRUID is based on a fact-checking task, meaning that it covers a limited set of domains. This may affect the conclusions based on the dataset, compared to other datasets situated in different domains.

Meanwhile, it is worth noting that no dataset is free of this problem – both CounterFact and ConflictQA have a limited scope by only focusing on WikiData triplets. DRUID is in comparison to these datasets better at covering different domains. Other more realistic datasets would also be limited by the characteristics of the domains they are situated in. For future work it would be interesting to *investigate whether one can control for and disentangle domain effects on context utilisation*, something the DRUID dataset should be useful for.

While we investigate the impact of many characteristics on context utilisation, it is not exhaustive. Future work could look into the impact of *other context characteristics on context utilisation*. For example, our study and dataset omit interesting context characteristics related to propaganda, simplified or manipulated content, anecdotal, mix of languages, multimodality, multi-hop reasoning, preciseness etc. (Piskorski et al., 2023; Wan et al., 2021; Jiang et al., 2020; Dufour et al., 2024). These properties would be relevant to study in future work.

In this work, we study the context utilisation behaviours of Llama 3.1 8B and Pythia 6.9B, two popular LMs used for RAG-evaluation studies. With this selection, we represent two families of models and can already reveal great disparities in context utilisation between synthetic and real-world datasets. For future work, it would be interesting to further investigate the context utilisation of *more model families and different model sizes*. All future studies are well-facilitated by the dataset and evaluation framework we introduce in this work.


While we include a comprehensive correlation analysis to identify the dependence between our studied characteristics and context usage, it does not give any information about causality. Future work could include a more *comprehensive causal analysis*. A causal analysis is necessary to fully understand the effects of different context characteristics on context utilisation (Feder et al., 2022). Given that our findings indicate that context utilisation cannot be predicted by one singleton characteristic, there are likely many potential confounders within DRUID, and all real, retrieved text. Future work on this could take inspiration from the studies by Gui and Veitch (2023). While our work provides a good starting point for RAG evaluations of context characteristics, our findings show that more work is needed to fully understand the complex behaviours governing context usage.

## 8 Ethical Considerations

Our work concerns the evaluation of RAG-based models on veracity prediction in a real-world setting. In the creation of the dataset, while we tried to maintain representativeness of the real world by including sources of data from different parts of the world, we introduced biases by selecting only English language sources. Consequently, our results only stand for claims and corresponding evidence sentences in English. For the annotation tasks, we do not retain any information about the annotators and pay them a fair wage as determined by the annotation platform. We also informed the annotators about how their data would be used and received their consent. However, for ease of understanding the subject matter and increasing chances of agreement, we screened the annotator pool to only include participants with at least an undergraduate degree, English fluency, no language-related disorders, and UK, US or Irish nationality. While this helped achieve higher-quality annotations, it limits the perspectives embedded in the dataset and may reinforce cultural biases, which we acknowledge as a potential risk.

Otherwise, we do not foresee any pressing potential risks with this work. We performed foundational research focused on evaluation, which should come with few implications for malicious use, environmental impact, security violations, etc.

## Acknowledgements

 This research was co-funded by the European Union (ERC, ExplainYourself, 101077481), by the Pioneer Centre for AI, DNRF grant number P1, as well as by The Villum Synergy Programme. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The research was also co-funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, as well as by WARA Media and Language, also a part of WASP. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis, partially funded by the Swedish Research Council through grant agreement no. 2022-06725. The work was also supported by compute credits from a Cohere

For AI Research Grant. Lastly, we would like to thank the anonymous reviewers for their helpful suggestions.

## References

- Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. [Where is your evidence: Improving fact-checking by justification modeling](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. [Fact checking with insufficient evidence](#). *Transactions of the Association for Computational Linguistics*, 10:746–763.
- Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. [MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usven Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.
- Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. [Generating literal and implied subquestions to fact-check complex claims](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. [Dense X retrieval: What retrieval granularity should we use?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.
- Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C. Park. 2024. [Typos that broke the RAG’s back: Genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2826–2844, Miami, Florida, USA. Association for Computational Linguistics.
- Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. [Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking](#). *arXiv preprint arXiv:2404.18971*.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leipold. 2020. [Climate-fever: A dataset for verification of real-world climate claims](#). *Preprint*, arXiv:2012.00614.
- Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. [Context versus prior knowledge in language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.
- Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Duffield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. [Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild](#). *Preprint*, arXiv:2405.11697.
- Martin Fajcik, Petr Motliceck, and Pavel Smrz. 2023. [Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10184–10205, Toronto, Canada. Association for Computational Linguistics.
- Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. [Generating fact checking briefs](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7147–7161, Online. Association for Computational Linguistics.
- Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. [Causal inference in natural language processing: Estimation, prediction, interpretation and beyond](#). *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. [Dissecting recall of factual associations in auto-regressive language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Toronto, Canada. Association for Computational Linguistics.



Singapore. Association for Computational Linguistics.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. [Missing counter-evidence renders NLP fact-checking unrealistic for misinformation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. [AmbiFC: Fact-Checking Ambiguous Claims with Evidence](#). *Transactions of the Association for Computational Linguistics*, 12:1–18.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. [Demystifying prompts in language models via perplexity estimation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Jun-teng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew

Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stéphane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Sweet, Gil Halpern, Grant



- Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojuan Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Lin Gui and Victor Veitch. 2023. Causal estimation for text data with (apparent) overlap violations. In *The Eleventh International Conference on Learning Representations*.
- Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. [A survey on automated fact-checking](#). *Transactions of the Association for Computational Linguistics*, 10:178–206.
- Lovisa Hagström, Youna Kim, Haeun Yu, Sang goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2025. [Cub: Benchmarking context utilisation techniques for language models](#). *Preprint*, arXiv:2505.16518.
- Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. [A richly annotated corpus for different tasks in automated fact-checking](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.
- Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. [CHEF: A pilot Chinese dataset for evidence-based fact-checking](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.
- Jumayel Islam, Lu Xiao, and Robert E. Mercer. 2020. [A lexicon-based approach for detecting hedges in informal text](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113, Marseille, France. European Language Resources Association.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojuan Jiang, Kang Liu, and Jun Zhao. 2024. [Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.
- Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and

- Kentaro Inui. 2023. [Realtime QA: What’s the answer right now?](#) In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Georgios Katsimpras and Georgios Paliouras. 2024. [GENRA: Enhancing zero-shot retrieval with rank aggregation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7566–7577, Miami, Florida, USA. Association for Computational Linguistics.
- Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. [WatClaimCheck: A new dataset for claim entailment and inference](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.
- Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. [Studying large language model behaviors under context-memory conflicts with real documents](#). In *First Conference on Language Modeling*.
- Bryan Li, Samar Haider, and Chris Callison-Burch. 2024a. [This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.
- Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. [Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. [DYNAMICQA: Tracing internal knowledge conflicts in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.
- Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. [Competition of mechanisms: Tracing how language models handle facts and counterfactuals](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.
- Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. [POTATO: The portable text annotation tool](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. [SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.
- Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. [On context utilization in summarization with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.
- Chris Samarin, Wynne Hsu, and Mong Li Lee. 2021. [Improving evidence retrieval for automated explainable fact-checking](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.
- Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. [Averitec: A dataset for real-world claim verification with evidence from the web](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

- Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.
- Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. [Towards debiasing fact verification models](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org.
- Jingyi Sun, Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. [Explaining sources of uncertainty in automated fact-checking](#). Preprint, arXiv:2505.17855.
- Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. [Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Juraj Vladika and Florian Matthes. 2023. [Scientific fact-checking: A survey of resources and approaches](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. [Fresh-LLMs: Refreshing large language models with search engine augmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Alexander Wan, Eric Wallace, and Dan Klein. 2024. [What evidence do language models find convincing?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.
- Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. [A DQN-based approach to finding precise evidences for fact verification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1030–1039, Online. Association for Computational Linguistics.
- Kexin Wang, Nils Reimers, and Iryna Gurevych. 2024a. [DAPR: A benchmark on document-aware passage retrieval](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4313–4330, Bangkok, Thailand. Association for Computational Linguistics.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024b. [REAR: A relevance-aware retrieval-augmented framework for open-domain question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626, Miami, Florida, USA. Association for Computational Linguistics.
- Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024c. [Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.
- Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. [Show me the work: Fact-checkers’ requirements for explainable automated fact-checking](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are rag models? quantifying the tug-of-war between rag and llms’ internal prior. *arXiv preprint arXiv:2404.10198*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in](#)



knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. [Characterizing mechanisms for factual recall in language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

## A Computational resources

All models are evaluated without fine-tuning on one A40 Nvidia GPU per evaluation. The total computational budget for the evaluation was about 100 GPU hours.

## B Use of AI assistants

AI assistants like Copilot and ChatGPT were intermittently used to generate template code and rephrase sentences in the paper, etc. However, no complete paper sections or code scripts have been generated by an AI assistant. All generated content has been inspected and verified by the authors.

## C DRUID

Dataset statistics for DRUID+ can be found in Table 2 and statistics for claims with inter-context conflicts can be found in Table 3. Inter-context conflicts are measured based on the annotated stance of each evidence piece; if we find evidence pieces with different stance (‘Refutes’ or ‘Supports’) for the same claim, we mark the claim as having conflicting evidence.

A comparison of DRUID to other fact-checking datasets can be found in Table 4. FEVER is marked as *synthetic* and not representative of realistic scenarios for context retrieval as the samples in the dataset have been artificially generated by the following process: 1) take a random sentence from Wikipedia, 2) give this sentence to an annotator and ask them to make up a set of claims based on

the sentence, 3) ask annotators to produce additional mutations of the claims and 4) fetch matching contexts mainly from the matching Wikipedia pages. This process is not representative of a realistic use case for context augmentation, as the claims do not match real information needs and the evidence is sourced from nearly perfect Wikipedia page matches. VitaminC and SciFact have been synthesised in a similar fashion, and are therefore also marked as synthetic.

Source	#claims	#samples
checkyourfact	300	6,653
climate/health/science.feedback	293	6,983
factcheckni	137	3,124
factly	299	6,443
politifact	300	7,954
srilanka.factcrescendo	173	4,093
borderlines	503	3,124
Total	2,005	48,517

Table 2: Statistics for the DRUID+ dataset.

Source	#claims	#confl. claims
checkyourfact	220	112
climate/health/science.feedback	220	77
factcheckni	109	25
factly	180	65
politifact	220	70
srilanka.factcrescendo	156	61
borderlines	224	41
Total	1,329	451

Table 3: Inter-context conflict statistics for the DRUID dataset (Xu et al., 2024). ‘#confl. claims’ denotes the number of claims for which we find inter-context conflicts, i.e. conflicting evidence pieces for which at least one evidence piece supports the claim and at least another refutes it.

## D Dataset creation

### D.1 Claim collection

Different fact-check organisations use different notations for the fact-check verdicts, ranging from ‘Pants on Fire’ to ‘Inaccurate’ (Augenstein et al., 2019). We only collect claims for which the verdict could be mapped to ‘True’, ‘False’ or ‘Half-true’ (See Appendix D). We collect claims from 7 fact-checking sources, with varying themes and countries of origin:

- checkyourfact.com
- (climatefeedback.org, healthfeedback.org, science.feedback.org)
- factcheckni.org
- factly.in



Dataset	Claim		Evidence		
	Source	Type	Sufficient	Unleaked	Retrieved
FEVER (Thorne et al., 2018)	W	Synthetic	✓	N/A	✓
VitaminC (Schuster et al., 2021)	W	Synthetic	✓	N/A	✓
SciFact (Wadden et al., 2020)	S	Synthetic	✓	N/A	✓
Liar-Plus (Alhindi et al., 2018)	FC	Real	✓	✗	✗
MultiFC (Augenstein et al., 2019)	FC	Real	✗	✗	✓
WatClaimCheck (Khan et al., 2022)	FC	Real	✗	✓	✗
ClaimDecomp (Chen et al., 2022)	FC	Real	✗	✓	✗
Snopes (Hanselowski et al., 2019)	FC	Real	✗	✓	✗
QABrief (Fan et al., 2020)	FC	Real	✗	✓	✗
CHEF (Hu et al., 2022)	FC	Real	✓	✗	✓
AVeriTeC (Schlichtkrull et al., 2023)	FC	Real	✓	✓	✓
Factcheck-Bench (Wang et al., 2024c)	LLM	Real/Synthetic	✓✗	✓	✓
DRUID	W, FC	Real	✓✗	✓✗	✓

Table 4: Comparison of related fact-checking datasets. *Source* indicates where the claims are collected from, such as Wikipedia, Fact-Checking articles, Scientific sources or LLM responses. *Type* indicates whether the claims are synthetic or real-world. *Sufficient* indicates whether the evidence can provide sufficient information. *Unleaked* means whether the evidence contains leaks from the future. *Retrieved* denotes whether the dataset involves evidence retrieval instead of relying on pre-retrieved passages e.g. the fact-checking article. ✓✗ indicates that both properties can be found and are annotated for.

- politifact.com
- srilanka.factcrescendo.com
- borderlines from Li et al. (2024a)

The method for sampling the claims is adapted to balance out the precedence of False and US-centric fact-checked claims. To this end, we sample the claims to ensure, to the extent possible, (1) an even distribution across the 7 fact-checking sources, (2) an even distribution across True, False and Half-true claims, and (3) an even distribution of claims posted before and after 2023 (to ensure we also obtain claims and evidence unlikely to be present in the assessed LM’s training data). For the sampling, we first prioritise (1), followed by (2) etc. Due to a shortage of some claims, we cannot achieve completely uniform distributions.

We remove all claims that mention a ‘photo’ or a ‘video’ to limit fact-verification to a single media.

## D.2 Claim veracity mappings

We map the claim veracity labels to ‘True’, ‘False’ or ‘Half-true’ as shown in Table 5.

## D.3 Automated evidence retrieval

Given a claim, the method is as follows:

**Use search engines to search the web for relevant web pages.** Fetch the top 20 search results for the claim from the Google and Bing search engines,

Our label	Incoming label
True	True
	TRUE
	ACCURATE
	ACCURATE WITH CONSIDERATION
	Correct
	Mostly accurate
	Accurate
Half-true	Half True
	PARTLY TRUE
	Correct But...
	Mostly_Accurate
	Partially correct
False	False
	FALSE
	MISLEADING
	Misleading
	Inaccurate
	Incorrect, Flawed_Reasoning
	INACCURATE
	INACCURATE WITH CONSIDERATION

Table 5: The claim veracity label mappings used for the creation of DRUID and DRUID+. Claims corresponding to verdict labels not listed in the table are dropped.

respectively.<sup>7</sup> The results are de-duplicated as they may overlap. From this step on, no regard is paid to the search engine ranks, while they are stored

<sup>7</sup>We used their respective APIs [customsearch.googleapis.com/customsearch/v1](https://customsearch.googleapis.com/customsearch/v1) and [api.bing.microsoft.com/v7.0/search](https://api.bing.microsoft.com/v7.0/search). Search results were retrieved in October 2024.

for potential future use cases.

**Chunk the content of each web page.** The search results consist of full web pages, for which not all page content can be expected to be immediately useful for the claim verification. Similarly to [Diggelmann et al. \(2020\)](#) we use an extractive approach based on chunking to get concise evidence that fits into the model context window. Each paragraph on the web page forms a chunk if it contains fewer than 200 words. Paragraphs longer than 200 words are split into multiple chunks of up to 200 words. This approach is based on manual tuning and inspection of some retrieved evidence.

**Get reranker scores for the chunks.** The search engines provide a quite coarse filter for relevant information with high recall but low precision. Moreover, the search engines cannot extract the relevant snippets from the search results. To get more precise and accurate retrieved contexts we use a reranker ([Diggelmann et al., 2020](#)). Specifically, we use the Cohere Rerank model<sup>8</sup> to get reranker scores for each chunk with respect to a claim.

To avoid claim repeats in the evidence, we also filter out sentences from paragraphs corresponding to  $RougeL(sentence, claim) > 0.8$  in the chunking step (step 3). Otherwise, the Cohere Rerank model was prone to fetch evidence that more or less only repeated the claim.

**Select web pages for evidence retrieval.** For DRUID we have a limited annotation budget and therefore select the four web pages for which we record the maximum reranker chunk scores and collect evidence from each of these. To represent the situation of not having access to fact-checking articles published after the claim was made, we adapt this selection to collect at minimum two pieces of evidence posted before the publication of the claim. This way, we ensure that at least half of the dataset contains unlearned information. For DRUID+ we select all webpage search results for the evidence collection.

**Collect evidence from the selected web pages.** Collect an evidence piece from each of the web pages selected in the previous step. This is done by aggregating the three top-ranked chunks from the web page via simple concatenation. If necessary, the number of chunks is decreased to ensure that

<sup>8</sup>rerank-english-v3.0 from <https://docs.cohere.com/v2/docs/rerank-2>.

no evidence piece is longer than 300 words. As a result, we have several pieces of evidence per claim, each representative of one web page.

## E Additional dataset details

### E.1 CounterFact

Column	Value
Claim	Geoffrey Hinton is employed by BBC.
Verdict	False
Evidence #1	Geoffrey Hinton is employed by BBC.
Relevant	True
Evidence stance	Supports
Evidence #2	Geoffrey Hinton is employed by Google.
Relevant	True
Evidence stance	Refutes

Table 6: A sample from CounterFact that has been recast to match the format of DRUID.

The CounterFact dataset referred to in this paper has been developed by [Ortu et al. \(2024\)](#) to study context usage under knowledge conflicts. It contains 10,000 samples based on fact triplets from WikiData. An example of a sample from the CounterFact dataset is “Redefine: Geoffrey Hinton is employed by BBC. Geoffrey Hinton is employed by”. A knowledge conflict is induced by the replacement of the correct answer (*Google*) with *BBC* in the context. We use the CounterFact split based on Pythia 6.9B.

To ensure alignment between the investigations for CounterFact and DRUID we first recast the CounterFact samples to a format that aligns with the DRUID dataset. This is exemplified in Table 6. The queries are recast to claims and we retain both the new knowledge conflicting context as well as the original correct context as evidence. By virtue of the synthetic nature of the dataset, we know beforehand that all claims are incorrect and that the new contexts support the claims. The opposite holds for the original, correct, contexts. We also know that all contexts are relevant to the claims.

### E.2 ConflictQA

We also inspect the ConflictQA dataset developed by [Xie et al. \(2024\)](#). The dataset contains ‘memory answers’ from an LM (based on its parametric memory) to prompts from PopQA (also based on WikiData fact triplets) together with ‘counter answers’ generated by an LLM instructed to produce

Column	Value
Memory answer	George Rankin is a lawyer.
Memory aligned evidence	George Rankin graduated from Harvard Law School in 2005 and has been practicing law for the past 15 years. He is a member of the American Bar Association and has been recognized as a leading lawyer in the field of intellectual property law by several prestigious legal publications. In addition, he currently serves as a partner at one of the top law firms in the country.
Counter memory aligned evidence	George Rankin Major General George James Rankin, (1 May 1887 - 28 December 1957) was an Australian soldier and politician. He served in both the House of Representatives and the Senate, representing the Country Party of Australia. Rankin was born at Bamawm, Victoria, the tenth child of Irish farmer James Rankin and Sarah, né Gallagher. He attended the local state school and became a farmer. In 1907, he joined the Militia, and was commissioned in the 9th Light Horse Regiment in 1909. He married Annie Isabella Oliver at Rochester, Victoria on 7 July 1912. In 1914, he was appointed a

Table 7: A sample from the ConflictQA dataset.

an answer that conflicts with the model answer. Each entry also contains corresponding evidence, one that is ‘parametric memory aligned’ and another that is ‘counter memory aligned’. These evidence pieces have been generated or sourced from Wikipedia/human annotation. We use the ConflictQA split based on Llama 2 7B. An example from the ConflictQA dataset can be found in Table 7.

We use the ‘memory answer’ (generated by Llama 2 7B) as the claim and the ‘parametric memory aligned evidence’ and ‘counter memory aligned evidence’ as supporting and refuting evidence corresponding to the claim.

The generated origin of the evidence is revealed at multiple instances (see below). Moreover, the generated evidence is many times directly on point which cannot be expected to be found in real-world scenarios.

## F Attention manipulation results on CounterFact after Reformatting

Figure 5 shows the results of pruning attention heads in Pythia for the original sentence completion task as studied by Ortu et al. (2024) compared to the same approach but recast to a claim verification task. The effects of attention head pruning are largely unaffected by the reformatting to a claim verification task, showing that LMs can be interpreted and manipulated for the claim verification task just as well as for the sentence completion task.

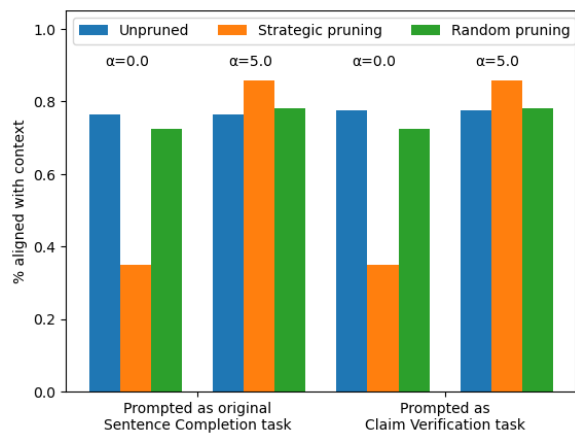


Figure 5: The results of pruning attention heads in Pythia for the original sentence completion task and for when the task has been recast to a claim verification task.

## G Jaccard similarity to proxy claim-evidence similarity

We use Jaccard similarity to proxy claim-evidence similarity. This is measured as follows.

$$J(C, E) = \frac{|W(C) \cap W(E)|}{|W(C) \cup W(E)|} \quad (3)$$

$W$  denotes the set of unique words, lowercased and ignoring punctuation or special characters like ‘-’, found in a claim  $C$  or evidence  $E$ .

## H Cohere: Refers to external source

The prompt used for the detection of references to external sources with Cohere Command R+ is as follows: “Does the following text refer to an external source or not? Admissible external sources are for example ‘a study’, ‘[1]’, ‘the BBC’, a news channel etc. Answer with a ‘Yes’ or ‘No’.\n\nText: <text>”.

Relevant	CounterFact	ConflictQA	DRUID
True	20,000	16,046	5,399
False	0	0	91

Table 8: Evidence relevance for each of the investigated datasets.

Evidence stance	CounterFact	ConflictQA	DRUID
refutes	10,000	8,023	1,760
insufficient	0	0	2,730
-refutes	0	0	557
-contradictory	0	0	410
-neutral	0	0	1,078
-supports	0	0	685
supports	10,000	8,023	909

Table 9: Evidence stance for each of the investigated datasets.

## I Context characteristics

The full statistics on the context characteristics for all datasets considered can be found in Tables 8 to 10.

## J Prompts

For each mode and model, we manually tune a prompt on 390 samples from DRUID to maximise context usage, using the balanced mean absolute error<sup>9</sup> as the objective function to be minimised. For mode (1) the gold labels are given by the claim veracity and for (2) the annotated evidence stances. We experimented with around 21 different prompts (0-shot, 2-shot and 3-shot) in total. The best-performing prompts were found to be 3-shot. All prompts request the model to say whether the claim, with or without evidence, is ‘True’ or ‘False’. Moreover, the model is instructed to respond ‘None’ if it is uncertain or cannot answer.

The tuned prompts used to evaluate Llama and Pythia can be found in Tables 12 to 14. The tables also list the prompts used for the 0-shot experiments described in the appendix.

## K Additional context usage results

Some cherry- and lemon-picked samples from the investigated datasets and corresponding model predictions can be found in Tables 15 to 20. Additional ACU scores for insufficient evidence from DRUID can be found in Figure 6c. We also investi-

<sup>9</sup>We used sci-kit learn’s `mean_absolute_error` with sample weights given by `compute_sample_weight` for the ‘balanced’ setting.

gate model context usage under zero-shot prompts. The context usage results for Llama and Pythia can be found in Figures 6b and 6d. We note that the ACU results change significantly under the zero-shot prompt compared to under the tuned 3-shot prompt.

We study the overarching trends shown by the averaged ACU scores in Table 11. We note how Llama shows better context usage scores compared to Pythia under both prompts, while Llama sees the greatest benefits from switching to the 3-shot prompt. Moreover, all models show improved average context usage under the 3-shot prompt compared to the 0-shot prompt, Pythia on CounterFact being the only exception.

We also look at changes in model prediction when the model is provided with evidence of a particular stance in Table 21 compared to when the model is provided with no evidence. In Table 22 we also list averaged  $\Delta P_M(t|C, E)$  stratified by evidence stance.

## L Dependence on context characteristics results

The results for Pythia corresponding to Figure 4 can be found in Figure 8a. Similarly, the results corresponding to insufficient evidence from DRUID for Llama and Pythia can be found in Figures 7c and 8c.

We also measure correlations between ACU and context characteristics under a 0-shot prompt. The results for Llama and Pythia can be found in Figures 7b and 8b. Similarly, we plot the zero-shot results for insufficient evidence from DRUID in Figures 7d and 8d. We note that while the ACU values changed significantly in Figure 6b under the 0-shot prompt compared to the 3-shot prompt, the dependencies on context characteristics are largely unchanged.

## M Annotation

### M.1 More details on the annotation

We screened the annotator pool to only include participants with at least an undergraduate degree, English fluency, no language-related disorders, and UK, US or Irish nationality. We were unable to obtain any additional details on e.g. the demographics of the annotation pool from Prolific as the group was too small to ensure anonymity if the information was shared.



Property	CounterFact	ConflictQA	DRUID+	DRUID
<b>Claim-evidence similarity</b>				
Jaccard similarity	<b>0.89 ± 0.12</b>	0.09 ± 0.04	0.09 ± 0.06	0.12 ± 0.08
Claim-evidence overlap	<b>0.93 ± 0.08</b>	0.76 ± 0.23	0.58 ± 0.25	0.66 ± 0.26
Repeats claim (%)	(50.00)	5.55	1.25	4.57
<b>Difficult to understand</b>				
Flesch reading ease score	61.65 ± 22.50	56.25 ± 12.00	53.17 ± 24.74	53.54 ± 16.60
Claim length	44.70 ± 11.90	43.64 ± 15.69	<b>84.08 ± 46.37</b>	<b>89.25 ± 46.15</b>
Evidence length	<b>44.63 ± 11.90</b>	570.46 ± 158.61	775.64 ± 407.40	745.39 ± 406.33
Llama: Perplexity	<b>172.94 ± 537.58</b>	7.55 ± 5.00	17.22 ± 124.59	16.08 ± 43.82
Pythia: Perplexity	<b>113.43 ± 1030.10</b>	9.29 ± 4.51	19.35 ± 122.17	18.13 ± 32.55
<b>Implicit</b>				
Claim entity overlap	0.75 ± 0.27	0.69 ± 0.36	0.46 ± 0.39	0.56 ± 0.40
<b>Refers external source</b>				
Detection by LLM (%)	(0.00)	27.37	-	<b>40.55</b>
<b>Unreliable</b>				
Unreliable source (%)	-	-	5.00	3.50
<b>Uncertain</b>				
Contains hedging (%)	<b>0.06</b>	15.34	36.61	36.54
Contains hedging discourse (%)	<b>0.03</b>	40.29	48.00	52.33
<b>Additional properties</b>				
Contains 'True'	0.00	1.99	2.57	4.06
Contains 'False'	0.00	0.10	4.27	9.02
Fact-check source (%)	-	-	14.41	<b>41.44</b>
Gold source (%)	-	-	4.13	17.21
Pub. after claim (%)	-	-	53.37	50.26
<b>Total instances</b>	20,000	16,046	48,517	5,490

Table 10: Statistics for the context characteristics in CounterFact (Yu et al., 2023), ConflictQA (Xie et al., 2024) and DRUID datasets. The characteristics and their detection are described in Sections 2 and 4.1, respectively. The values indicate the mean ± the standard deviation or the percentage of claim-evidence samples affected, denoted with (%). The LLM-detected properties for CounterFact indicated with a (parenthesis) were not LLM detected but automatically detected for each sample, leveraging its synthetic nature. Outliers are marked in **bold**.

Dataset	Model	Prompt	ACU
CounterFact	Llama	<b>3-shot</b>	<b>0.74</b>
		0-shot	0.57
	Pythia	3-shot	-0.05
ConflictQA	Llama	<b>3-shot</b>	<b>0.32</b>
		0-shot	0.71
	Pythia	<b>3-shot</b>	<b>0.55</b>
		0-shot	0.50
DRUID	Llama	<b>3-shot</b>	<b>0.84</b>
		0-shot	0.34
	Pythia	<b>3-shot</b>	<b>0.25</b>
		0-shot	0.15

Table 11: The averaged ACU scores across all evidence stances for each dataset, model and prompt. The better performing prompt and corresponding ACU value is marked in **bold**.

For the annotation of relevance, the annotators can choose between the labels ‘relevant’ or ‘not relevant’. Relevant evidence is double annotated for stance, where the annotators can choose between ⟨supports, insufficient-supports, insufficient-neutral, insufficient-contradictory, insufficient-refutes, refutes⟩. Insufficient evidence denotes evidence lacking sufficient information to assess the veracity of a given claim (Atanasova et al., 2022). There may be different levels of insufficiency, e.g. some evidence may not be sufficient, while it can lean towards being in support of a given claim (insufficient-supports). The stance labels found

in most fact-checking datasets are ⟨supports, not enough info, refutes⟩, for which ‘not enough info’ is essentially the same as ‘insufficient’ (Thorne et al., 2018). Schlichtkrull et al. (2023) also include the label ‘conflicting evidence/cherry-picking’ for their annotation task, which has some similarities to our label ‘insufficient-contradictory’. We expand on the labels by adding more nuances to insufficient evidence, and find this to improve annotator agreement.

## M.2 Annotation guidelines and annotation interface

The annotation guidelines and two examples annotation pages from the annotation interface can be found in Figures 9 to 18.

Model	#shots	Prompt
Llama	3	<p>Are the following claims True or False? Answer None if you are not sure or cannot answer.</p> <p>Claimant: Joe Biden  Claim: ""One quarter" of today's \$31.4 trillion federal debt "was accumulated in the four years of my predecessor," Donald Trump."  Answer: True</p> <p>Claimant: Viral post  Claim: "the new coronavirus has HIV proteins that indicate it was genetically modified in a laboratory."  Answer: False</p> <p>Claimant: Sara Daniels  Claim: "Blackpink released the single 'You me too' in 2026."  Answer: None</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Answer:</p>
Pythia	3	<p>Are the following claims True or False? Answer None if you are not sure or cannot answer.</p> <p>Claimant: Joe Biden  Claim: ""One quarter" of today's \$31.4 trillion federal debt "was accumulated in the four years of my predecessor," Donald Trump."  Answer: True</p> <p>Claimant: Viral post  Claim: "5G causes cancer."  Answer: False</p> <p>Claimant: Sara Daniels  Claim: "Blackpink released the single 'You me too' in 2026."  Answer: None</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Answer:</p>
Llama & Pythia	0	<p>Is the following claim True or False? Answer None if you are not sure or cannot answer.</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Answer:</p>

Table 12: Prompts used to evaluate Pythia and Llama in a claim-only setting. Values in <brackets> are replaced by the actual entry for the evaluated sample. For CounterFact and ConflictQA we omit the 'Claimant' lines as there are no claimant entries for these datasets.

Model	#shots	Prompt
Llama	3	<p>Here are some claims and corresponding evidence. Does the evidence Support or Refute the claim? Answer None if there is not enough information in the evidence to decide.</p> <p>Claimant: Joe Biden  Claim: ""One quarter" of today's \$31.4 trillion federal debt "was accumulated in the four years of my predecessor," Donald Trump."  Evidence: "Biden's number is accurate; about one-fourth of the total debt incurred to date came on Trump's watch. However, assigning debt to a particular president is tricky, because so much of the spending was approved by decades-old, bipartisan legislation that set the parameters for Social Security and Medicare. A different calculation shows more debt stemming from former President Barack Obama, with whom Biden served as vice president."  Answer: Support</p> <p>Claimant: Viral post  Claim: "the new coronavirus has HIV proteins that indicate it was genetically modified in a laboratory." Evidence: "Microbiologists say the spike proteins found in the new coronavirus are different from the ones found in HIV. [...] There is no evidence to suggest the coronavirus was genetically modified."  Answer: Refute</p> <p>Claimant: Sara Daniels  Claim: "Blackpink released the single 'You me too' in 2026."  Evidence: "Blackpink released their album 'Born Pink' in 2022."  Answer: None</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Evidence: "&lt;evidence&gt;"  Answer:</p>

Table 13: Prompts used to evaluate Pythia and Llama in a setting with provided claim and evidence. Values in <brackets> are replaced by the actual entry for the evaluated sample. For CounterFact and ConflictQA we omit the 'Claimant' lines as there are no claimant entries for these datasets.

Model	#shots	Prompt
Pythia	3	<p>Are the claims True or False based on the accompanying evidence? If you are not sure or cannot answer, say None.</p> <p>Claimant: Joe Biden  Claim: ""One quarter" of today's \$31.4 trillion federal debt "was accumulated in the four years of my predecessor," Donald Trump."  Evidence: "Biden's number is accurate; about one-fourth of the total debt incurred to date came on Trump's watch. However, assigning debt to a particular president is tricky, because so much of the spending was approved by decades-old, bipartisan legislation that set the parameters for Social Security and Medicare. A different calculation shows more debt stemming from former President Barack Obama, with whom Biden served as vice president."  Answer: True</p> <p>Claimant: Viral post  Claim: "the new coronavirus has HIV proteins that indicate it was genetically modified in a laboratory." Evidence: "Microbiologists say the spike proteins found in the new coronavirus are different from the ones found in HIV. [...] There is no evidence to suggest the coronavirus was genetically modified."  Answer: False</p> <p>Claimant: Sara Daniels  Claim: "Blackpink released the single 'You me too' in 2026."  Evidence: "Blackpink released their album 'Born Pink' in 2022."  Answer: None</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Evidence: "&lt;evidence&gt;"  Answer:</p>
Llama, Pythia	0	<p>Based on the provided evidence, is the claim True or False? If you are not sure or cannot answer, say None.</p> <p>Claimant: &lt;claimant&gt;  Claim: "&lt;claim&gt;"  Evidence: "&lt;evidence&gt;"  Answer:</p>

Table 14: Prompts used to evaluate Pythia and Llama in a setting with provided claim and evidence. Values in <brackets> are replaced by the actual entry for the evaluated sample. For CounterFact and ConflictQA we omit the 'Claimant' lines as there are no claimant entries for these datasets.

<b>Dataset</b>	CounterFact		
<b>Claim</b>	Danish Outdoor Council is headquartered in Perth.		
<b>Evidence</b>	Danish Outdoor Council is headquartered in Copenhagen.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	1.51		
$P_{Llama}(False C)$	0.69	$P_{Llama}(False C, E)$	0.84
$P_{Llama}(None C)$	0.17	$P_{Llama}(None C, E)$	0.15
$P_{Llama}(True C)$	0.14	$P_{Llama}(True C, E)$	0.01
<b>ACU<sub>Pythia</sub></b>	0.13		
$P_{Pythia}(False C)$	0.36	$P_{Pythia}(False C, E)$	0.34
$P_{Pythia}(None C)$	0.03	$P_{Pythia}(None C, E)$	0.26
$P_{Pythia}(True C)$	0.58	$P_{Pythia}(True C, E)$	0.34
<b>Dataset</b>	CounterFact		
<b>Claim</b>	Yahoo! Screen is owned by Sony.		
<b>Evidence</b>	Yahoo! Screen is owned by Sony.		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	1.77		
$P_{Llama}(False C)$	0.65	$P_{Llama}(False C, E)$	0.05
$P_{Llama}(None C)$	0.10	$P_{Llama}(None C, E)$	0.09
$P_{Llama}(True C)$	0.25	$P_{Llama}(True C, E)$	0.84
<b>ACU<sub>Pythia</sub></b>	0.16		
$P_{Pythia}(False C)$	0.51	$P_{Pythia}(False C, E)$	0.12
$P_{Pythia}(None C)$	0.04	$P_{Pythia}(None C, E)$	0.47
$P_{Pythia}(True C)$	0.42	$P_{Pythia}(True C, E)$	0.36
<b>Dataset</b>	CounterFact		
<b>Claim</b>	The Voice debuted on CNN.		
<b>Evidence</b>	The Voice debuted on CNN.		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	1.44		
$P_{Llama}(False C)$	0.58	$P_{Llama}(False C, E)$	0.09
$P_{Llama}(None C)$	0.13	$P_{Llama}(None C, E)$	0.17
$P_{Llama}(True C)$	0.28	$P_{Llama}(True C, E)$	0.73
<b>ACU<sub>Pythia</sub></b>	1.01		
$P_{Pythia}(False C)$	0.43	$P_{Pythia}(False C, E)$	0.12
$P_{Pythia}(None C)$	0.09	$P_{Pythia}(None C, E)$	0.19
$P_{Pythia}(True C)$	0.45	$P_{Pythia}(True C, E)$	0.66

Table 15: Cherry-picked ACU samples for Llama and/or Pythia on CounterFact.

<b>Dataset</b>	CounterFact		
<b>Claim</b>	Satchel Paige is a professional basketball.		
<b>Evidence</b>	Satchel Paige is a professional baseball.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	0.04		
$P_{Llama}(False C)$	0.75	$P_{Llama}(False C, E)$	0.55
$P_{Llama}(None C)$	0.10	$P_{Llama}(None C, E)$	0.39
$P_{Llama}(True C)$	0.14	$P_{Llama}(True C, E)$	0.05
<b>ACU<sub>Pythia</sub></b>	-0.14		
$P_{Pythia}(False C)$	0.46	$P_{Pythia}(False C, E)$	0.35
$P_{Pythia}(None C)$	0.04	$P_{Pythia}(None C, E)$	0.27
$P_{Pythia}(True C)$	0.47	$P_{Pythia}(True C, E)$	0.31
<b>Dataset</b>	CounterFact		
<b>Claim</b>	Honda SFX, produced by Airbus.		
<b>Evidence</b>	Honda SFX, produced by Honda.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	-0.73		
$P_{Llama}(False C)$	0.64	$P_{Llama}(False C, E)$	0.37
$P_{Llama}(None C)$	0.16	$P_{Llama}(None C, E)$	0.42
$P_{Llama}(True C)$	0.20	$P_{Llama}(True C, E)$	0.20
<b>ACU<sub>Pythia</sub></b>	-0.73		
$P_{Pythia}(False C)$	0.44	$P_{Pythia}(False C, E)$	0.18
$P_{Pythia}(None C)$	0.14	$P_{Pythia}(None C, E)$	0.52
$P_{Pythia}(True C)$	0.36	$P_{Pythia}(True C, E)$	0.25
<b>Dataset</b>	CounterFact		
<b>Claim</b>	iPad, developed by Douglas.		
<b>Evidence</b>	iPad, developed by Douglas.		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	0.44		
$P_{Llama}(False C)$	0.49	$P_{Llama}(False C, E)$	0.14
$P_{Llama}(None C)$	0.17	$P_{Llama}(None C, E)$	0.46
$P_{Llama}(True C)$	0.33	$P_{Llama}(True C, E)$	0.38
<b>ACU<sub>Pythia</sub></b>	-0.10		
$P_{Pythia}(False C)$	0.40	$P_{Pythia}(False C, E)$	0.10
$P_{Pythia}(None C)$	0.15	$P_{Pythia}(None C, E)$	0.60
$P_{Pythia}(True C)$	0.36	$P_{Pythia}(True C, E)$	0.24

Table 16: Lemon-picked ACU samples for Llama and/or Pythia on CounterFact.



<b>Dataset</b>	ConflictQA		
<b>Claim</b>	The screenwriter for Highway was Imtiaz Ali.		
<b>Evidence</b>	Highway is a 2014 Indian Hindi-language road drama film written and directed by Imtiaz Ali and produced by Sajid Nadiadwala. The film stars Alia Bhatt and Randeep Hooda. Screened in the Panorama section of the 2014 Berlin International Film Festival, the film released worldwide on 21 February 2014. The film is based on the episode of the same name from the Zee TV anthology series Rishtey, starring Aditya Srivastava and Kartika Rane, which was also written and directed by Imtiaz Ali. It tells the story of a girl (Alia Bhatt) who, for reasons later revealed, discovers freedom after being kidnapped.		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	1.94		
$P_{Llama}(False C)$	0.37	$P_{Llama}(False C, E)$	0.03
$P_{Llama}(None C)$	0.07	$P_{Llama}(None C, E)$	0.05
$P_{Llama}(True C)$	0.55	$P_{Llama}(True C, E)$	0.90
<b>ACU<sub>Pythia</sub></b>	0.89		
$P_{Pythia}(False C)$	0.39	$P_{Pythia}(False C, E)$	0.22
$P_{Pythia}(None C)$	0.09	$P_{Pythia}(None C, E)$	0.07
$P_{Pythia}(True C)$	0.47	$P_{Pythia}(True C, E)$	0.59
<b>Dataset</b>	ConflictQA		
<b>Claim</b>	The composer of The Nose was Dmitri Shostakovich.		
<b>Evidence</b>	Michael Figgis is indeed the composer of The Nose. Figgis is a highly respected composer, having won numerous awards for his film scores, and his work on The Nose has been praised by both critics and audiences. In an interview with Film Score Monthly, Figgis stated that he was inspired by the surrealism of the story and the absurdist humor in Gogol's writing, and that he wanted to create a score that captured the feeling of disorientation and confusion that is so prevalent in the story. He also discussed the challenges of translating the story's unique tone and atmosphere into music, but ultimately felt that he was able to find the right balance. Overall, Figgis's work on The Nose is a testament to his skill as a composer and his ability to bring unique and complex stories to life through music.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	1.01		
$P_{Llama}(False C)$	0.46	$P_{Llama}(False C, E)$	0.65
$P_{Llama}(None C)$	0.05	$P_{Llama}(None C, E)$	0.31
$P_{Llama}(True C)$	0.48	$P_{Llama}(True C, E)$	0.03
<b>ACU<sub>Pythia</sub></b>	-0.66		
$P_{Pythia}(False C)$	0.37	$P_{Pythia}(False C, E)$	0.11
$P_{Pythia}(None C)$	0.05	$P_{Pythia}(None C, E)$	0.01
$P_{Pythia}(True C)$	0.54	$P_{Pythia}(True C, E)$	0.84

Table 17: Cherry-picked ACU samples for Llama and/or Pythia on ConflictQA.

<b>Dataset</b>	ConflictQA		
<b>Claim</b>	The Canada women's national field hockey team plays field hockey.		
<b>Evidence</b>	Contrary to popular belief, the Canada women's national field hockey team plays football as well. In fact, many field hockey players also have a background in football, as the two sports share similar skills such as agility, speed, and endurance. According to a recent interview with team captain Sarah Jullien, she stated that "I started playing football when I was young and it has definitely helped me improve my performance on the field hockey pitch." Additionally, the team's official website lists football as one of the recommended cross-training sports for players looking to improve their game.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	0.48		
$P_{Llama}(False C)$	0.16	$P_{Llama}(False C, E)$	0.34
$P_{Llama}(None C)$	0.03	$P_{Llama}(None C, E)$	0.20
$P_{Llama}(True C)$	0.80	$P_{Llama}(True C, E)$	0.46
<b>ACU<sub>Pythia</sub></b>	-0.29		
$P_{Pythia}(False C)$	0.40	$P_{Pythia}(False C, E)$	0.29
$P_{Pythia}(None C)$	0.05	$P_{Pythia}(None C, E)$	0.04
$P_{Pythia}(True C)$	0.51	$P_{Pythia}(True C, E)$	0.64
<b>Dataset</b>	ConflictQA		
<b>Claim</b>	Domašov is located in the Czech Republic.		
<b>Evidence</b>	not live in these communities, but they are members of the Miles Jesu family. It was reported in 2004 that there were 27 Miles Jesu houses in 14 countries. The latest (January 2012) information indicates that there are domus communities in 9 countries and vinculum members in an additional 3 countries. Domus communities are found in the following countries (with date of first foundation): United States (1964), India (1984), Spain (1985), Nigeria (1987), Italy (1988) Czech Republic (1990), Ukraine (1990), Poland (1991), and Slovakia (2004). The three additional countries are Puerto Rico, England and Austria. The members in the Ukraine		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	-0.53		
$P_{Llama}(False C)$	0.15	$P_{Llama}(False C, E)$	0.14
$P_{Llama}(None C)$	0.04	$P_{Llama}(None C, E)$	0.31
$P_{Llama}(True C)$	0.81	$P_{Llama}(True C, E)$	0.54
<b>ACU<sub>Pythia</sub></b>	0.03		
$P_{Pythia}(False C)$	0.39	$P_{Pythia}(False C, E)$	0.23
$P_{Pythia}(None C)$	0.05	$P_{Pythia}(None C, E)$	0.16
$P_{Pythia}(True C)$	0.54	$P_{Pythia}(True C, E)$	0.40

Table 18: Lemon-picked ACU samples for Llama and/or Pythia on ConflictQA.

<b>Dataset</b>	DRUID		
<b>Claim</b>	Vandana Tiwari, Sister of Bageshwar Dham Dhirendra Shastri, is marrying a Muslim man		
<b>Evidence</b>	A photo of a couple in traditional attire is viral on social media, claiming that the woman seen in it is Bageshwar Dham Dhirendra Shastri's sister 'Vandana Tiwari'. A post that shares this photo claims that she is marrying a Muslim man without the knowledge of her brother. Let's verify the truth behind these claims through this fact-checking article. [...] According to BBC, Bageshwar Dham's Dhirendra Shastri has a sister named Rita Garg, and she is already married. Regarding this issue, Bageshwar Dham's PRO, Kamal Awasthi, told Aaj Tak that she married a Hindu man called Kamlesh Chauraha in 2015. All of this makes it evident that the viral photo is being misquoted as a picture of Dhirendra Shastri's sister and her Muslim husband while it actually features Actress Gehana Vasisth and her husband.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	2.04		
$P_{Llama}(False C)$	0.42	$P_{Llama}(False C, E)$	0.78
$P_{Llama}(None C)$	0.36	$P_{Llama}(None C, E)$	0.19
$P_{Llama}(True C)$	0.20	$P_{Llama}(True C, E)$	0.01
<b>ACU<sub>Pythia</sub></b>	-0.12		
$P_{Pythia}(False C)$	0.55	$P_{Pythia}(False C, E)$	0.52
$P_{Pythia}(None C)$	0.04	$P_{Pythia}(None C, E)$	0.08
$P_{Pythia}(True C)$	0.33	$P_{Pythia}(True C, E)$	0.34
<b>Dataset</b>	DRUID		
<b>Claim</b>	Sapodilla Cay is a territory of Guatemala		
<b>Evidence</b>	Guatemala recalls in its application for permission to intervene that on November 16, 2022, Belize initiated proceedings against the Republic of Honduras over "sovereignty over the Sapodilla Cays or Cayes, a cluster of islands in the Gulf of Honduras, which Guatemala also claims." Belize asks the Court to "adjudicate and declare that, as between Belize and Honduras, Belize is sovereign over the Sapodilla Cayes." [...] (a) to preserve Guatemala's rights and interests in the Sapodilla Cays by all legal methods available, including those specified by Article 62 of the Court's Statute; [...] Belize stated in its Application for Initiation of Proceedings that the Sapodilla Cayes have been part of the territory of Belize since the early nineteenth century, first as part of the settlement of Belize and later as part of the colony of British Honduras, and since 1981 as part of the independent State of Belize.		
<b>Evidence stance</b>	insufficient-neutral		
<b>ACU<sub>Llama</sub></b>	0.75		
$P_{Llama}(False C)$	0.63	$P_{Llama}(False C, E)$	0.19
$P_{Llama}(None C)$	0.17	$P_{Llama}(None C, E)$	0.41
$P_{Llama}(True C)$	0.20	$P_{Llama}(True C, E)$	0.39
<b>ACU<sub>Pythia</sub></b>	1.25		
$P_{Pythia}(False C)$	0.37	$P_{Pythia}(False C, E)$	0.05
$P_{Pythia}(None C)$	0.03	$P_{Pythia}(None C, E)$	0.12
$P_{Pythia}(True C)$	0.57	$P_{Pythia}(True C, E)$	0.40

Table 19: Cherry-picked ACU samples for Llama and/or Pythia on DRUID.

<b>Dataset</b>	DRUID		
<b>Claim</b>	Blocks of color printed on toothpaste tubes indicate whether the toothpaste is made of safe ingredients		
<b>Evidence</b>	The truth is: the toothpaste color-coding system simply doesn't exist. Oral care companies don't mark their toothpastes with colored squares to try to trick consumers and hide ingredients from them. We're sure you're wondering, so why are there color blocks on toothpaste tubes then? We're happy to report that they do, in fact, have a purpose! They actually help in the manufacturing of the toothpaste tubes by telling light sensors where the end of the tube is so that it can be cut and sealed properly. We know, it's not as exciting as a secret code, but we think the truth is pretty cool too. [...] If you want to know what kind of ingredients your toothpaste has, don't look for a colored block at the end of the tube. Instead, take a look at the packaging for a comprehensive list of ingredients. You can talk with your dentist to learn more about how each ingredient works to keep your mouth healthy and what kind of toothpaste would be best to meet your needs.		
<b>Evidence stance</b>	refutes		
<b>ACU<sub>Llama</sub></b>	1.16		
$P_{Llama}(False C)$	0.64	$P_{Llama}(False C, E)$	0.76
$P_{Llama}(None C)$	0.21	$P_{Llama}(None C, E)$	0.21
$P_{Llama}(True C)$	0.14	$P_{Llama}(True C, E)$	0.02
<b>ACU<sub>Pythia</sub></b>	-0.75		
$P_{Pythia}(False C)$	0.49	$P_{Pythia}(False C, E)$	0.29
$P_{Pythia}(None C)$	0.03	$P_{Pythia}(None C, E)$	0.07
$P_{Pythia}(True C)$	0.42	$P_{Pythia}(True C, E)$	0.60
<b>Dataset</b>	DRUID		
<b>Claim</b>	CO2 concentrations are increasing in Earth's atmosphere faster than they have in the last 50,000 years.		
<b>Evidence</b>	Atmospheric CO2 concentrations rising faster today than the last 50,000 years, as accurately claimed in recent social media posts Atmospheric carbon dioxide (CO2) concentrations and their current rate of increase is unprecedented in the last 50,000 years, based on ice core data. The highest increase in CO2 in that period occurred over the span of 50 years, but the same increase occurred in only the last five years – which is 10 times faster. As human emissions of CO2 increase, global temperatures rise in response through the greenhouse effect. [...] In May 2024, a number of articles and Facebook posts claimed that carbon dioxide (CO2) concentrations are increasing in Earth's atmosphere faster than they have in the last 50,000 years. So what sparked this claim?		
<b>Evidence stance</b>	supports		
<b>ACU<sub>Llama</sub></b>	-1.09		
$P_{Llama}(False C)$	0.15	$P_{Llama}(False C, E)$	0.35
$P_{Llama}(None C)$	0.16	$P_{Llama}(None C, E)$	0.37
$P_{Llama}(True C)$	0.67	$P_{Llama}(True C, E)$	0.26
<b>ACU<sub>Pythia</sub></b>	-0.86		
$P_{Pythia}(False C)$	0.40	$P_{Pythia}(False C, E)$	0.49
$P_{Pythia}(None C)$	0.02	$P_{Pythia}(None C, E)$	0.12
$P_{Pythia}(True C)$	0.56	$P_{Pythia}(True C, E)$	0.23

Table 20: Lemon-picked ACU samples for Llama and/or Pythia on DRUID.

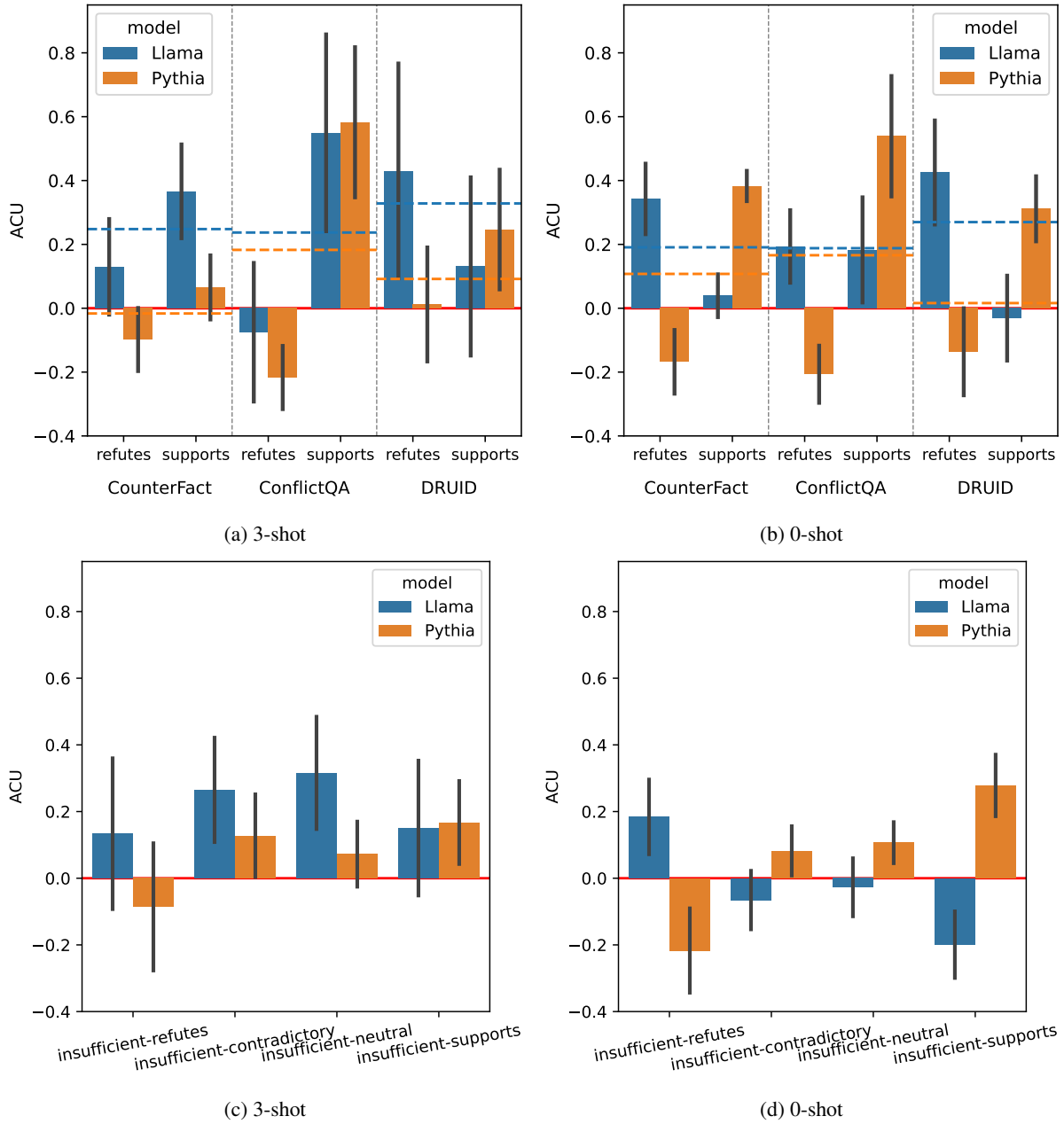


Figure 6: Accumulated context usage (Equation (2)) each model on sufficient evidence from all datasets (Figures 6a and 6b) and insufficient evidence from DRUID (Figures 6c and 6d) under different prompts. Figure 3 is included again for comparison (Figure 6a). The error bars indicate the standard deviation. The maximum and minimum context usage value possible is 3 and -3, respectively. Values under the red line indicate 'context-repulsion'.

Dataset	Model	Prediction stance	Evidence stance			$\sum \Delta N_D$	Memory conflict (%)
			False	None	True		
CounterFact	Llama	False $\uparrow\downarrow$	7,166 (-2575)	2,826 (+2,826)	8 (-251)	-5,150	2.59
		True $\downarrow\uparrow$	0 (-9,741)	2,557 (+2,557)	7,443 (+7,184)	14,358	97.41
	Pythia	False $\uparrow\downarrow$	1,608 (-2,364)	4,026 (+3,967)	4,366 (-1,591)	-4,740	59.57
		True $\downarrow\uparrow$	0 (-3,972)	4,095 (+4,036)	5,905 (-52)	-116	39.72
ConflictQA	Llama	False $\uparrow\downarrow$	1,048 (-1,265)	4,149 (+4,148)	2,826 (-2,883)	-2,530	71.16
		True $\downarrow\uparrow$	5 (-2,308)	350 (+349)	7,668 (+1,959)	3,918	28.83
	Pythia	False $\uparrow\downarrow$	170 (-1,890)	78 (+42)	7,766 (+1,839)	-3,771	73.88
		True $\downarrow\uparrow$	17 (-2,043)	29 (-7)	7,972 (+2,045)	4,095	25.68
DRUID	Llama	False $\uparrow\downarrow$	1,528 (-86)	202 (+181)	30 (-95)	-172	7.10
		None $\downarrow\uparrow$	600 (-1,126)	1,226 (+1,199)	219 (-73)	2,398	0.00
		True $\downarrow\uparrow$	124 (-404)	285 (+274)	500 (+130)	260	58.09
	Pythia	False $\uparrow\downarrow$	1,212 (-101)	2 (0)	543 (+98)	-199	25.28
		None $\downarrow\uparrow$	450 (-509)	20 (+20)	1,569 (+483)	46	0.00
		True $\downarrow\uparrow$	83 (-242)	3 (+2)	822 (+239)	479	35.75

Table 21: Model predictions for the task of claim verification based on provided evidence, stratified by evidence stance. Values in (parenthesis) indicate the change in model predictions compared to when the model is prompted without context. The arrows indicate the desirable direction for maximum context usage for each of the possible output labels (False, None, True). For example, given that a model has predicted ‘False’, we ideally want it to do this only on evidence with the stance ‘False’. Numbers in green indicate that the model generally is following the context and numbers in red indicate the opposite, based on whether the total model predictions change to align more with the desirable direction when evidence is introduced.  $\sum \Delta N_D$  indicates the accumulated number of desirable switches minus the number of undesirable switches in model prediction when provided with evidence of a certain stance. ‘Memory conflict’ indicates the share of samples for which the stance of the provided evidence conflicts with the parametric model prediction (no context or evidence provided). ‘None’ evidence stances or parametric predictions are not considered to correspond to memory conflicts.

Dataset	Model	Evidence stance	$\Delta P_M(\text{False} C, E)$	$\Delta P_M(\text{None} C, E)$	$\Delta P_M(\text{True} C, E)$	ACU
CounterFact	Llama	refutes $\uparrow\downarrow$	$-0.13 \pm 0.24$	$0.31 \pm 0.13$	$-0.84 \pm 0.18$	0.40
		supports $\downarrow\uparrow$	$-0.83 \pm 0.07$	$0.18 \pm 0.19$	$0.45 \pm 0.20$	1.10
	Pythia	refutes $\uparrow\downarrow$	$-0.30 \pm 0.19$	$0.27 \pm 0.11$	$-0.27 \pm 0.17$	-0.30
		supports $\downarrow\uparrow$	$-0.62 \pm 0.11$	$0.33 \pm 0.12$	$-0.10 \pm 0.20$	0.19
ConflictQA	Llama	refutes $\uparrow\downarrow$	$-0.28 \pm 0.34$	$0.33 \pm 0.17$	$-0.39 \pm 0.44$	-0.22
		supports $\downarrow\uparrow$	$-0.81 \pm 0.22$	$-0.26 \pm 0.39$	$0.58 \pm 0.38$	1.65
	Pythia	refutes $\uparrow\downarrow$	$-0.51 \pm 0.20$	$-0.11 \pm 0.28$	$0.26 \pm 0.29$	-0.66
		supports $\downarrow\uparrow$	$-0.70 \pm 0.17$	$-0.47 \pm 0.33$	$0.58 \pm 0.26$	1.75
DRUID	Llama	refutes $\uparrow\downarrow$	$0.35 \pm 0.41$	$-0.22 \pm 0.37$	$-0.71 \pm 0.35$	1.28
		insufficient-refutes $\uparrow\uparrow$	$-0.07 \pm 0.46$	$0.07 \pm 0.34$	$-0.40 \pm 0.44$	0.40
		insufficient-contr. $\downarrow\uparrow$	$-0.41 \pm 0.33$	$0.25 \pm 0.23$	$-0.13 \pm 0.37$	0.79
		insufficient-neutral $\downarrow\downarrow$	$-0.35 \pm 0.29$	$0.31 \pm 0.20$	$-0.30 \pm 0.37$	0.96
		insufficient-supports $\downarrow\uparrow$	$-0.37 \pm 0.29$	$0.16 \pm 0.27$	$-0.08 \pm 0.38$	0.45
		supports $\downarrow\uparrow$	$-0.40 \pm 0.30$	$0.05 \pm 0.32$	$0.04 \pm 0.39$	0.39
	Pythia	refutes $\uparrow\downarrow$	$-0.07 \pm 0.27$	$0.07 \pm 0.10$	$-0.17 \pm 0.33$	0.03
		insufficient-refutes $\uparrow\uparrow$	$-0.25 \pm 0.26$	$0.03 \pm 0.13$	$0.04 \pm 0.29$	-0.26
		insufficient-contr. $\downarrow\uparrow$	$-0.40 \pm 0.21$	$0.04 \pm 0.11$	$0.06 \pm 0.26$	0.38
		insufficient-neutral $\downarrow\downarrow$	$-0.33 \pm 0.21$	$0.01 \pm 0.13$	$0.13 \pm 0.25$	0.21
		insufficient-supports $\downarrow\uparrow$	$-0.35 \pm 0.22$	$-0.03 \pm 0.17$	$0.19 \pm 0.24$	0.51
		supports $\downarrow\uparrow$	$-0.42 \pm 0.22$	$-0.06 \pm 0.20$	$0.26 \pm 0.24$	0.74

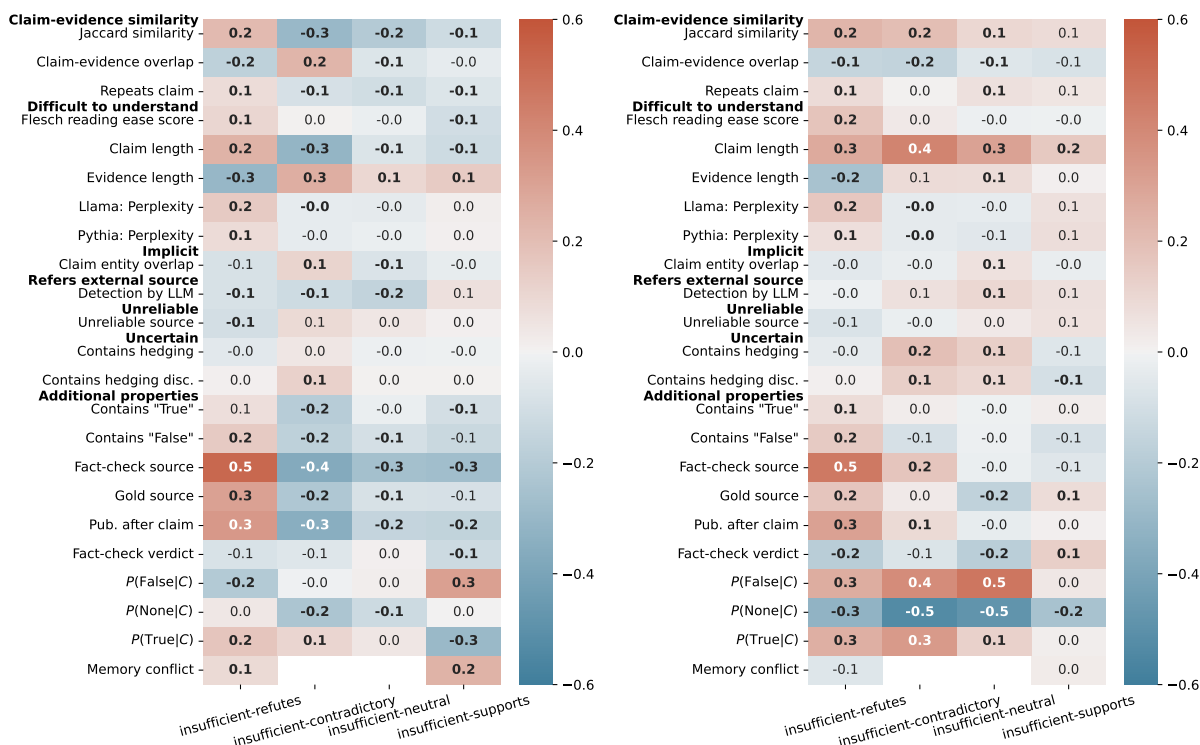
Table 22: Averages and standard deviations for differences in prediction probabilities (scaled) when evidence is introduced for all datasets. The arrows indicate the desirable direction for maximum context usage. Numbers in green indicate that the model generally is following the context and numbers in red indicate the opposite, based on the total change in prediction probability as evidence is introduced. ACU is defined in Equation (2).





(a) 3-shot

(b) 0-shot



(c) 3-shot

(d) 0-shot

Figure 7: Spearman correlations between ACU and different sample features for Llama under a tuned 3-shot prompt and a zero-shot prompt. We also show the results on insufficient evidence from DRUID. Significant correlation values (p-value 0.05) are marked in bold.

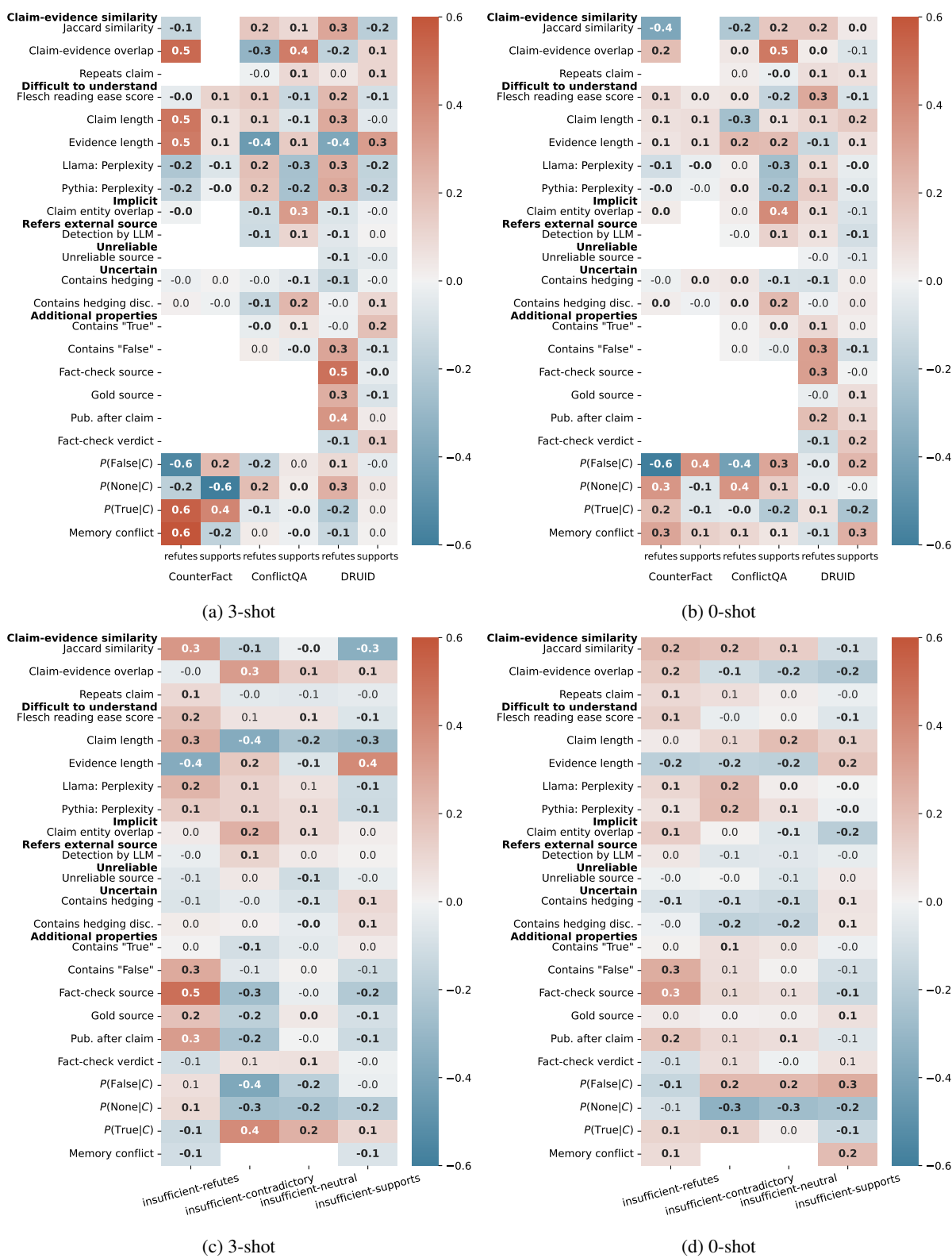


Figure 8: Spearman correlations between ACU and different sample features for Pythia under a tuned 3-shot prompt and a zero-shot prompt. We also show the results on insufficient evidence from DRUID. Significant correlation values (p-value 0.05) are marked in bold.

intro

## Introduction

For this annotation task, you are to consider **claims** and corresponding **evidence pieces**. Your task is to annotate the **relevance** and **stance** of evidence pieces:

- A **claim** is some statement about the world, see the example below.
- Additional information is typically necessary to verify the truthfulness of a claim — this is referred to as **evidence** or **evidence piece**. An evidence piece consists of one or several sentences extracted from an external source for the particular claim. Each evidence shown comes from a webpage. Some evidence may contain a '['...' or new paragraphs to indicate that the subsequent text comes from the same webpage but a bit further down, do not concern yourself with what might have been written in between. All text following "Evidence:" is a part of the evidence.
- Relevant evidence has a **stance** with respect to the given claim. The evidence in the example below has the stance **refutes**, i.e. according to the evidence the claim is false.

## Example

<b>Claim</b>	Hate speech is not protected by the first amendment.
<b>Claimant</b>	John Doe
<b>Claim date</b>	2024-05-04
<b>Evidence #1</b>	There are some exceptions to the free speech clause in the First Amendment, but "hate speech" is not one of them. The Supreme Court has repeatedly held up the right of an individual or group to engage in speech that much of the public likely finds offensive, like displaying swastikas, burning crosses, or protesting a soldier's funeral.
<b>Evidence date</b>	2024-07-01
<b>Relevant</b>	<b>true</b>
<b>Stance</b>	<b>refutes</b>

Example of a claim with corresponding evidence and annotations. Fields in pink are to be annotated and fields in gray are provided for the annotation task. 'Claim date' denotes the publish date of the claim. 'Evidence date' denotes the publish date of the evidence.

## General guidelines

- Only consider the provided claim (including claim date and claimant) and evidence (including evidence date) for your annotation. **Do not involve any prior or external knowledge related to the claim** apart from what is in the given claim and evidence. Disregard any personal opinions, whether you disagree or agree with the claim or evidence. Also, disregard evidence that you might have observed previously for the same or different claims. The only exception to this rule is when you are explicitly instructed to consider external knowledge for specific annotations.
- For each task you have the option of flagging samples with **quality issues**. If there is any issue with a sample that prevents you from annotating it as instructed, leave a comment shortly describing the issue in the corresponding quality issue textbox, mark the evidence stance as 'not\_applicable' and move on. Do not annotate the sample. For example, a quality issue could arise if there is some issue with the provided claim that makes it impossible to annotate evidence properties (such as an incomplete, hard to understand or ambiguous claim), or if the evidence is empty. Ideally, you should not need to use this textbox frequently.
- This annotation task is focused on analysing information expressed as **text only**. If you encounter a claim or evidence that seems to discuss a photo or video, treat it as a quality issue.
- The claims and evidence presented in the annotation task are not necessarily accurate. Moreover, the information presented in the task does not represent the opinion of the task administrators.

## General information

Figure 9: Page 1 of 10 depicting the annotation interface.

- This study will store your Prolific ID and annotation results. None of your personal information will be stored. The annotations may later on be released in a public dataset to aid other researchers in their work. This dataset would not reveal your Prolific ID or any other personal information that could be connected to you.
- After having completed the assigned annotations, you will be redirected to the Prolific webpage to finalise the task. You will be compensated for your work after the quality of your annotations has been verified.
- You can drop out of the study any time you like by just leaving this webpage. However, you will only receive your reward after having completed all requested annotations.
- If you drop out of the study, please return this submission on Prolific by selecting the 'stop without completing' button.
- Some of the samples shown to you may contain information that is politically loaded or misinforming.
- At the end of the study, you will have the option of providing feedback.
- Your participation will help us develop more accurate and reliable systems for automated claim verification, an important subtask within automated fact checking.

Keyboard Input:

→  
Move forward

Move forward

---

Copyright © 2022 [Blablalab](#)

[Fork on GitHub](#) | [Cite Us](#)

Figure 10: Page 2 of 10 depicting the annotation interface.



## Instructions

Your task is to annotate the **relevance** and **stance** of evidence pieces. Examples of different evidence pieces and corresponding annotations can be found at the bottom of this page.

### Step 1: Is the evidence relevant?

Does the evidence contain any information that

1. directly supports or refutes the claim,
2. is topically related to the topic or entities of the claim or claimant (same people, places, organisations, etc.), or
3. can be seen as implicitly referring to the claim?

If yes, annotate relevant as **true**.

Also, assign a value of **true** for this property if the evidence discusses the same subject as the claim, regardless of whether it fully supports or refutes the claim. Even if the evidence is insufficient to determine the claim's veracity, it should still be considered relevant if it addresses the topic of the claim or discusses the same entities (e.g., Evidence #3 and Additional example #1). Also, assign a value of **true** if the evidence could be seen as implicitly referring to the claim (e.g., Evidence #7). If the evidence does not change your confidence in the verdict but is on-topic, it should still be marked as relevant.

Assign a value of **false** if the evidence is entirely unrelated to the claim or its subject matter (e.g., Evidence #2).

### Step 2: What is the stance of the evidence?

Each provided evidence should correspond to one of the stances listed below with respect to the given claim. Only one stance can be given per claim-evidence pair. Evidence marked as relevant: **false** cannot correspond to a stance. You can always assume that the provided claim has been made by the claimant. Do not annotate the stance with respect to whether the claimant made the claim or not (e.g., Evidence #8).

- **supports**: The evidence contains sufficient information that supports the given claim. I.e., the claim can be considered true based on the information in the corresponding evidence. Repetitions or quotes of the claim in the evidence do not count as support if it is clear that they are there simply to show that the claim is being addressed (e.g., Evidence #9). For claims consisting of several parts, all parts need to be supported by the evidence to mark it as supporting.
- **refutes**: The evidence contains sufficient information that refutes the given claim. I.e., the claim can be considered false based on the information in the corresponding evidence. For claims consisting of several parts, it is enough if only one part of the claim is refuted by the evidence to mark it as refuting (e.g., Additional example #2).
- **insufficient**: The evidence contains relevant information but is not sufficient for verifying the claim. For samples with this stance, we ask you to use the following subcategories:
  - **insufficient-supports**: The evidence is insufficient, but leaning towards supporting the claim.
  - **insufficient-refutes**: The evidence is insufficient, but leaning towards refuting the claim.
  - **insufficient-contradictory**: The evidence contains information that is contradictory, such that no final verdict can be reached. For example, some part of the evidence could support the claim while another part refutes it.
  - **insufficient-neutral**: The evidence is insufficient and not in particular support or opposition to the claim.
- **not\_applicable**: If you annotated relevant as **false** in the previous task, mark the stance as **not\_applicable**.

Figure 11: Page 3 of 10 depicting the annotation interface.

## Examples of annotations

---

<b>Claim</b>	Hate speech is not protected by the first amendment.
<b>Claimant</b>	John Doe
<b>Claim date</b>	2024-05-04

---

<b>Evidence #1</b>	There are some exceptions to the free speech clause in the First Amendment, but "hate speech" is not one of them. The Supreme Court has repeatedly held up the right of an individual or group to engage in speech that much of the public likely finds offensive, like displaying swastikas, burning crosses, or protesting a soldier's funeral.
<b>Evidence date</b>	2024-07-01
<b>Relevant</b>	true
<b>Stance</b>	refutes

---

<b>Evidence #2</b>	Rising temperatures push bees to their physiological limits, and could cause the extinction of bee populations.
<b>Evidence date</b>	2024-06-13
<b>Relevant</b>	false
<b>Stance</b>	not_applicable

---

<b>Evidence #3</b>	There are some exceptions to the free speech clause in the First Amendment.
<b>Evidence date</b>	2020-02-22
<b>Relevant</b>	true
<b>Stance</b>	insufficient-neutral

---

<b>Evidence #4</b>	The first amendment states that hate speech is not protected in most situations.
<b>Evidence date</b>	2008-12-10
<b>Relevant</b>	true
<b>Stance</b>	insufficient-supports

---

<b>Evidence #5</b>	The first amendment states that hate speech is protected in some situations.
<b>Evidence date</b>	2011-09-05
<b>Relevant</b>	true
<b>Stance</b>	insufficient-refutes

---

<b>Evidence #6</b>	Hate speech is protected by the first amendment, he stated. Actually, hate speech is not protected by the first amendment, she stated.
<b>Evidence date</b>	1999-03-08
<b>Relevant</b>	true
<b>Stance</b>	insufficient-contradictory

---

<b>Evidence #7</b>	This claim is false.
<b>Evidence date</b>	2015-08-28
<b>Relevant</b>	true
<b>Stance</b>	refutes

---

<b>Evidence #8</b>	John Doe stated in a Facebook post that hate speech is not protected by the first amendment.
<b>Evidence date</b>	2015-08-28
<b>Relevant</b>	true
<b>Stance</b>	insufficient-neutral

---

<b>Evidence #9</b>	Hate speech is protected by the first amendment. [...] A social media post stated that hate speech is not protected by the first amendment.
--------------------	---

Figure 12: Page 4 of 10 depicting the annotation interface.

---

<b>Claim</b>	Hate speech is not protected by the first amendment.
<b>Claimant</b>	John Doe
<b>Claim date</b>	2024-05-04

---

<b>Evidence date</b>	2024-07-01
<b>Relevant</b>	true
<b>Stance</b>	refutes

---

Examples of a claim with corresponding evidence pieces and annotations. Fields in pink are to be annotated and fields in gray are provided for the annotation task. 'Claim date' denotes the publish date of the claim. 'Evidence date' denotes the publish date of the evidence.

### Additional example #1

---

<b>Claim</b>	Says gubernatorial candidate Rebecca Kleefisch "worked with Scott Walker to sign five abortion restrictions into law that took away services and threatened doctors with prison time for providing safe and legal abortions."
<b>Claimant</b>	Kelda Roys
<b>Claim date</b>	2021-09-29

---

<b>Evidence</b>	Rebecca Kleefisch, Former Lt. Gov. Under Scott Walker, Launches Gubernatorial Campaign MADISON, Wis. (AP) — Republican Rebecca Kleefisch, who spent eight years as lieutenant governor under Scott Walker, launched her campaign for Wisconsin governor on Thursday, casting herself as someone who will "fight for you" while deriding the Democratic incumbent as a weak failure.
<b>Evidence date</b>	2021-09-09
<b>Relevant</b>	true
<b>Stance</b>	insufficient-neutral

---

### Additional example #2

---

<b>Claim</b>	"Vitamin D3 is radiated sheep's wool mixed with chloroform", "Vitamin D3 IS RAT POISON !!", "D levels are low because of lack of light and high vitamin A diets."
<b>Claimant</b>	Evan Torrens
<b>Claim date</b>	2024-05-24

---

<b>Evidence</b>	Vitamin D3 supplements are safe for treating vitamin D deficiency; comparison to rat poison is misleading
<b>Evidence date</b>	2024-05-24
<b>Relevant</b>	true
<b>Stance</b>	refutes

---

Keyboard Input:

←	→
Move backward	Move forward

Move backward	Move forward
---------------	--------------

Figure 13: Page 5 of 10 depicting the annotation interface.

consent

## Consent

Please indicate your consent towards participating in this study. If you do not consent and wish to drop out of the study, please return this submission on Prolific by selecting the 'stop without completing' button' and close this page.

I have read and understood the information about the study.

Yes  
 No

I have read and understood the study instructions.

Yes  
 No

I want to participate in this research and continue with the study.

Yes  
 No

Keyboard Input:

←	→
Move backward	Move forward

Figure 14: Page 6 of 10 depicting the annotation interface.



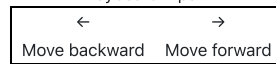
pep\_talk

## Before you continue

We are happy that you want to participate in this study! Before you continue, we wanted to share some advice to help make the annotation process as smooth as possible:

- You can go back and forth between pages and samples using the buttons at the bottom of the webpage. **Do not use the browser 'back' button to go back to a page** as it will make this webpage behave weirdly.
- By now, you should have been presented with an introduction page with general guidelines, an instruction page with annotation examples and a page to indicate your consent to participate in this study. If you have missed any of these pages, we recommend that you go back now to make sure that you have seen all necessary pages.
- We recommend that you **store all instructions to make it easier to go back to them**. This can for example be done by printing the webpage to a PDF file (select 'Print' in your browser and then 'Save as PDF') or taking a screenshot.
- There are **tooltips** for each answer alternative. When you hover over an alternative with your cursor, more information about that answer alternative will be shown.
- **Take your time with the annotations.** We understand that the instructions and samples presented may be time-consuming to process. Feel free to go back to the instructions whenever you are uncertain about some sample. We are thankful for your effort and will make sure to adapt the compensation for the annotation work in case it turns out to take more time than initially estimated.

Keyboard Input:



---

Copyright © 2022 [Blablablab](#)

[Fork on GitHub](#) | [Cite Us](#)

Figure 15: Page 7 of 10 depicting the annotation interface.

**Claimant:** Facebook posts  
**Claim date:** 2021-03-18  
**Claim:** "Pelosi's \$1.9 trillion bailout gives EVERY federal employee a \$21,000 bonus check... they never lost their job!"  
**Evidence date:** 2021-03-18  
**Evidence:** The law allocates money for an expanded paid-leave fund for federal workers dealing with certain COVID-19-related matters. There is no bonus check. It covers leave that would otherwise be unpaid.

Is the evidence relevant? Does the evidence contain any information that 1) directly supports or refutes the claim, 2) is topically related to the topic or entities of the claim or claimant (same people, places, organisations, etc.), or 3) can be seen as implicitly referring to the claim?

True  
 False

What is the stance of the evidence? Each provided evidence should correspond to one of the stances listed below. Evidence marked as relevant=False should be annotated as 'not\_applicable'.

supports  
 insufficient-supports  
 insufficient-neutral  
 insufficient-contradictory  
 insufficient-refutes  
 refutes  
 not\_applicable

Was there a quality issue with this sample that prevented you from annotating it as instructed? If so, shortly describe the issue here. Leave this box empty if there was no issue.

Keyboard Input:

←	→	1	2	3	4	5	6	7	8	9
Move backward	Move forward	True	False	supports	insufficient-supports	insufficient-neutral	insufficient-contradictory	insufficient-refutes	refutes	not_applicable

Copyright © 2022 [Blablalab](#)

[Fork on GitHub](#) | [Cite Us](#)

Figure 16: Page 8 of 10 depicting the annotation interface.

**Claimant:** Chervin Jafariieh, Social media user  
**Claim date:** 2024-06-06  
**Claim:** "Cat parasite" causes toxoplasmosis; toxoplasmosis increases ghrelin hormone levels  
**Evidence date:** 2024-06-13  
**Evidence:** Toxoplasmosis is caused by an infection with the parasite *Toxoplasma gondii*. This parasite is most commonly found in contaminated food (especially raw and undercooked meat), but it can also be found in cat feces. Most *T. gondii* infections in people with healthy immune systems are asymptomatic and resolve on their own within a few weeks or months. Ghrelin is a hormone produced by the stomach and often called the "hunger hormone". There is currently no evidence supporting an association between toxoplasmosis and increased ghrelin hormone levels.

A Facebook reel shared on 6 June 2024 claimed that a so-called "cat parasite" causes toxoplasmosis. It also claimed that toxoplasmosis leads to an increased release of the ghrelin hormone (commonly known as the "hunger hormone"), which in turn leads a host of other health issues including "depression, highs and lows, irregular heartbeat, muscle twitching, brittle hair, dry skin, psoriasis, [and] unregulated sleep cycles".

The "cat parasite" claim has received media attention in the past, primarily addressing the potential link between *T. gondii*, mental illness, and risk-taking behavior. However, there is no apparent evidence for an association between toxoplasmosis and an increase in the ghrelin hormone. We explain below.

<p>Is the evidence relevant? Does the evidence contain any information that 1) directly supports or refutes the claim, 2) is topically related to the topic or entities of the claim or claimant (same people, places, organisations, etc.), or 3) can be seen as implicitly referring to the claim?</p> <p><input type="radio"/> True  <input type="radio"/> False</p>	<p>What is the stance of the evidence? Each provided evidence should correspond to one of the stances listed below. Evidence marked as relevant=False should be annotated as 'not_applicable'.</p> <p><input type="radio"/> supports  <input type="radio"/> insufficient-supports  <input type="radio"/> insufficient-neutral  <input type="radio"/> insufficient-contradictory  <input type="radio"/> insufficient-refutes  <input type="radio"/> refutes  <input type="radio"/> not_applicable</p>	<p>Was there a quality issue with this sample that prevented you from annotating it as instructed? If so, shortly describe the issue here. Leave this box empty if there was no issue.</p> <div style="border: 1px solid black; height: 20px; width: 100%;"></div>
---	--	--

Keyboard Input:

←	→	1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---	---	---

Figure 17: Page 9 of 10 depicting the annotation interface.

Move backward	Move forward	True	False	supports	insufficient-supports	insufficient-neutral	insufficient-contradictory	insufficient-refutes	refutes	not_applicable
---------------	--------------	------	-------	----------	-----------------------	----------------------	----------------------------	----------------------	---------	----------------

---

Copyright © 2022 [Blablalab](#)

[Fork on GitHub](#) | [Cite Us](#)

Figure 18: Page 10 of 10 depicting the annotation interface.