CUB: Benchmarking Context Utilisation Techniques for Language Models

Anonymous ACL submission

Abstract

Incorporating external knowledge is crucial for knowledge-intensive tasks, such as question answering and fact checking. However, language models (LMs) may ignore relevant information that contradicts outdated parametric memory or be distracted by irrelevant contexts. While many context utilisation manipulation techniques (CMTs) have recently been proposed to alleviate these issues, few have seen systematic comparison. In this paper, we develop CUB (Context Utilisation Benchmark) – the first comprehensive benchmark designed to help practitioners within retrieval-augmented generation (RAG) diagnose CMTs under different context conditions. With this benchmark, we conduct the most extensive evaluation to date of seven state-of-the-art methods, representative of the main categories of CMTs, across three diverse datasets and tasks, applied to nine LMs. Our results reveal that most existing CMTs struggle to handle the full spectrum of context types encountered in real-world retrieval-augmented scenarios. We also find that many CMTs display inflated performance on simple synthesised datasets, compared to more realistic datasets with naturally occurring samples. Our findings expose critical gaps in current CMT evaluation practices and demonstrate the need for holistic testing and the development of CMTs that can robustly handle multiple context types.

1 Introduction

Context utilisation is a key component of language models (LMs) used for retrieval-augmented generation (RAG), as the benefits of retrieving external information are only realised if the generative model makes adequate use of the retrieved information. While recent research has identified many benefits of augmenting LMs with retrieved information (Shuster et al., 2021; Hagström et al., 2023), it has also identified weaknesses of LMs used for RAG, of which many are associated with context utilisation. For example, LMs can easily be distracted by irrelevant contexts (Shi et al., 2023) or ignore relevant contexts due to memory-context conflicts (Xu et al., 2024). The robustness of LMs to irrelevant contexts is important as information retrieval systems used for RAG are not guaranteed to always retrieve relevant information. Moreover, as information may be updated to conflict with the training data of the LM, the model should prioritise the most recently updated information.

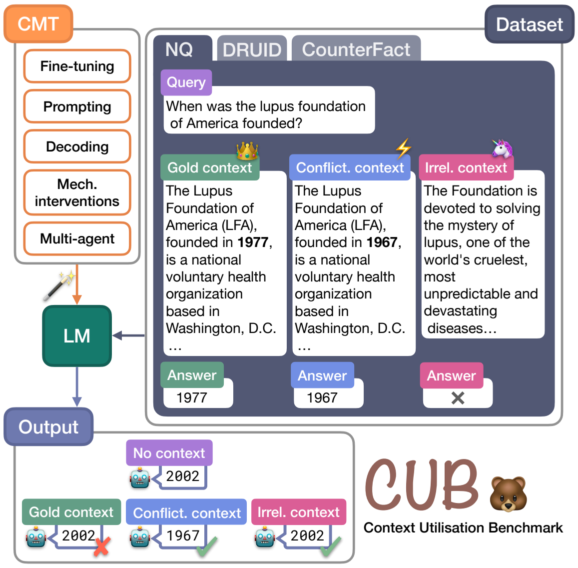

Figure 1: The Context Utilisation Benchmark. We evaluate a range of LMs under different CMTs on samples from NQ, DRUID and CounterFact for gold, conflicting and irrelevant contexts.

As a consequence, many different methods for increasing or suppressing LM context utilisation, henceforth referred to as *CMTs* (Context utilisation Manipulation Techniques), have been proposed. The methods encompass a broad range of approaches, from different decoding methods (Shi et al., 2024; Kim et al., 2024) to fine-tuning methods (Li et al., 2023), prompting (Liu et al., 2023), multi-agent (Feng et al., 2024; Du et al., 2024), and mechanistic interventions (Ortu et al., 2024; Jin et al., 2024). Many of these CMTs are developed with distinct objectives in mind, often targeting a specific aspect of context utilisation – such as improving robustness to irrelevant contexts or enhancing faithfulness to conflicting information. While each method yields promising results in isolation, their evaluation is often limited to settings that closely match the method's target objective, making it unclear whether these approaches will generalise or remain effective in real-world RAG scenarios. To address this pressing evaluation gap, we develop the first comprehensive CMT benchmark to systematically test and compare different CMTs across datasets that truly represent the diversity of real-world domains and tasks (Figure 1). Our **contributions** are as follows:

- We develop CUB (Context Utilisation Benchmark) – the first comprehensive benchmark designed to enable rigorous evaluation and comparison of CMTs under various context conditions (§3).[1] CUB evaluates CMTs across gold, conflicting, and irrelevant contexts that capture key RAG challenges.
- We conduct the most extensive evaluation to date of CMTs, assessing seven state-of-the-art methods representative of the main categories of CMTs (§4) across our benchmark (§6). The study encompasses approximately 750 experimental data points, spanning 9 LMs, 7 CMTs, 3 datasets, and 3 different context types.
- We provide insights into what CMT works best for different scenarios and identify critical areas for improvement. Our analysis reveals that existing CMTs face fundamental trade-offs – they struggle to optimise performance across all context types. This points to a need for next-generation CMTs that can robustly handle diverse context conditions.

2 Related Work

Context-intensive datasets We consider two main categories of context-intensive datasets: 1) datasets representing *knowledge-intensive tasks*, i.e. tasks for which access to external context is crucial, and 2) datasets designed to *diagnose* model adaptability to external knowledge. Examples of datasets representative of knowledge-intensive tasks are Natural Questions (NQ), DRUID, the KILT datasets and PubMedQA (Kwiatkowski et al., 2019; Hagström et al., 2024; Petroni et al., 2021; Jin et al., 2019). Examples of diagnostic datasets representative of the latter category are CounterFact and ConflictQA (Meng et al., 2022; Xie et al., 2024a). These datasets contain synthesised queries based on fact triplets from LAMA (Petroni et al., 2019) (e.g. Thomas Ong-citizen of-Singapore) for which 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"). Diagnostic 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).

Previous work has typically evaluated different CMTs on either of the dataset categories in isolation, creating a fragmented understanding of CMT effectiveness. CUB bridges this critical gap by incorporating datasets representative of both knowledge-intensive tasks and diagnostic datasets, enabling the first comprehensive evaluation of CMTs across truly diverse settings.

CMTs Many context utilisation manipulation techniques have recently been proposed. Existing CMTs can be categorised into one of four main groups based on *intervention level*, i.e. what aspect of the model they manipulate. 1) *fine-tuning* CMTs update model parameters to modify context utilisation. For example, fine-tuning on distracting contexts was found to yield improved robustness to distracting contexts (Li et al., 2023; Shen et al., 2024; Yoran et al., 2024). Fang et al. (2024) specifically focus on different types of retrieval noise likely to be encountered in real-world environments and develop a fine-tuning approach to handle these. 2) *prompting techniques* modify the input to the LM to improve context utilisation, representing minimally modified settings. 3) *mechanistic interventions* on the LM modify certain model components at inference time to alter context utilisation. Examples involve attention modification (Ortu et al., 2024; Jin et al., 2024) and SpARe interventions (Zhao et al., 2025). Lastly, 4) *decoding methods* involve a modified decoding approach, applied to the output logits, to manipulate context utilisation. Examples include context-aware contrastive decoding (Yuan et al., 2024; Kim et al., 2024; Shi et al., 2024; Wang et al., 2024; Zhao et al., 2024) and lookback lens decoding (Chuang et al., 2024).

Despite significant progress in developing new

[1] Code will be made available upon publication.

CMTs, systematic evaluations remain scarce. Prior studies have largely assessed individual CMTs in isolation, likely due to the lack of a unified benchmark. In this paper, we address this gap by conducting the first comprehensive evaluation of seven representative CMTs across the four main categories of CMTs, using the CUB dataset.

Benchmarks To our knowledge, there is not yet a benchmark specifically designed for CMTs, representing a significant research gap. The closest examples of existing benchmarks are RAG-Bench by Fang et al. (2024), KILT by Petroni et al. (2021) and AxBench by Wu et al. (2025). However, these serve different purposes: the first evaluates the retrieval-noise robustness of LMs, the second assesses performance of RAG systems as a whole, and the latter focuses on steering techniques for LMs, emphasising safety and reliability rather than context utilisation. CUB addresses this critical infrastructure gap by creating the first comprehensive and purpose-built benchmark specifically for the evaluation of CMTs, taking inspiration from these existing benchmarks while addressing the unique challenges of CMT evaluation.

3 CUB: A Context Utilisation Benchmark

Given a CMT, CUB is designed to systematically test the technique across different datasets, models and metrics, providing rigorous evaluation. To ensure fair and meaningful comparisons, CUB incorporates a carefully designed pre-defined method for hyperparameter search, eliminating potential bias in method comparisons.

3.1 Language Models

CUB evaluates the model sensitivity of CMTs across a range of nine different LMs, providing insights into how CMT effectiveness varies with model architecture and scale. The comprehensive model coverage includes both open-sourced models – GPT-2 XL, Pythia (6.9B), Qwen2.5 (1.5B, 7B, and 32B) (Radford et al., 2019; Biderman et al., 2023; Yang et al., 2024) – and API-based Cohere Command A (111B), which is specialised for RAG.[2] For the Qwen models, we strategically include both base and instruction-tuned variants to capture the impact of instruction tuning on context utilisation. This carefully curated model selection enables comprehensive comparisons across model families, model sizes, instruction-tuning approaches, and deployment paradigms.

[2] https://cohere.com/blog/command-a

Dataset	Split	#samples	%Gold	%Conflict.	%Irrel.
CounterFact	*dev*	198	33.3	33.3	33.3
	test	2,499	33.3	33.3	33.3
NQ	*dev*	198	33.3	33.3	33.3
	test	4,945	33.4	33.1	33.4
DRUID	*dev*	198	33.3	33.3	33.3
	test	4,302	43.5	56.1	0.4

Table 1: Statistics of the datasets that form CUB. 'Conflict.' and 'Irrel.' denote conflicting and irrelevant contexts, respectively.

3.2 Datasets

To systematically evaluate how CMTs respond to different types of contextual information, CUB evaluates each CMT across CounterFact, NQ and DRUID (see Table 1). The inclusion of these datasets is based on three key criteria that ensure comprehensive coverage: (i) diversity in task difficulty, (ii) diversity in realistic versus synthesised RAG scenarios, and (iii) high utilisation in related work, ensuring our findings connect to the broader literature. CounterFact represents a controlled causal language modelling task with carefully synthesised counterfactual contexts designed to conflict with model memory. NQ represents a popular and more realistic setup focused on RAG for open-domain QA of substantially greater difficulty, with contexts sampled from Wikipedia. DRUID is a cutting-edge dataset representing another critical RAG task – automated fact-checking – which requires sophisticated reasoning based on naturally occurring claims and evidence from the internet.

For each dataset, we curate samples representative of the three types of contexts that are fundamental to realistic RAG scenarios: 1) **gold** contexts that are relevant, 2) **conflicting** contexts that are relevant but always contradict the LM's memory, and 3) **irrelevant** contexts that provide no information to solve the given question (Fang et al., 2024). This approach ensures that CUB captures the full spectrum of challenges to RAG in real-world deployments. More details on the data collection process can be found below and in Appendix B.

CounterFact To construct a CounterFact dataset with counterfactual contexts, we first identify samples from LAMA that have been memorised by Pythia 6.9B, following the approach by Saynova et al. (2025). We base the CounterFact dataset on

Pythia to obtain a set of samples likely to have been memorised by all CUB models, since LMs have been found to memorise more facts as they grow in size (Saynova et al., 2025). We confirm this in Appendix B; all CUB LMs are found to have memorised at least 70% of the CounterFact samples. Based on the known fact triplets, we sample conflicting contexts following the approach of Meng et al. (2022). We also sample gold contexts that simply state the correct triplet. For the irrelevant contexts, we randomly sample fact triplets unrelated to the sample query.

NQ The gold context samples are simply the original NQ samples. For the collection of samples with conflicting contexts, we follow a substitution approach inspired by the method of Longpre et al. (2021). For the collection of samples with irrelevant contexts, we apply a LM re-ranker to identify the most relevant non-gold paragraph from the Wikipedia page in which the gold context was found. With this approach, we collect irrelevant adversarial contexts representative of real-world RAG scenarios.

DRUID The ⟨claim, evidence⟩ samples of DRUID have been manually annotated for stance of the evidence (supports, refutes, insufficient or irrelevant). We map stance to context type as described in Appendix B. Since DRUID represents a reasoning task, asking the model whether provided evidence supports the claim under consideration (True or False), or is insufficient (None), the output space for the samples is limited to three tokens.

3.3 Metrics

Similarly to Jin et al. (2024) we use a binary score to measure context utilisation. We refer to it as the *binary context utilisation* (BCU) score and define it as follows. For relevant contexts (gold and conflicting) the score is 1 if the LM prediction is the same as the token promoted by the context, t_C, and 0 otherwise. For irrelevant contexts the score is 1 if the LM prediction is the same as the memory token, t_M, (i.e. the prediction made by the model before any context has been introduced) and 0 otherwise. We report the averaged BCU score per context type. To assess the relative effectiveness of CMTs, we also report the net gain of each CMT, compared to when no CMT is applied, using BCU score ($\Delta = \text{BCU}_{\texttt{CMT}} - \text{BCU}_{\texttt{Regular}}$). We also consider *continuous context utilisation*, CCU, a more fine-grained metric that measures the change in outputted token probabilities as context is introduced, further described in Appendix C.

Methods	Objective	Level	Tuning Cost	Inference Cost
Fine-tuning	Both	Fine-tuning	High	Low
Prompting	Both	Prompt.	Low	Mid
Multi-agent	Both	Prompt.	None	High
PH3 +context	Faith	Mech.	High	Low
COIECD	Faith	Decoding	Mid	Mid
PH3 +memory	Robust	Mech.	High	Low
ACD	Robust	Decoding	None	Mid

Table 2: Comparison of CMTs by objective, intervention level, and cost. The CMTs are coloured by objective with warm colours for 'Both', blue for 'Faith' and green for 'Robust'. 'Mech.' denotes mechanistic interventions.

3.4 Hyperparameter Search

For CMTs requiring hyperparameter tuning, we use the validation set of each dataset to select values that maximise the average BCU across all context types, unless a method-specific tuning procedure is explicitly specified (see Appendix D). This ensures a fair comparison between CMTs.

4 Context Utilisation Manipulation Techniques

We benchmark a comprehensive set of seven state-of-the-art CMTs on CUB, strategically selected to represent all major categories of CMTs and provide unprecedented coverage of the CMT landscape. Table 2 provides a detailed overview of the key characteristics of each CMT, including their primary objectives, intervention levels, and computational costs for both tuning and inference. We adaptively select the appropriate set of LMs for each CMT based on technical compatibility. As a baseline for comparison, we also evaluate regular LMs on identical inputs with no CMT applied (`Regular`), enabling direct assessment of CMT effectiveness.

Fine-tuning We adapt the approach of Li et al. (2023), which fine-tunes LMs to ensure the usage of relevant contexts (see Appendix E). It considers four different types of contexts: relevant, irrelevant, empty, and conflicting contexts.

Prompting We curate a set of 12 prompts for each evaluation dataset, tailoring prompt selection to each evaluated model's characteristics. Each prompt set combines human expertise and AI generation: 6 prompts are curated by human experts following established methodologies (Jin et al., 2024), while 6 prompts are generated by advanced LLMs,[3] following established prompt engineering approaches (Wu et al., 2025).

Multi-agent Inspired by LM agents and self-refinement (Du et al., 2024; Feng et al., 2024; Madaan et al., 2023), which are widely adopted techniques in reasoning tasks, we decompose context utilisation into two components – relevance and context faithfulness – and assign each as a separate task to an individual LM agent (see Appendix F). We aim to examine whether LMs are capable of accurately evaluating context relevance and answer faithfulness, to subsequently self-correct themselves for improved faithfulness to relevant contexts.

Mechanistic interventions: PH3 We adopt the PH3 method by Jin et al. (2024) (see Appendix G). PH3 can be used in two different modes – suppressing context attention heads (`PH3 +memory`) or suppressing memory attention heads (`PH3 +context`).

Context-aware contrastive decoding: ACD and COIECD Contrastive decoding approaches adjust the model's output distribution based on two distributions: one for which only the query is given as input and one for which the context also is included. Among them, *contextual information-entropy constraint decoding* (`COIECD`; Yuan et al., 2024) is designed to detect the presence of knowledge conflicts and selectively resolve them, aiming to improve faithfulness to conflicting context without compromising performance when no conflict exists. In contrast, *adaptive contrastive decoding* (`ACD`; Kim et al., 2024) addresses the challenge of irrelevant context by using entropy-based weighting to adaptively ensemble parametric and contextual distributions. We test both on CUB to cover the nuance in decoding approaches.

5 Features Impacting Context Utilisation

To deepen our understanding of the results on CUB, we complement the benchmark with a sophisticated analysis of features likely to impact context utilisation. Our goal is to uncover the fundamental factors that determine *why* certain CMTs and LMs achieve superior performance. We systematically investigate features at both model and input levels, providing unprecedented insights into the mechanisms underlying context utilisation.

[3] Mainly by ChatGPT, but also by Microsoft Co-pilot.

5.1 Model Features

By virtue of the unprecedented LM coverage in CUB, we can systematically measure multiple salient model features. We analyse **model size**, whether the model is **instruction-tuned**, and **strength of model memory**. To control for external confounders related to model family and implementation differences, we carefully measure correlations with model size and instruction-tuning exclusively across Qwen models. Strength of model memory is quantified as the softmaxed logits for the top token predicted by the LM when only the query is provided (without context).

5.2 Input Features

We measure multiple input characteristics found to impact context utilisation for humans and/or LMs (see Appendix H). By considering **context length** and **Flesch reading ease score**, we aim to measure whether the context is *difficult to understand* (Gao et al., 2024; Vladika and Matthes, 2023). Using **distractor rate**, we aim to measure whether the context contains *distracting information* (Shaier et al., 2024). With **query-context overlap** we also aim to measure *query-context similarity* (Wan et al., 2024). Lastly, we check the **answer position** (Liu et al., 2024) and if the evaluated LMs find the context **relevant**.

5.3 Metric for Feature Impact

By virtue of the unified and controlled setup of CUB, we can systematically study correlation coefficients to investigate the impact of different input and model features with minimal risk of confounding factors. We employ Spearman's ρ to measure the impact of features on context utilisation, proxied by BCU, providing robust statistical insights into the factors that drive CMT effectiveness.

6 Main Results on CUB

The CUB results can be found in Figures 2 and 3; CCU scores together with additional results and analyses can be found in Appendix A. We structure our analysis around a set of key findings about CMT effectiveness.

6.1 Overall Trends

We first note that the BCU and CCU scores in Figures 2 and 4, respectively, support the same trends and focus the analysis on the BCU results.

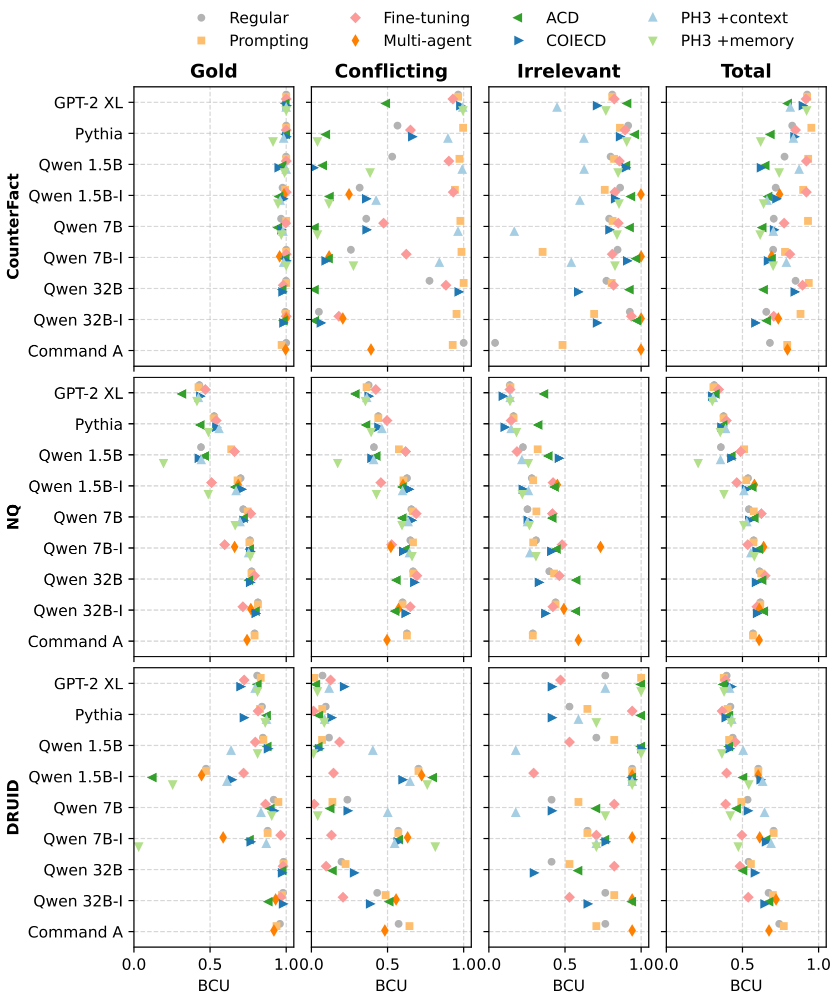

Figure 2: BCU scores for the evaluated context utilisation manipulation methods applied to the evaluated models and datasets. 'Total' denotes the averaged performance across all context types.

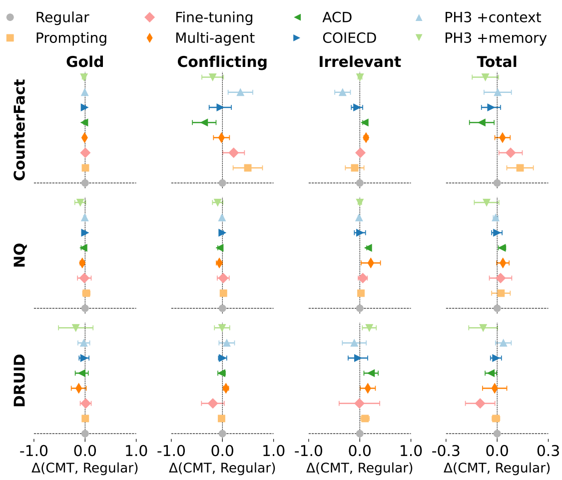

Figure 3: Model-averaged relative performance (Δ) of each CMT compared to `Regular` across datasets and context types. The horizontal bars represent the standard deviation.

Context utilisation scales with model size – but with surprising exceptions. From Figure 2, we observe that larger `Regular` LMs generally achieve superior performance across all context types for NQ and DRUID, following expected scaling laws. On NQ, the best performing model is Qwen 32B, while on DRUID, Command A leads performance. Remarkably, our results reveal that applying the right CMT to a smaller LM can achieve context utilisation performance on par with much larger regular LMs – for instance, `Fine-tuning` applied to Qwen 7B matches the performance of `Regular` Qwen 32B on NQ. However, CounterFact reveals a striking counter-intuitive pattern: `Regular` model performance across all contexts actually *decreases* as model size increases. We attribute this unexpected phenomenon to the artificial nature of the dataset, which appears to confuse larger LMs. This finding, combined with the fact that NQ and DRUID datasets demand greater model capacity, underscores a critical insight: evaluating context utilisation solely on simple datasets like CounterFact provides a fundamentally misleading picture of CMT effectiveness.

Most CMTs exhibit inflated performance on conflicting CounterFact contexts. We observe a striking phenomenon: all LMs that do not already achieve perfect BCU scores on conflicting CounterFact contexts leap to a perfect score of 1.0 under `Prompting`, `PH3 +context`, and `Fine-tuning`. However, these same CMTs fail to deliver comparable improvements on the more realistic NQ or DRUID datasets. These results expose a critical flaw in current evaluation practices – CMTs that appear highly effective in simplified settings may fail to generalise to real-world complexity. This finding demonstrates the essential need for holistic evaluation frameworks like CUB.

6.2 CMT Comparison

We further assess whether the CMTs consistently outperform `Regular` across different context types. Figure 3 shows the average Δ of each CMT, aggregated over all evaluated models. A value above zero indicates that the CMT yields a net improvement over `Regular`, whereas a negative value highlights cases where the CMT degrades performance.

We uncover a fundamental trade-off between sensitivity to relevant contexts versus robustness to irrelevant contexts. Our analysis reveals a striking pattern: each CMT exhibits inherent trade-offs across context types, with overall Δ values (Total) converging to near zero across NQ and DRUID. This convergence exposes that *no single CMT emerges as universally superior*. For instance, `PH3 +context` demonstrates consistent improvements over `Regular` in conflicting contexts, but significantly underperforms when applied to irrelevant contexts. Conversely, `ACD`, while effectively handling irrelevant contexts, shows degraded performance in conflicting context scenarios.

In realistic RAG scenarios, the type of context provided to the LM cannot be predicted in advance. Therefore, the ideal CMT must perform optimally across *all* context types. Our comprehensive evaluation reveals that while we have identified CMTs that excel with either relevant or irrelevant contexts in isolation, *no existing CMT can robustly handle both relevant and irrelevant contexts simultaneously*. This represents a fundamental gap in current CMT capabilities that needs to be addressed.

Meanwhile, prompting-based CMTs, such as `Prompting` and `Multi-agent`, demonstrate remarkably stable performance across context types, avoiding the dramatic performance drops observed with other approaches. Compared to more complex CMTs, they offer this robustness with lower optimisation and implementation costs, making them attractive for practical deployment. `Multi-agent` shows particularly promising results with clear gains in irrelevant contexts, though its efficacy remains limited in gold and conflicting settings. This pattern suggests that while LMs demonstrate competence in identifying irrelevant contexts, they face fundamental limitations in effectively utilising relevant information.

6.3 Impact of Model and Input Features

See Tables 3 and 7 for Spearman's ρ between BCU and the features described in Section 5. Our findings are as follows.

Larger LMs perform better on NQ and DRUID. Corroborating our findings in Section 6.1, we observe a positive correlation with model size ($\rho \approx 0.3$) on DRUID gold contexts. `Multi-agent` also works significantly better with bigger LMs on DRUID gold contexts ($\rho = 0.42$). In addition, we observe a positive correlation with model size on NQ gold contexts ($\rho \in [0.20, 0.37]$).

Instruction-tuning is beneficial for conflicting and irrelevant DRUID contexts. We note how instruction tuning generally correlates with improved performance on conflicting and irrelevant DRUID contexts ($\rho \in [0.29, 0.77]$ depending on CMT). The conflicting DRUID contexts frequently require the LM to be able to abstain (i.e. respond with a 'None') when presented with insufficient contexts, which is something instruction-tuned models may be more adept at.

Conversely, instruction-tuning is clearly detrimental for conflicting CounterFact contexts ($\rho \leq -0.36$), potentially because this makes the LMs more critical of unreliable information, as opposed to performing pure causal language modelling.

A strong model memory corresponds to high performance on irrelevant contexts from NQ and CounterFact. We observe high correlations ($\rho \approx 0.36$) between memory strength and robustness to irrelevant contexts for `Regular` on CounterFact and NQ. These correlations increase when `Fine-tuning`, `ACD` or `Prompting` is applied. Fur-

Dataset	Context	CMT	Corr.
Model size			
DRUID	Gold	Multi-agent	**0.42**
DRUID	Gold	ACD	**0.41**
NQ	Gold	PH3 +memory	**0.37**
DRUID	Gold	Regular	**0.36**
DRUID	Gold	Prompting	**0.36**
NQ	Conflicting	PH3 +memory	**0.33**
NQ	Gold	Regular	**0.20**
NQ	Irrelevant	Regular	**0.14**
NQ	Conflicting	Regular	**0.09**
CounterFact	Gold	Regular	**0.04**
CounterFact	Irrelevant	Regular	0.02
CounterFact	Conflicting	Regular	-0.01
DRUID	Conflicting	Regular	**-0.08**
DRUID	Irrelevant	Regular	**-0.20**
DRUID	Irrelevant	PH3 +memory	**-0.33**
CounterFact	Conflicting	Fine-tuning	**-0.33**
DRUID	Irrelevant	COIECD	**-0.44**
Instruct tuned			
DRUID	Conflicting	PH3 +memory	**0.77**
DRUID	Irrelevant	PH3 +context	**0.65**
DRUID	Conflicting	ACD	**0.54**
DRUID	Conflicting	Prompting	**0.46**
DRUID	Conflicting	Regular	**0.40**
DRUID	Conflicting	COIECD	**0.34**
DRUID	Irrelevant	Regular	**0.29**
NQ	Gold	Regular	**0.13**
CounterFact	Irrelevant	Regular	**0.12**
NQ	Irrelevant	Regular	**0.06**
NQ	Conflicting	Regular	**0.05**
CounterFact	Gold	Regular	0.01
DRUID	Gold	Regular	**-0.19**
CounterFact	Conflicting	Regular	**-0.36**
DRUID	Gold	ACD	**-0.38**
CounterFact	Conflicting	PH3 +context	**-0.43**
DRUID	Gold	PH3 +memory	**-0.72**
Strength of memory			
DRUID	Conflicting	PH3 +memory	**0.54**
NQ	Irrelevant	Fine-tuning	**0.47**
NQ	Irrelevant	ACD	**0.39**
CounterFact	Irrelevant	Fine-tuning	**0.39**
NQ	Irrelevant	Prompting	**0.39**
NQ	Irrelevant	COIECD	**0.38**
DRUID	Conflicting	ACD	**0.37**
NQ	Irrelevant	Regular	**0.37**
CounterFact	Irrelevant	Regular	**0.35**
DRUID	Conflicting	Prompting	**0.34**
CounterFact	Irrelevant	ACD	**0.32**
CounterFact	Irrelevant	PH3 +memory	**0.31**
CounterFact	Irrelevant	COIECD	**0.30**
DRUID	Conflicting	Regular	**0.26**
NQ	Gold	Regular	**0.18**
DRUID	Irrelevant	Regular	0.15
NQ	Conflicting	Regular	**0.09**
CounterFact	Gold	Regular	**0.04**
DRUID	Gold	Regular	0.02
CounterFact	Conflicting	ACD	**-0.31**
CounterFact	Conflicting	COIECD	**-0.42**
DRUID	Gold	PH3 +memory	**-0.43**
CounterFact	Conflicting	Regular	**-0.44**

Table 3: Spearman's ρ between BCU and model features. Correlation values for `Regular` or with an absolute value above 0.3 are shown. Correlation values with an absolute value below 0.3 are marked in gray. Significant correlation values (p-value < 0.05) are marked in **bold**. Results are measured across models.

thermore, we observe for CounterFact how strong `Regular` model memory correlates with low performance on conflicting contexts ($\rho = -0.44$). This is expected – previous work has already shown how LMs are resistant to synthesised contexts that contradict the internal model memory (Longpre et al., 2021; Xie et al., 2024a).

Answer position matters little for context utilisation. We measure low correlation values (below 0.3) across all settings for answer position in the context and Flesch reading ease score, and have thus omitted them in Table 7. Previous work has already found the Flesch reading ease score to show low correlations with LM context utilisation; our work further supports this finding (Hagström et al., 2024). Liu et al. (2024) found the answer position impactful for the utilisation of long contexts. CUB does not contain equally long contexts, which potentially explains why we do not see the same impact of answer position.

Context utilisation on gold NQ contexts is degraded on long contexts with high distractor rates. We measure weak negative correlations with context length ($\rho = -0.23$) and distractor rate ($\rho = -0.19$) with respect to `Regular` performance on gold NQ contexts. This is expected – long gold contexts or contexts with a high rate of distractors should be more difficult to process and utilise. We hypothesise the fairly low correlation levels are a consequence of each feature alone not being sufficiently predictive of model performance.

7 Conclusion

We present CUB, the first comprehensive benchmark for systematically evaluating CMTs across diverse contexts, datasets, and model architectures. Using CUB, we conduct the most extensive comparison to date of seven representative CMTs, revealing a core trade-off between robustness to irrelevant context and effective use of relevant information. Our analysis shows that model characteristics play a dominant role in context utilisation, while input features alone are weak predictors. Crucially, we uncover major flaws in current evaluation practices – many CMTs perform well on synthetic datasets but falter in realistic settings. These findings underscore the need for holistic evaluation frameworks and set a new standard for developing context manipulation techniques in real-world RAG environments.

Limitations

While context can be characterised by various dimensions, including its length or the number of documents provided, our current study specifically focuses on context type. We view the investigation of longer or multi-document contexts as important fields for further research, yet orthogonal to our current focus, as explained below.

Long-context settings (e.g., over 4k tokens) have recently garnered significant attention (Liu et al., 2024), and evaluating CMTs in these scenarios is an important direction for future research. However, this paper focuses on a different set of challenges – those associated with standard-length contexts – making long-context evaluation out of scope. Long-context tasks involve distinct issues in context utilisation and require different evaluation strategies and CMTs (Shaham et al., 2023; Zhang et al., 2024a; Min et al., 2023; Zhang et al., 2024b). Our goal with CUB was to introduce a foundational benchmark that enables rigorous comparisons between CMTs across diverse types of contexts and domains. Starting with standard-length contexts was a natural choice, as they remain both prevalent and crucial. Even within this scope, CUB uncovers valuable insights into CMT behavior. Future work can extend on the foundation provided by CUB to include long-context benchmarks.

The *multi-document* setting – where multiple contexts are provided to the model – is another important area for future research. However, CUB focuses on single-document settings to allow controlled and meaningful evaluation of diverse CMTs across different context types. Evaluating CMTs in a multi-document setup introduces significant complexity, as it becomes challenging to disentangle the influence of individual documents from their combined effects. Furthermore, many of the recent and widely used CMTs included in CUB were designed specifically for single-document scenarios and are not yet compatible with multi-document inputs. As such, CUB provides a necessary and robust foundation, which future work can build upon to address the more intricate challenges of multi-document settings.

CUB is built on a carefully curated selection of nine models, chosen to balance recency, diversity in model size, instruction tuning, and relevance to RAG-specific applications. While a limitation of CUB is its *coverage of only a subset of available models*, the thoughtful selection ensures that the insights it provides are broadly informative and likely to generalise. Additionally, CUB is designed to be extensible, allowing future inclusion of a wider range of models.

While the CMTs included in CUB represent the current state of the art and cover all major categories, the rapidly evolving landscape means that *many promising methods remain to be benchmarked*. However, by open-sourcing CUB, we provide the research community with a robust foundation for systematic evaluation of both existing and future CMTs.

The datasets used in CUB were carefully selected to span a range of task difficulties and reflect common RAG scenarios. However, the *insights derived from* CUB *are inherently constrained by the characteristics of these underlying datasets*. Additionally, since all datasets are in English, it remains an open question whether the findings generalise to other languages (Chirkova et al., 2024). Future work could extend CUB to cover more datasets as well as non-English datasets.

Finally, CUB does not explicitly address datasets involving *temporal dynamics*, which represent an interesting avenue for future research. Time-sensitive information can introduce natural conflicts within the context, offering a richer setting to analyse context utilisation (Loureiro et al., 2022; Xiong et al., 2024). Additionally, while the context types included in CUB were carefully selected based on prior research, it is possible that other context types have been overlooked.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5476, Singapore. Association for Computational Linguistics.

Lovisa Hagström, Sara Vera Marjanović, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2024. A reality check on context utilisation for retrieval-augmented generation. *Preprint*, arXiv:2412.17031.

Lovisa Hagström, Ercong Nie, Ruben Halifa, Helmut Schmid, Richard Johansson, and Alexander Junge. 2025. Language model re-rankers are steered by lexical similarities. *Preprint*, arXiv:2502.17036.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian 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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2421–2431, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

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.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Denitsa Saynova, Lovisa Hagström, Moa Johansson, Richard Johansson, and Marco Kuhlmann. 2025. Fact recall, heuristics or pure guesswork? precise interpretations of language models for fact completion. *Preprint*, arXiv:2410.14405.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Preprint*, arXiv:2305.13117.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Sagi Shaier, Lawrence Hunter, and Katharina von der Wense. 2024. Desiderata for the context use of question answering systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 777–792, St. Julian's, Malta. Association for Computational Linguistics.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. 2024. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8988–9003, Miami, Florida, USA. 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.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. 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.

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.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *Preprint*, arXiv:2409.07394.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *Preprint*, arXiv:2501.17148.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024a. 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. 2024b. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922, Bangkok, Thailand. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024b. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Preprint*, arXiv:2403.04797.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4225–4237, Mexico City, Mexico. Association for Computational Linguistics.

A Additional results

A.1 CUB results

The exact BCU scores on CUB can be found in Table 4. We also report the accuracy scores on CUB in Table 5. For CounterFact and DRUID, accuracy is measured based on whether the first generated token is the same as the first gold token. For NQ, for which the correct answer may be different permutations of the same set of tokens, we measure accuracy based on whether the first output token (e.g. "July") matches any of the tokens in the answer (e.g. "15 July").

CCU scores can be found in Figure 4. For the CCU scores, we note that they generally follow the same trends as the BCU scores in Figure 2; some CMTs perform better on gold, conflicting or irrelevant contexts, while none are superior when all context types are taken into consideration. The only disparate trend at odds with the BCU scores is that `Fine-tuning` Qwen models that have been instruction-tuned stand out by performing extra poorly with respect to CCU score. We hypothesise that this is a consequence of an increase in $P_M(t_C|Q)$ (i.e. prediction probability without context) from the fine-tuning, yielding less room for improvement in prediction confidence when context is introduced.

A.2 Analysis of inflated CMT performance on CounterFact

The inflated performance on CounterFact, observed in Figures 2 and 3, can potentially be explained by a suboptimal default prompt for CounterFact. Following previous work, the default prompt only contained the example to be completed, without any additional instructions or few-shot examples. For NQ and DRUID, the default prompt contained task instructions and few-shot examples. Furthermore, we observe how `Prompting` performs best on CounterFact on average, with a near perfect performance, indicating that a better default prompt may have neutralised any additional improvements from other CMTs. This raises the question of whether certain CMTs only address low context utilisation when caused by poor prompting, finding no leverage if the prompt already is adequate.

A.3 Quality check of irrelevant NQ contexts

For the CUB evaluation, we find 244 (14%) NQ samples with the context type 'irrelevant' for which at least 5 of the 9 evaluated LMs switch prediction to the gold answer *after* having seen the sample context. This indicates that some of the irrelevant contexts may actually be gold, as a result of quality issues with the annotation for NQ (in our sampling we assume that Wikipedia paragraphs not annotated as gold are not gold). However, we also note for some of these 244 samples that the context may simply be the heading of a Wikipedia page with the same title as the gold answer (e.g. "<H1> Scythe </H1>" when the gold answer is "scythe" for the query "what is the name of the weapon the grim reaper carries?"), without providing sufficient evidence with respect to the question, raising the question of whether they should be considered relevant by the model.

A.4 Performance of Relevance Judgement

For the `Multi-agent` technique, we investigate whether instruction-tuned LMs are capable of identifying irrelevant context when explicitly prompted to do so. According to Table 6, the `Multi-agent` approach demonstrates strong performance in detecting irrelevant contexts and in recognising gold contexts as relevant. Although it does not reliably maintain a closed-book response when directly generating responses (i.e. `Regular`), it can accurately detect irrelevance when equipped with an explicit relevance assessment setup.

The prediction accuracy of relevance assessment on conflicting contexts is consistently lower than that on other contexts. This discrepancy is particularly evident in the conflicting contexts of the CounterFact dataset. For instance, we found that LMs often generate feedback such as: "X is Y, not Z. Therefore, the context is irrelevant". This suggests that LM interprets factual inconsistency with its internal knowledge as a signal of irrelevance, even when instructed to ignore its own memory.

One possible explanation for this behaviour lies in the nature of the CounterFact dataset itself. Contexts in CounterFact are typically composed of single-sentence facts, which may lack sufficient surrounding information to render the context trustworthy from the model's perspective. Such behaviour is less pronounced in NQ and DRUID datasets, where the provided contexts are relatively longer and richer, offering more semantic cues that

	Dataset	CounterFact				NQ				DRUID			
Model	Method	Gold	Conflict.	Irrel.	Tot.	Gold	Conflict.	Irrel.	Tot.	Gold	Conflict.	Irrel.	Tot.
GPT-2 XL	Regular	**100.0**	96.4	81.0	92.5	43.0	37.6	13.7	31.4	80.9	7.3	76.5	39.6
	Fine-tuning	**100.0**	92.9	82.4	91.8	46.9	42.3	13.9	34.3	72.4	12.6	47.1	38.7
	Prompting	**100.0**	96.4	81.0	92.5	42.4	36.2	14.2	30.9	83.3	1.9	**100.0**	37.7
	PH3 +context	**100.0**	99.4	44.8	81.4	42.3	36.4	14.0	30.9	79.6	11.6	76.5	41.5
	PH3 +memory	**100.0**	99.5	76.8	92.1	41.4	35.4	13.9	30.2	81.1	3.9	**100.0**	37.9
	COIECD	**100.0**	97.6	70.8	89.5	43.4	37.4	9.0	29.9	69.8	21.3	41.2	42.4
	ACD	99.6	49.1	91.0	79.9	31.8	29.1	36.4	32.4	81.3	3.2	**100.0**	37.6
Pythia 6.9B	Regular	**100.0**	56.5	91.5	82.7	52.7	43.9	16.2	37.6	84.1	9.4	52.9	42.1
	Fine-tuning	**100.0**	65.1	89.4	84.8	54.0	49.6	14.6	39.4	81.5	1.4	94.1	36.6
	Prompting	**100.0**	99.6	86.1	**95.2**	52.7	43.9	16.2	37.6	82.8	7.1	64.7	40.3
	PH3 +context	98.3	89.7	62.4	83.5	55.9	46.3	14.6	38.9	87.1	8.7	58.8	43.0
	PH3 +memory	91.4	4.0	90.5	61.9	48.9	39.2	18.1	35.4	86.2	8.4	70.6	42.5
	COIECD	99.9	66.0	86.0	84.0	53.9	43.8	10.2	35.9	72.0	13.0	41.2	38.8
	ACD	**100.0**	9.7	96.0	68.6	43.8	36.1	32.6	37.5	87.4	5.2	**100.0**	41.3
Qwen2.5 1.5B	Regular	99.9	53.1	80.0	77.6	44.0	41.1	22.4	35.8	84.7	11.6	70.6	43.6
	Fine-tuning	**100.0**	90.3	85.7	92.0	66.1	61.9	18.5	48.8	79.7	18.5	52.9	45.3
	Prompting	**100.0**	97.2	82.2	93.2	63.9	57.5	32.1	51.1	85.0	7.0	82.4	41.2
	PH3 +context	**100.0**	99.0	62.5	87.2	44.2	40.9	21.7	35.6	63.8	40.4	17.6	50.5
	PH3 +memory	98.9	38.5	84.9	74.1	19.4	17.3	26.0	20.9	81.2	1.4	**100.0**	36.5
	COIECD	94.8	1.2	89.8	61.9	42.4	39.2	45.8	42.5	87.8	4.8	**100.0**	41.3
	ACD	97.6	7.7	90.3	65.2	46.7	42.8	39.3	42.9	87.8	4.8	**100.0**	41.3
Qwen2.5 1.5B *Instruct*	Regular	97.6	31.7	86.2	71.8	70.1	62.8	28.2	53.7	47.3	70.3	94.1	60.4
	Fine-tuning	**100.0**	93.2	82.7	92.0	51.0	45.6	42.2	46.3	72.0	14.5	29.4	39.6
	Prompting	99.3	94.2	76.1	89.9	68.1	60.5	29.1	52.5	47.3	70.3	94.1	60.4
	Multi-agent	98.6	24.7	99.9	74.4	68.5	60.2	45.0	57.9	44.4	72.4	94.1	60.3
	PH3 +context	96.0	42.5	59.8	66.1	67.1	59.9	26.0	51.0	61.1	64.7	94.1	63.2
	PH3 +memory	94.6	11.5	85.5	63.9	48.8	42.7	22.0	37.8	25.4	76.1	94.1	54.1
	COIECD	97.8	35.8	82.7	72.1	70.5	63.9	22.1	52.1	64.1	59.6	94.1	61.7
	ACD	95.6	12.1	93.5	67.1	66.7	60.0	43.4	56.7	12.3	79.9	94.1	50.6
Qwen2.5 7B	Regular	96.6	36.0	79.0	70.5	71.7	65.6	25.3	54.2	91.8	23.6	41.2	53.3
	Fine-tuning	99.6	47.4	85.0	77.4	76.7	68.8	41.7	62.4	86.4	1.8	82.4	39.0
	Prompting	**100.0**	97.8	81.3	93.0	74.7	66.5	31.2	57.5	94.9	13.8	58.8	49.3
	PH3 +context	97.8	96.3	16.7	70.3	69.7	63.6	25.3	52.8	83.4	50.1	17.6	64.5
	PH3 +memory	96.8	4.0	84.2	61.6	66.5	59.5	26.6	50.8	90.5	4.1	76.5	42.0
	COIECD	96.6	36.0	79.0	70.5	71.7	65.6	25.3	54.2	91.8	23.6	41.2	53.3
	ACD	94.7	2.3	92.7	63.2	72.3	59.9	41.9	58.0	89.8	12.6	70.6	46.4
Qwen2.5 7B *Instruct*	Regular	**100.0**	25.9	84.5	70.1	76.2	65.0	31.0	57.4	87.8	57.1	64.7	70.5
	Fine-tuning	**100.0**	62.3	81.0	81.1	59.6	52.7	48.1	53.5	96.4	13.2	70.6	49.6
	Prompting	**100.0**	98.6	35.3	78.0	75.8	66.7	29.1	57.2	87.8	57.1	64.7	70.5
	Multi-agent	95.7	11.6	**100.0**	69.1	66.1	52.2	**73.3**	63.9	58.6	63.2	94.1	61.3
	PH3 +context	98.3	84.0	54.1	78.8	75.3	64.4	26.9	55.5	86.9	54.7	70.6	68.8
	PH3 +memory	**100.0**	27.6	82.8	70.1	76.4	66.1	30.9	57.8	3.1	**81.4**	70.6	47.3
	COIECD	99.9	9.1	90.6	66.5	76.2	60.1	40.8	59.0	76.4	56.5	76.5	65.2
	ACD	99.6	11.5	96.9	69.3	76.3	62.1	44.6	61.0	76.2	57.6	76.5	65.8
Qwen2.5 32B	Regular	99.9	77.6	77.2	84.9	77.3	66.7	39.7	61.2	**98.2**	19.8	41.2	54.0
	Fine-tuning	98.1	88.4	81.9	89.4	79.2	**69.2**	46.3	**64.9**	98.0	9.7	82.4	48.4
	Prompting	**100.0**	**100.0**	80.7	93.6	77.2	66.9	42.8	62.3	**98.2**	22.5	52.9	55.6
	COIECD	97.4	96.5	58.5	84.1	76.1	67.4	32.7	58.7	97.1	27.8	29.4	57.9
	ACD	97.6	2.3	92.6	64.1	75.7	56.1	57.6	63.1	97.6	14.1	58.8	50.6
Qwen2.5 32B *Instruct*	Regular	99.4	4.9	92.6	65.6	**81.4**	59.9	43.8	61.7	97.9	43.2	76.5	67.2
	Fine-tuning	**100.0**	18.0	93.6	70.5	71.6	64.9	42.0	59.5	96.4	20.8	52.9	53.8
	Prompting	99.9	95.3	69.1	88.1	**81.4**	59.9	43.8	61.7	97.2	48.7	82.4	70.0
	Multi-agent	**100.0**	20.6	**100.0**	73.5	76.8	57.2	49.2	61.1	93.1	55.6	94.1	72.1
	COIECD	98.0	6.0	70.8	58.3	79.7	61.6	36.8	59.4	97.7	38.3	64.7	64.3
	ACD	98.4	2.5	97.5	66.1	80.1	55.2	57.4	64.2	88.5	51.4	94.1	67.7
Command A	Regular	**100.0**	**100.0**	4.1	68.0	79.2	62.7	28.9	56.9	95.9	57.3	76.5	74.2
	Prompting	97.0	92.8	48.4	79.4	79.2	62.7	28.9	56.9	93.6	64.4	70.6	**77.2**
	Multi-agent	99.6	39.1	99.9	79.6	74.3	49.7	58.8	61.0	91.9	48.2	94.1	67.4

Table 4: BCU scores on CUB. A high BCU score is desirable regardless of context type. Gold denotes relevant contexts that also contain the gold answer. Conflict. denotes 'Conflicting' – relevant contexts that contain a conflicting answer, dissimilar from the correct answer or model memory. Irrel. denotes irrelevant contexts. Tot. denotes the average performance across all context types. Values marked in **bold** indicate the top CMT score across LMs for each dataset and context type.

	Dataset	CounterFact				NQ				DRUID			
Model	Method	Gold	Conflict.	Irrel.	Tot.	Gold	Conflict.	Irrel.	Tot.	Gold	Conflict.	Irrel.	Tot.
GPT-2 XL	Regular	**100.0**	2.9	69.7	57.5	43.0	8.1	20.8	24.0	80.9	69.0	64.7	74.2
	Fine-tuning	**100.0**	3.2	70.6	57.9	46.9	7.7	23.8	26.2	72.4	65.5	41.2	68.4
	Prompting	**100.0**	2.9	69.7	57.5	42.4	7.5	20.3	23.5	83.3	73.8	76.5	78.0
	PH3 +context	**100.0**	0.4	29.8	43.4	42.3	7.8	20.4	23.6	79.6	65.7	52.9	71.7
	PH3 +memory	**100.0**	0.4	65.1	55.1	41.4	7.4	20.1	23.0	81.1	72.6	76.5	76.3
	COIECD	**100.0**	2.3	67.7	56.7	43.4	7.1	19.4	23.3	69.8	51.0	47.1	59.1
	ACD	99.6	29.4	72.3	67.1	31.8	7.7	18.1	19.2	81.3	73.0	76.5	76.6
Pythia 6.9B	Regular	**100.0**	37.2	91.4	76.2	52.7	9.8	29.6	30.8	84.1	49.9	47.1	64.7
	Fine-tuning	**100.0**	26.5	91.8	72.8	54.0	5.6	26.6	28.8	81.5	74.4	70.6	77.5
	Prompting	**100.0**	0.5	86.1	62.2	52.7	9.8	29.6	30.8	82.8	57.1	47.1	68.3
	PH3 +context	98.3	2.5	62.1	54.3	55.9	8.4	30.0	31.5	87.1	55.2	52.9	69.0
	PH3 +memory	91.4	86.0	90.4	89.2	48.9	11.5	29.7	30.1	86.2	55.1	64.7	68.7
	COIECD	99.9	27.3	86.0	71.0	53.9	9.8	27.4	30.4	72.0	32.9	35.3	50.0
	ACD	**100.0**	77.6	**95.9**	91.2	43.8	12.1	29.7	28.6	87.4	69.2	**82.4**	77.2
Qwen2.5 1.5B	Regular	99.9	41.9	74.2	72.0	44.0	7.7	22.0	24.6	84.7	63.5	52.9	72.7
	Fine-tuning	**100.0**	5.5	77.0	60.8	66.1	18.8	42.4	42.5	79.7	60.3	58.8	68.7
	Prompting	**100.0**	1.6	79.7	60.4	63.9	17.0	38.5	39.8	85.0	69.8	58.8	76.4
	PH3 +context	**100.0**	0.7	50.1	50.3	44.2	12.6	25.5	27.5	63.8	26.9	11.8	42.9
	PH3 +memory	98.9	52.8	78.0	76.6	19.4	8.1	10.4	12.7	81.2	74.5	70.6	77.4
	COIECD	94.8	71.9	79.0	81.9	42.4	16.3	27.6	28.8	87.8	72.7	70.6	79.3
	ACD	97.6	70.8	79.4	82.6	46.7	15.5	28.0	30.1	87.8	72.7	70.6	79.3
Qwen2.5 1.5B *Instruct*	Regular	97.6	54.5	79.6	77.2	70.1	16.1	37.1	41.2	47.3	11.1	0.0	26.8
	Fine-tuning	**100.0**	7.0	78.0	61.7	51.0	7.6	27.8	28.8	72.0	28.5	47.1	47.5
	Prompting	99.3	5.4	74.1	59.6	68.1	15.7	38.8	41.0	47.3	11.1	0.0	26.8
	Multi-agent	98.6	68.7	83.0	83.4	68.5	16.9	36.1	40.6	44.4	10.0	0.0	24.9
	PH3 +context	96.0	35.9	58.2	63.4	67.1	15.4	34.7	39.1	61.1	18.9	0.0	37.2
	PH3 +memory	94.6	68.9	78.3	80.6	48.8	13.1	25.8	29.3	25.4	7.2	0.0	15.1
	COIECD	97.8	50.4	77.1	75.1	70.5	15.5	35.9	40.7	64.1	19.2	0.0	38.7
	ACD	95.6	77.7	82.1	85.1	66.7	19.0	39.0	41.6	12.3	3.6	0.0	7.4
Qwen2.5 7B	Regular	96.6	52.2	72.6	73.8	71.7	16.7	39.0	42.6	91.8	57.6	23.5	72.3
	Fine-tuning	99.6	45.1	77.1	73.9	76.7	18.5	50.5	48.6	86.4	**74.8**	70.6	79.8
	Prompting	**100.0**	2.4	86.2	62.9	74.7	17.9	44.6	45.8	94.9	64.2	35.3	77.4
	PH3 +context	97.8	0.2	6.0	34.7	69.7	17.0	38.7	41.9	83.4	30.5	5.9	53.4
	PH3 +memory	96.8	88.6	79.4	88.2	66.5	17.6	37.7	40.6	90.5	73.4	70.6	**80.8**
	COIECD	96.6	52.2	72.6	73.8	71.7	16.7	39.0	42.6	91.8	57.6	23.5	72.3
	ACD	94.7	85.5	80.4	86.9	72.3	23.9	47.2	47.8	89.8	68.1	47.1	77.5
Qwen2.5 7B *Instruct*	Regular	**100.0**	42.0	85.4	75.8	76.2	19.8	47.1	47.8	87.8	28.3	0.0	54.1
	Fine-tuning	**100.0**	34.8	88.0	74.3	59.6	8.1	35.3	34.4	96.4	65.0	64.7	78.6
	Prompting	**100.0**	1.9	37.5	46.5	75.8	20.3	46.0	47.4	87.8	28.3	0.0	54.1
	Multi-agent	95.7	85.5	94.0	91.7	66.1	21.4	40.9	42.9	58.6	18.5	29.4	36.0
	PH3 +context	98.3	12.5	55.6	55.5	75.3	18.5	44.1	46.0	86.9	31.5	0.0	55.5
	PH3 +memory	**100.0**	50.9	83.8	78.2	76.4	20.1	47.7	48.1	3.1	2.5	0.0	2.7
	COIECD	99.9	75.0	90.8	88.6	76.2	25.8	48.2	50.1	76.4	29.2	5.9	49.7
	ACD	99.6	85.1	94.0	92.9	76.3	25.0	49.3	50.3	76.2	29.1	5.9	49.5
Qwen2.5 32B	Regular	99.9	21.4	75.0	65.4	77.3	20.8	47.7	48.7	**98.2**	58.5	29.4	75.7
	Fine-tuning	98.1	9.8	77.2	61.7	79.2	20.3	**55.9**	51.9	98.0	66.6	64.7	80.3
	Prompting	**100.0**	0.2	80.7	60.3	77.2	19.9	50.2	49.2	**98.2**	57.5	41.2	75.2
	COIECD	97.4	3.2	59.7	53.4	76.1	18.8	43.9	46.3	97.1	47.4	17.6	68.9
	ACD	97.6	85.7	81.3	88.2	75.7	31.4	53.3	53.5	97.6	66.1	47.1	79.8
Qwen2.5 32B *Instruct*	Regular	99.4	81.0	93.5	91.3	**81.4**	28.6	52.2	54.2	97.9	41.8	29.4	66.2
	Fine-tuning	**100.0**	78.5	92.2	90.2	71.6	13.3	44.3	43.2	96.4	61.8	47.1	76.8
	Prompting	99.9	3.2	70.6	57.9	**81.4**	28.6	52.2	54.2	97.2	36.2	11.8	62.6
	Multi-agent	**100.0**	78.5	94.7	91.1	76.8	22.7	40.7	46.8	93.1	31.7	17.6	58.4
	COIECD	98.0	9.7	72.4	60.0	79.7	23.4	49.4	50.9	97.7	43.3	29.4	66.9
	ACD	98.4	**94.7**	95.4	**96.2**	80.1	**35.3**	55.4	**57.0**	88.5	36.0	17.6	58.8
Command A	Regular	**100.0**	0.0	4.4	34.8	79.2	12.3	33.8	41.9	95.9	30.3	5.9	58.8
	Prompting	97.0	0.7	47.8	48.5	79.2	12.3	33.8	41.9	93.6	23.3	0.0	53.8
	Multi-agent	99.6	32.2	90.2	74.0	74.3	13.5	40.4	42.8	91.9	33.2	23.5	58.7

Table 5: Accuracy with respect to gold label on CUB. Gold denotes relevant contexts that also contain the gold answer. Conflict. denotes ‘Conflicting’ – relevant contexts that contain a conflicting answer, dissimilar from the correct answer or model memory. Irrel. denotes irrelevant contexts. Tot. denotes the average performance across all context types. Values marked in **bold** indicate the top CMT score across LMs on each dataset and context type.

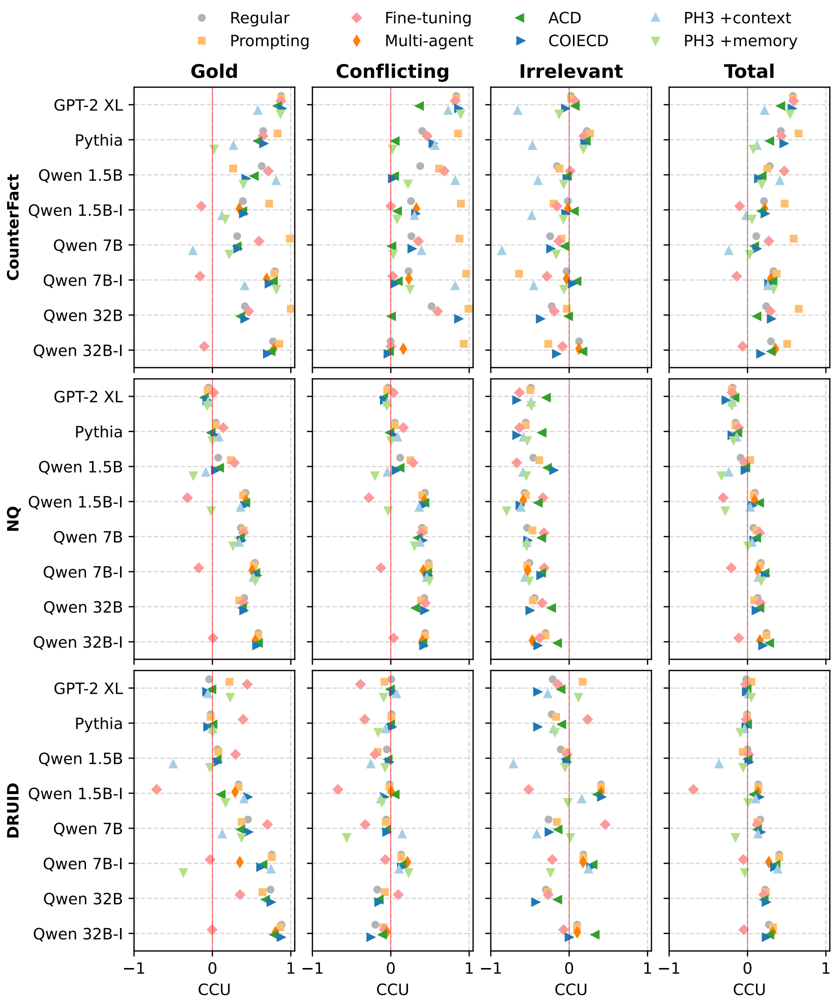

Figure 4: CCU scores for the evaluated context utilisation manipulation methods applied to the evaluated models and datasets. ‘Total’ denotes the averaged performance across all context types. A high CCU score is desirable regardless of context type. The red vertical lines indicate scores of 0.

may help the LM interpret the information as contextually anchored (Xie et al., 2024b).

The performance of relevance assessment is particularly low on the NQ dataset compared to other datasets. Since irrelevant contexts of NQ dataset are sampled from the same document and may be topically or semantically similar to the question, distinguishing relevance may become more challenging.

A.5 Features Impacting Context Utilisation

See Table 3 for the correlation values between model features and context utilisation. See Table 7 for the correlation values between input features and context utilisation.

B Data Collection

We collect three datasets for CUB: CounterFact, NQ, and DRUID. For each dataset, we carefully construct validation and test splits to enable fair and unified hyperparameter tuning for each CMT, as well as unbiased final evaluation metrics. The following subsections describe the construction and characteristics of each dataset in detail.

B.1 CounterFact

Samples from the CounterFact dataset can be found in Table 8. The relations covered by the dataset are *capital of* (80%), *country of origin* (9%), *location of formation* (9%), *field of work* (1%) and *country of citizenship* (1%).

Rate of memorisation of CUB models We evaluate all `Regular` LMs on the samples from CUB CounterFact without context. The results can be found in Table 9. We observe rates above 70% for all models. As expected, the highest memorisation rate is found for Pythia. The lowest is found for

	Gold	Conflict.	Irrel.	All
		QWEN2.5 1.5B-I		
CounterFact	98.56	24.25	99.88	74.23
NQ	92.44	91.89	26.26	70.13
DRUID	93.27	96.52	17.65	94.79
		QWEN2.5 7B-I		
CounterFact	99.16	10.68	99.88	69.91
NQ	80.70	76.14	59.35	72.05
DRUID	82.53	65.56	94.12	73.06
		QWEN2.5 32B-I		
CounterFact	99.64	19.57	99.40	72.87
NQ	94.74	92.50	25.77	70.94
DRUID	98.66	76.25	88.24	86.05
		COMMAND A		
CounterFact	100.00	99.88	99.88	99.92
NQ	94.31	91.82	37.69	74.56
DRUID	93.11	68.55	88.24	79.31

Table 6: Multi-agent: Relevance assessment accuracy

GPT-2 XL, which can be expected as the model is quite small and old.

Prompt templates Following the same approach as previous work, no specific prompt template was used for the LMs evaluated on CounterFact. The LMs were evaluated in a simple sentence completion format as shown in Table 8.

However, since the sentence completion format is less compatible with the instruction-tuned models, we added a small prompt template for the evaluation of the instruction-tuned Qwen models on CounterFact, as follows.

Prompt without context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are
    a helpful assistant.<|im_end|>
<|im_start|>user
Complete the following sentence. Only answer
    with the next word.
<prompt><|im_end|>
<|im_start|>assistant
```

Prompt with context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are
    a helpful assistant.<|im_end|>
<|im_start|>user
Complete the following sentence. Only answer
    with the next word.
Fact: <context>
<prompt><|im_end|>
<|im_start|>assistant
```

Dataset	Context	CMT	Corr.
	Context length		
CounterFact	Irrelevant	Regular	0.06
CounterFact	Conflicting	Regular	0.04
CounterFact	Gold	Regular	0.02
DRUID	Conflicting	Regular	-0.02
DRUID	Irrelevant	Regular	-0.02
NQ	Irrelevant	Regular	-0.06
DRUID	Gold	Regular	-0.08
NQ	Conflicting	Regular	-0.22
NQ	Gold	Regular	-0.23
DRUID	Irrelevant	Multi-agent	**-0.32**
	Query-context overlap		
DRUID	Gold	Regular	0.02
DRUID	Irrelevant	Regular	-0.03
NQ	Gold	Regular	-0.06
NQ	Conflicting	Regular	-0.08
NQ	Irrelevant	Regular	-0.08
DRUID	Conflicting	Regular	-0.13
DRUID	Irrelevant	Multi-agent	**-0.30**
	Distractor rate		
CounterFact	Gold	Regular	0.00
NQ	Conflicting	Regular	-0.19
NQ	Gold	Regular	-0.19
CounterFact	Conflicting	Regular	-0.22
CounterFact	Conflicting	ACD	**-0.34**
CounterFact	Conflicting	Multi-agent	**-0.49**
	Relevance judgement		
CounterFact	Conflicting	Multi-agent	**0.53**
CounterFact	Conflicting	Regular	0.17
NQ	Irrelevant	Regular	0.11
DRUID	Irrelevant	Regular	0.05
NQ	Gold	Regular	0.04
DRUID	Gold	Regular	0.03
NQ	Conflicting	Regular	0.02
CounterFact	Irrelevant	Regular	0.01
CounterFact	Gold	Regular	-0.01
DRUID	Conflicting	Regular	-0.15
NQ	Irrelevant	Multi-agent	**-0.36**
DRUID	Irrelevant	Multi-agent	**-0.49**

Table 7: Spearman's ρ between BCU and different input aspects. Correlation vallues for `Regular` or with an absolute value above 0.3 are shown. Correlation values with an absolute value below 0.3 are marked in gray. Significant correlation values (p-value < 0.05) are marked in **bold**. Results are measured across models.

Prompt	Type
Fact: Athens, the capital city of Greece. Athens, the capital city of	Gold
Fact: Thomas Ong is a citizen of Pakistan. Thomas Ong is a citizen of	Conflicting
Fact: Melbourne, that is the capital of Jordan. Prince Oscar Bernadotte is a citizen of	Irrelevant

Table 8: CounterFact prompts with contexts and corresponding context types. For prompts without context, the first line (starting with "Fact:") is simply removed.

B.2 NQ

We retain all samples from the development set of NQ[4] for which a short answer of fewer than five tokens is identified in the raw HTML of the

[4] https://console.cloud.google.com/storage/browser/natural_questions/v1.0/dev

Model	Accuracy
GPT-2 XL	71.8
Pythia	99.6
Qwen 1.5B	77.0
Qwen 1.5B-I	83.1
Qwen 7B	79.7
Qwen 7B-I	93.6
Qwen 32B	78.0
Qwen 32B-I	94.5
Command A	90.6

Table 9: Accuracy, proxying memorisation rate, on samples from CounterFact without context.

corresponding Wikipedia pages. Samples from the NQ dataset can be found in Table 10.

Sampling of conflicting contexts We create conflicting contexts that promote a different answer simply by taking the gold context and substituting the gold answer in the context. The substitute answer is sampled to yield coherent conflicting contexts, and to have a different meaning compared to the gold answer.

For a given question, context and short answer, we perform the following steps to identify substitute answers for conflicting contexts:

1. Check if the short answer is a date[5]. If so, sample a new random date in the interval [1900, 2030) and format it in the same way as the gold date.
2. If the short answer is not a date, prompt an LLM[6] with the question and short answer to provide a substitute answer of the same format. If the proposed answer is already found in the sample context, prompt the model, for a maximum of 20 times, to generate another answer until a substitute answer not already found in the context has been generated.

The prompt used to query an LLM for a substitute answer was as follows:

Prompt for getting substitute answers.

```
## Instructions
Please provide an incorrect answer to the
    example below.
The incorrect answer should be incorrect in the
    sense that it should be significantly
    different from the original answer. At the
    same time, it should be a plausible answer
    to the given question.
The incorrect answer should follow the same
    formatting as the original answer such that
    it should be possible to directly replace
    the original answer with the incorrect
    answer in any context.
The incorrect answer should be a single word or
    a short phrase.
Only output the incorrect answer.

## Example
Question: <question>
Original answer:<target_true>
Incorrect answer:
```

[5] Using the `dateutil.parser` in Python.

[6] The Cohere model `command-r-plus-08-2024` from https://docs.cohere.com/v2/docs/command-r-plus.

In the event that the model generated a substitute answer that already could be found in the context, the previous model answer was added to the chat history together with the following new user query:

Prompt for getting another substitute answer.

```
Please provide another incorrect answer
    following the same format as the original
    answer. Only output the incorrect answer.
```

Quality of conflicting contexts A manual inspection of 200 samples found the method reliable for producing adequate conflicting contexts with an accuracy of 90% (11 samples corresponded to poor formatting, 4 were too similar to gold, and 4 were dropped due to data formatting issues or the LLM being unable to generate a substitute answer not already found in the context). In addition, we inspect the CUB results to ascertain the quality of the conflicting context sampling, see Appendix A.

We also experimented with a method based on named entities and random sampling for producing substitute answers for the conflicting contexts. In the method, the entity type of the answer to be replaced was detected and another named entity of the same type was randomly sampled from a NE dataset as the replacement. We found this method to work poorly compared to the LLM based approach. Mainly because the detected NEs lacked sufficient information for a successful sampling of replacements (e.g. "2024" and "last year" may both be labelled as time entities, while they are not interchangeable in all contexts).

Sampling of irrelevant contexts Given a query and a corresponding Wikipedia page, the NQ annotators were instructed to mark the first paragraph in the Wikipedia page that contains an answer to the query. Therefore, to ensure that we only sample irrelevant contexts, we perform the sampling over all paragraphs before the gold paragraph in the given Wikipedia page.

We use the Jina Reranker v2[7] to identify the most relevant non-gold paragraph. It is a modern

[7] `jinaai/jina-reranker-v2-base-multilingual`

Question	Short answer	Context	Type
when did the movie napoleon dynamite come out?	June 11, 2004	<Table> <Tr> <Th colspan="2"> Napoleon Dynamite </Th> </Tr> <Tr> <Td colspan="2"> Theatrical release poster </Td> </Tr> <Tr> <Th> Directed by </Th> <Td> Jared Hess </Td> </Tr> <Tr> <Th> Produced by </Th> <Td> <Ul> <Li> Jeremy Coon </Li> <Li> Chris Wyatt </Li> <Li> Sean Covel </Li> <Li> Jory Weitz </Li> </Ul> </Td> </Tr> <Tr> <Th> Screenplay by </Th> <Td> <Ul> <Li> Jared Hess </Li> <Li> Jerusha Hess </Li> </Ul> </Td> </Tr> <Tr> <Th> Based on </Th> <Td> Peluca by Jared Hess </Td> </Tr> <Tr> <Th> Starring </Th> <Td> <Ul> <Li> Jon Heder </Li> <Li> Jon Gries </Li> <Li> Efren Ramirez </Li> <Li> Tina Majorino </Li> <Li> Aaron Ruell </Li> <Li> Diedrich Bader </Li> <Li> Haylie Duff </Li> </Ul> </Td> </Tr> <Tr> <Th> Music by </Th> <Td> John Swihart </Td> </Tr> <Tr> <Th> Cinematography </Th> <Td> Munn Powell </Td> </Tr> <Tr> <Th> Edited by </Th> <Td> Jeremy Coon </Td> </Tr> <Tr> <Th> Production company </Th> <Td> <Ul> <Li> MTV Films </Li> <Li> Napoleon Pictures </Li> <Li> Access Films </Li> </Ul> </Td> </Tr> <Tr> <Th> Distributed by </Th> <Td> <Ul> <Li> Fox Searchlight Pictures (North America) </Li> <Li> Paramount Pictures (International) </Li> </Ul> </Td> </Tr> <Tr> <Th> Release date </Th> <Td> <Ul> <Li> January 17, 2004 (2004 - 01 - 17) (Sundance) </Li> <Li> **June 11, 2004** (2004 - 06 - 11) (United States) </Li> <Li> </Li> <Li> </Li> <Li> </Li> </Ul> </Td> </Tr> <Tr> <Th> Running time </Th> <Td> 95 minutes </Td> </Tr> <Tr> <Th> Country </Th> <Td> United States </Td> </Tr> <Tr> <Th> Language </Th> <Td> English </Td> </Tr> <Tr> <Th> Budget </Th> <Td> $400,000 </Td> </Tr> <Tr> <Th> Box office </Th> <Td> $46.1 million </Td> </Tr> </Table>	Gold
when was the lupus foundation of america founded?	1977	<P> The Lupus Foundation of America (LFA), founded in **1967**, is a national voluntary health organization based in Washington, D.C. with a network of chapters, offices and support groups located in communities throughout the United States . The Foundation is devoted to solving the mystery of lupus, one of the world's cruelest, most unpredictable and devastating diseases, while giving caring support to those who suffer from its brutal impact . Its mission is to improve the quality of life for all people affected by lupus through programs of research, education, support and advocacy . </P>	Conflicting
who has scored the most tries in rugby union?	Daisuke Ohata	<P> This is a list of the leading try scorers in rugby union test matches . It includes players with a minimum of 30 test tries . </P>	Irrelevant

Table 10: NQ samples and corresponding context types.

LM re-ranker that has been proven to work well on NQ (Hagström et al., 2025).

Prompt templates The 2-shot prompts used to evaluate the LMs on NQ were as follows.

Prompt without context.

```
Answer the following questions.
Question: When is the first episode of House of the Dragon released?
Answer: August 21, 2022

Question: In what country will the 2026 Winter Olympics be held?
Answer: Italy

Question: <question>
Answer:
```

Prompt with context.

```
Answer the following questions based on the context below.
Question: When is the first episode of House of the Dragon released?
Context: <Table> <Tr> <Th> Season </Th> <Th> Episodes </Th> <Th> First released </Th> <Th> Last released </Th> </Tr> <Tr> <Td> 1 </Td> <Td> 10 </Td> <Td> August 21, 2022 </Td> <Td> October 23, 2022 </Td> </Tr> <Tr> <Td> 2 </Td> <Td> 8 </Td> <Td> June 16, 2024 </Td> <Td> August 4, 2024
</Td> </Tr> </Table
Answer: August 21, 2022

Question: Where will the 2026 Winter Olympics be held?
Context: <P> The 2026 Winter Olympics (Italian: Olimpiadi invernali del 2026), officially the XXV Olympic Winter Games and commonly known as Milano Cortina 2026, is an upcoming international multi-sport event scheduled to take place from 6 to 22 February 2026 at sites across Lombardy and Northeast Italy.
</P>
Answer: Lombardy and Northeast Italy

Question: <question>
Context: <context>
Answer:
```

For the instruction-tuned Qwen models, a chat template with slightly different prompt templates was used. The 2-shot prompt templates for the instruction-tuned models were as follows.

Prompt without context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.<|im_end|>
<|im_start|>user
Answer the question. Only answer with the answer. Examples of questions and desired answers are given below.

# Example 1
Question: When is the first episode of House of the Dragon released?
Answer: August 21, 2022

# Example 2
Question: In what country will the 2026 Winter Olympics be held?
Answer: Italy

# Now, answer the following question (only with the answer):
Question:<question>
```

```
Answer:<|im_end|>
<|im_start|>assistant
```

Prompt with context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are
    a helpful assistant.<|im_end|>
<|im_start|>user
Answer the question based on the provided
    context. Only answer with the answer.
    Examples of questions and desired answers
    are given below.

# Example 1
Question: When is the first episode of House of
    the Dragon released?
Context: <Table> <Tr> <Th> Season </Th> <Th>
    Episodes </Th> <Th> First released </Th> <Th
    > Last released </Th> </Tr> <Tr> <Td> 1 </Td
    > <Td> 10 </Td> <Td> August 21, 2022 </Td> <
    Td> October 23, 2022 </Td> </Tr> <Tr> <Td> 2
     </Td> <Td> 8 </Td> <Td> June 16, 2024 </Td>
     <Td> August 4, 2024
</Td> </Tr> </Table
Answer: August 21, 2022

# Example 2
Question: Where will the 2026 Winter Olympics be
     held?
Context: <P> The 2026 Winter Olympics (Italian:
    Olimpiadi invernali del 2026), officially
    the XXV Olympic Winter Games and commonly
    known as Milano Cortina 2026, is an upcoming
     international multi-sport event scheduled
    to take place from 6 to 22 February 2026 at
    sites across Lombardy and Northeast Italy.
    </P>
Answer: Lombardy and Northeast Italy

# Now, answer the following question (only with
    the answer):
Question: <question>
Context: <context>
Answer:<|im_end|>
<|im_start|>assistant
```

B.3 DRUID

We map the stances of DRUID to context type using the following approach:

1. Gold: If the evidence is relevant and the stance of the evidence aligns with the claim verdict reached by the fact-check site (here considered gold). This automatically encompasses most samples with evidence that has been sampled from a fact-check site, as the stance of the evidence is likely to align with the FC verdict.
2. Conflicting: If the evidence is relevant and the stance of the evidence does not align with the claim verdict. This automatically encompasses all samples with insufficient evidence, as the original FC verdicts always are True, Half True or False.
3. Irrelevant: If the evidence is irrelevant.

Samples from the DRUID dataset can be found in Table 11. The evidence stance and fact-check verdict distributions per context type can be found in Tables 12 and 13.

Differently from CounterFact and NQ, no context synthesis is necessary for the DRUID samples as they, by virtue of utilising naturally occurring samples from a RAG pipeline, already contain samples representative of gold, conflicting and irrelevant contexts.

Prompt templates The 2-shot prompts used for evaluating the LMs on DRUID were as follows.

Prompt without context.

```
Are the following claims True or False? Answer
    None if you are not sure or cannot answer.

Claimant: Viral post
Claim: "the new coronavirus has HIV proteins
    that indicate it was genetically modified in
     a laboratory."
Answer: False

Claimant: Sara Daniels
Claim: "Blackpink released the single 'You me
    too' in 2026."
Answer: None

Claimant: <claimant>
Claim: "<claim>"
Answer:
```

Prompt with context.

```
Are the claims True or False based on the
    accompanying evidence? If you are not sure
    or cannot answer, say None.

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

Claimant: Sara Daniels
Claim: "Blackpink released the single 'You me
    too' in 2026."
Evidence: "Blackpink released their album 'Born
    Pink' in 2022."
Answer: None

Claimant: <claimant>
Claim: "<claim>"
Evidence: "<evidence>"
Answer:
```

For the instruction-tuned Qwen models, a chat template with slightly different prompt templates was

Claimant	Claim	Verdict	Evidence	Type
Viral Claim	Harvard professor Charles Lieber was arrested for manufacturing and selling the new coronavirus to China	False	Lieber was arrested on January 28 for "making false statements to the agency of the United States Government," or lying to federal authorities about his ties to China, as per the fact-check report. The channel added that prosecutors have never alleged that Lieber was involved in manufacturing and/or selling a virus to China. The full federal court complaint against Dr Lieber can be read here.</p>.<p>The report also clarified Lieber's links to Wuhan. The report stated, "Lieber travelled to WUT (Wuhan University of Technology) in mid-November 2011 ostensibly in order to participate in a Nano-Energy Materials Forum."</p>.<p>On July 29, Dr Lieber's attorney Marc Mukasey told WCVB Channel 5 that he didn't hide anything or get paid as the government alleges.</p>.<p>Thus, the social media claim that Harvard professor Dr Charles Lieber "made and sold" the Covid-19 virus to China is false.</p>	Gold
FACEBOOK POST	WikiLeaks has published the 1st list of black money holders in Swiss banks.	False	(See attached file: List of Black Money Holders from Wiki	Conflict.
Irish Congress of Trade Unions (ICTU)	One in five school staff in Northern Ireland are assaulted at least once a week.	False	Finnegan, who died in January 2002, had also abused boys at St. Colman's College, a prestigious Catholic boys' secondary school in Newry, Northern Ireland. He taught there from 1967 to 1971 and again from 1973 to 1976, when he was appointed president of the school. He served in that post until 1987. [...] Admitted on October 9, 2014 to sample charges of indecently assaulting four boys as young as 10 at St Mary's CBS primary school in Mullingar between 1984 and 1987. Jailed for two years at Mullingar Circuit Court sitting in Tullamore. This concluded a ten-year investigation by detectives in Mullingar. [...] When Smyth returned to Kilnacrott in 1983, he again began abusing children in Belfast, including the girl who, on February 23, 1990, would meet with a social worker at the Catholic Family Welfare Society in Belfast and start all the Smyth revelations.	Irrel.

Table 11: DRUID samples and corresponding context types. 'Conflict.' and 'Irrel.' denote conflicting and irrelevant context types, respectively.

Context	Evidence stance	Count
Gold	Refutes	1,579
	Supports	359
Conflicting	Refutes	35
	Insufficient-refutes	437
	Insufficient-contradictory	163
	Insufficient-neutral	892
	Insufficient-supports	585
	Supports	367
Irrelevant	`not applicable`	83

Table 12: Stance distribution per context type for DRUID.

Context	FC verdict	Count
Gold	False	1,579
	True	359
Conflicting	False	1,842
	Half True	276
	True	361
Irrelevant	False	54
	Half True	13
	True	16

Table 13: Fact-check verdict distribution per context type for DRUID.

used for compatibility. The 2-shot prompt templates for the instruction-tuned models were as follows.

Prompt without context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are
    a helpful assistant.<|im_end|>
<|im_start|>user
Is the claim True or False? Answer None if you
    are not sure or cannot answer. Only answer
    with True, False or None. Examples of claims
     and desired answers are given below.

# Example 1
Claimant: Viral post
Claim: "the new coronavirus has HIV proteins
    that indicate it was genetically modified in
     a laboratory."
Answer: False

# Example 2
Claimant: Sara Daniels
Claim: "Blackpink released the single 'You me
    too' in 2026."
Answer: None

# Now, answer for the following claim:
Claimant: <claimant>
Claim: "<claim>"
Answer (True, False or None):<|im_end|>
<|im_start|>assistant
```

Prompt with context for instruction-tuned LMs.

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are
    a helpful assistant.<|im_end|>
<|im_start|>user
Is the claim True or False based on the
    accompanying evidence? If you are not sure
    or cannot answer, say None. Only answer with
     True, False or None. Examples of claims,
    evidence and desired answers are given below
    .

# Example 1
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

# Example 2
Claimant: Sara Daniels
Claim: "Blackpink released the single 'You me
    too' in 2026."
Evidence: "Blackpink released their album 'Born
    Pink' in 2022."
Answer: None

# Now, answer for the following claim:
Claimant: <claimant>
Claim: "<claim>"
Evidence: "<evidence>"
Answer (True, False or None):<|im_end|>
<|im_start|>assistant
```

C CCU metric

BCU cannot measure the difference in model behaviour when context is introduced, as it does not take model behaviour without context into consideration. To address this, we introduce CCU. Given a query Q and context C, CCU measures the change in probability for token t as follows.

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

For relevant contexts C we record $\mathrm{CCU}(t_C)$, i.e. the scores for the token promoted by the context. For irrelevant contexts we record the $\mathrm{CCU}(t_M)$, i.e. the scores for the top token predicted by the model when prompted without context (memory). The range of CCU is $[-1, 1]$, for which a value of -1 denotes that the model goes completely *against* the context when the context is relevant or against its memory when the context is irrelevant, and vice versa for CCU values of 1. We report the averaged CCU per context type.

By measuring the token probabilities before and after context is introduced, the CCU metric more accurately captures how the LM is impacted by context. However, this metric excludes the Command A model, which does not provide the output logits necessary to compute CCU scores.

D Hyperparameter Search

D.1 Prompting

The tuned prompt found for each model and dataset can be found in Table 14. Different sets of prompts were experimented with depending on dataset and model type. A set of 11 to 12 prompts were produced for each of CounterFact, NQ and DRUID for the three different model types (causal LM, instruction-tuned LMs and Command A), respectively. Prompts with the same number are similar to each other across model types (e.g. Prompt #2 for Qwen2.5 on DRUID is similar to Prompt #2 for instruction-tuned Qwen2.5 on DRUID). Prompt sets across different datasets are dissimilar as they are adapted to align the instructions and few-shot examples to the given dataset. Prompt sets across different model types for the same dataset are dissimilar as small tweaks need to be applied for the instruction-tuned models that work less well in a purely causal language modelling setting, and for Command A that is a chat-based model. All prompts will be possible to view in the code repository of the paper.

Dataset	Model		Prompt
CounterFact	GPT2-XL	1.5B	*default*
	PYTHIA	6.9B	Prompt #10 (ChatGPT)
	QWEN2.5	1.5B	Prompt #1 (Jin et al. (2024))
		7B	Prompt #11 (ChatGPT)
		32B	Prompt #8 (ChatGPT)
	QWEN2.5-I	1.5B	Instruct-prompt #4 (manual)
		7B	Instruct-prompt #11 (ChatGPT)
		32B	Instruct-prompt #3 (manual)
	COMMAND A		Prompt #5 (ChatGPT)
NQ	GPT2-XL	1.5B	Prompt #2 (manual)
	PYTHIA	6.9B	*default*
	QWEN2.5	1.5B	Prompt #7 (ChatGPT)
		7B	Prompt #6 (ChatGPT)
		32B	Prompt #5 (manual)
	QWEN2.5-I	1.5B	Prompt #5 (manual)
		7B	Prompt #3 (manual)
		32B	*default*
	COMMAND A		*default*
DRUID	GPT2-XL	1.5B	Prompt #8 (ChatGPT)
	PYTHIA	6.9B	Prompt #2 (manual)
	QWEN2.5	1.5B	Prompt #2 (manual)
		7B	Prompt #11 (Microsoft Copilot)
		32B	Prompt #1 (manual)
	QWEN2.5-I	1.5B	*default*
		7B	*default*
		32B	Prompt #2 (manual)
	COMMAND A		Prompt #1 (manual)

Table 14: The tuned prompts for each LM. *default* denotes that the original prompt template (seen in Appendix B) worked best. "-I" denotes instruction-tuned model versions. The source of the prompt is indicated in parenthesis.

D.2 PH3

The tuned attention head configurations for PH3 can be found in Table 15. The head configurations are grouped by the top number of identified attention heads to consider and to what extent we allow mixing between context and memory heads. E.g. #25 `all` denotes all top-25 context and memory heads detected, #3 `memory` denotes the top-3 memory heads, allowing for overlap with context heads, and #1 `only memory` denotes memory heads detected without overlap with context heads when considering the top-1 context and memory heads.

D.3 Context-aware Contrastive Decoding: COIECD

Unlike other CMTs, the hyperparameters used in COIECD, α and λ, are selected following the original paper, Yuan et al. (2024), using the gold context from the validation set of NQ dataset. This deviation is necessary, as optimising COIECD's hyperparameters by maximising the average BCU across all context types causes the model to converge to using only the output distribution without context in the decoding step. This outcome arises from the nature of COIECD, where always relying on the distribution without context results in a BCU score of 1.0 for irrelevant contexts, while also causing the model to ignore context, including gold and conflicting contexts. To prevent COIECD from collapsing into regular generation without context and to enable meaningful comparison with other CMTs, we follow the hyperparameter search from the original paper. While Yuan et al. (2024) uses the same hyperparameter values across all models, our models exhibit different tendencies during hyperparameter search. Therefore, we tune the hyperparameters separately for each model to ensure a fair comparison with other methods. We search α in the range [0.0, 2.0] and λ in the range [0.1, 1.0], and the hyperparameters for each model are in Table 16.

E Implementation Details of Fine-tuning

To align the domain with our evaluation data, we curate the fine-tuning data with two QA datasets (Joshi et al., 2017; Rajpurkar et al., 2018), one FC dataset (Schlichtkrull et al., 2023), and one sentence completion dataset (Marjanovic et al., 2024). Before fine-tuning each LM, we elicit its parametric answers by querying without contexts. We then select the questions that the LM answered correctly and pair them with irrelevant and empty contexts. The fine-tuning data thus contains contexts that can be irrelevant, counterfactual, or empty. During fine-tuning, we train the LM to generate answers aligned with the provided context. When the context is irrelevant, we train the LM to be robust, i.e. ignore the context and output its parametric answer. Due to the computational costs associated with fine-tuning billion-sized LMs, we use the Low-Rank Adaptation method (Hu et al., 2021).

The LMs are fine-tuned with a learning rate of 5e-5,[8] using warm-up. To avoid overfitting, we use early stopping based on the loss on the validation set. For QA datasets, we use the train split from SQuAD 2.0 (Rajpurkar et al., 2018), and TriviaQA (Joshi et al., 2017). For a FC dataset, we take the train split from AVeriTeC (Schlichtkrull et al., 2023). For a sentence completion dataset, we take the static partition of the DYNAMICQA (Marjanovic et al., 2024). We only create counterfactual training examples with DYNAMICQA dataset. The detailed statistics for mixing the selected datasets can be found in Table 17.

F Additional Details of Multi-agent

Algorithm 1 Multi-agent

Given: question q, context c
Stage1: Relevance Assessment
Predict $f_{\text{rel}} \sim \text{LM}_{\text{rel}}(f_{\text{rel}} \mid q, c)$
if $f_{\text{rel}} =$ `Relevant` **then**
    Proceed to Stage 2
else
    return $\text{LM}(a \mid q)$ ▷ Answer w/o c
end if
Stage 2: Context-Faithfulness
Predict $a_c \sim \text{LM}(a_c \mid q, c)$
Predict $f_{\text{faith}} \sim \text{LM}_{\text{faith}}(f_{\text{faith}} \mid q, c, a_c)$
if $f_{\text{faith}} =$ `Faithful` **then**
    return a_c ▷ Answer w/ c
else
    Proceed to Stage 3
end if
Stage 3: Self-Refinement
return $\text{LM}(a \mid q, c, a_c, f_{\text{faith}})$ ▷ Self-Refined

As illustrated in the algorithm and Figure 5, we first assess relevance using the relevance agent to determine whether the provided context should be used. Then, the faithfulness agent provides feedback on the model response that was generated with context. If the feedback indicates that the initial answer is unfaithful, the model generates a self-refined answer based on that feedback. Given that these tasks require instruction-following capabilities, we restrict our evaluation to instruction-tuned

[8] Experiments with other learning rates yielded insignificant changes in performance on the validation set.

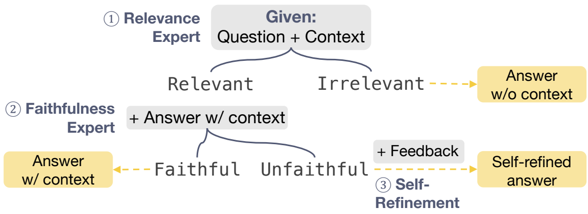

Figure 5: Overview of the multi-agent approach.

or chat LMs.

We design the `Multi-agent` approach to investigate whether LMs can explicitly handle the two objectives of context utilisation: (1) being robust to irrelevant context and (2) being faithful to relevant context. Rather than directly generating an answer, an LM is guided to perform intermediate reasoning steps, each handled by a dedicated LM agent. This decomposition allows us to understand whether LMs can explicitly recognise when the context should be used and whether their answer aligns with it when it is. While self-refinement and LM agent have been used broadly in reasoning tasks (Du et al., 2024; Feng et al., 2024; Madaan et al., 2023), our motivation is grounded in examining two components of context utilisation separately. Notably, self-refinement is only applied when the context is assessed as relevant but the answer is assessed as unfaithful, reflecting our focus on improving the usage of relevant context. By structuring the problem in this way, we aim to better understand the extent to which LMs can reason about context relevance and faithfulness.

Figure 5 and Algorithm 1 outline the `Multi-agent` procedure employed in our framework. Given a question and the context, the model first undergoes a relevance assessment stage, where it is explicitly instructed to determine whether the context is relevant to the question (Shen et al., 2024). If assessed as irrelevant, the model answers without the context; if relevant, it incorporates the context to generate the initial answer and proceeds to the next stage. In the context faithfulness assessment, the model is instructed to provide feedback on whether its answer faithfully reflects the provided context. If deemed faithful, the answer is retained as the final answer. If the prediction is assessed as unfaithful, the model is instructed to refine its answer using the question, context, initial answer, and feedback derived from the faithfulness assessment. This self-refinement stage encourages the model to self-correct based on its own feedback. To ensure consistency in output formatting during refinement, we incorporate two-shot demonstrations.

The templates for relevance assessment, context faithfulness, and self-refinement are presented below. Task-specific templates for each dataset are available in the released code.

Relevance Assessment (NQ)

```
You are a relevance assessment expert. Your task
     is to evaluate whether the provided context
     is relevant to the question.

Context: {context}
Question: {question}

If the provided context is relevant to the
    question, answer "Relevant", otherwise
    answer "Irrelevant". Do not rely on your own
     knowledge or judge the factual accuracy of
    the context.
Answer:
```

Context faithfulness (CounterFact and NQ)

```
You are a context-faithfulness expert. Your task
     is to evaluate whether the proposed answer
    faithfully uses the information in the
    provided context.

Context: {context}
Question: {question}
Proposed answer: {response}

Does the answer faithfully reflect the content
    of the context? Do not rely on your own
    knowledge or judge the factual accuracy of
    the context. Please explain briefly.

Feedback:
```

Self-refinement (NQ)

```
Your task is to generate the best possible final
     answer to the question, based on the expert
     feedback.
You may keep the original proposed answer if it
    is correct, or revise it if the feedback
    suggests it is incorrect or unsupported.
Generate only the final answer. Do not include
    any explanation or repeat the prompt.

{Two demonstrations}

Context: {context}
Question: {question}
Proposed answer: {response}
Feedback on context faithfulness: {feedback}
Final answer:
```

G Additional Details on PH3

The PH3 method is implemented in two steps: 1) identification of attention heads responsible for context or memory reliance via path patching and 2) pruning the identified attention heads for increased memory or context usage. To identify attention heads, we use the CounterFact datasets with samples that elicit exact fact recall in each studied model (Saynova et al., 2025). For the evaluation on our studied datasets, we tune the number of heads to prune on the validation splits of each evaluation dataset, similarly to the approach by Jin et al. (2024). The attention head configuration is tuned for each mode (`PH3 +memory` and `PH3 +context`, respectively).

H Input Features

We detect the input features described in Section 5.2 as follows:

- Context length is measured by the number of characters in the context.
- Flesch reading ease score is measured with the `textstat`[9] module.
- Query-context overlap is measured as the size of the set of words that form the intersection of the set of words in the query and context, respectively, normalised by the size of the set of query words. CounterFact is excluded from this analysis as its synthetic samples yield trivial results for this feature.
- The answer position is measured as the index of the answer in the context normalised by context length. This feature is only detectable for gold and conflicting contexts for CounterFact and NQ.
- The distractor rate is measured as the number of answer entities found in the context, divided by the total number of entities in the context with an entity type that matches the answer entity type(s).[10] This feature is similarly only measurable for gold and conflicting contexts from CounterFact and NQ.
- Relevance is given by the relevance agent based on Qwen 32B Instruct from the Multi-agent setup. It labels context as either 'relevant' or 'irrelevant'.

I Computational Resources

GPT2-XL was evaluated using one Nvidia T4 GPU. Pythia, Qwen 1.5B and Qwen 7B using one A40 GPU. Qwen 32B was evaluated using four A40 GPUs. The compute budget for all CMTs was about 14 hours per model for CounterFact, 28 hours per model for NQ and 21 hours per model for DRUID, amounting to a total of about 900 GPU hours.

The costs for the experiments with Cohere Command A amounted to a total of about 120 USD.

J 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.

[9] https://github.com/textstat/textstat

[10] Named entities are detected using spaCy and `en_core_web_trf`.

Model	Mode	CounterFact	NQ	DRUID
GPT2-XL	+context	#25 `all` L18H10, L21H10, L21H7, L22H18, L22H20, L24H6, L26H14, L26H20, L26H8, L27H15, L27H5, L28H15, L29H5, L29H9, L30H21, L30H8, L31H0, L31H3, L31H8, L32H13, L33H14, L33H18, L33H2, L33H7, L34H17, L34H20, L35H17, L35H19, L35H21, L36H17, L36H2, L37H7, L38H24, L38H7, L39H12, L39H9, L40H13, L40H23, L41H5, L41H9, L42H24, L43H15, L47H0	#1 `all` L28H15, L35H19	#5 `only memory`* L32H13, L35H19, L42H24, L43H15
	+memory	#12 `memory` L26H14, L26H8, L32H13, L33H14, L35H19, L38H24, L40H23, L41H5, L42H24, L43H15, L47H0, L30H8	#7 `only context` L27H15, L28H15, L29H9, L33H2, L34H17, L37H7	#22 `all` L21H10, L22H20, L24H6, L26H14, L26H20, L26H8, L27H15, L27H5, L28H15, L29H9, L30H21, L30H8, L31H0, L31H3, L31H8, L32H13, L33H14, L33H18, L33H2, L33H7, L34H17, L34H20, L35H17, L35H19, L36H17, L36H2, L37H7, L38H24, L38H7, L39H12, L39H9, L40H13, L40H23, L41H5, L42H24, L43H15, L47H0
PYTHIA 6.9B	+context	#15 `memory` L10H27, L14H6, L16H16, L17H28, L19H11, L19H21, L20H11, L20H18, L21H8, L27H22, L18H7, L19H28, L20H2, L20H8, L24H5	#17 `only memory` L10H27, L14H28, L14H6, L16H16, L17H28, L19H11, L19H21, L20H11, L20H18, L21H8, L22H12, L27H22	#10 `only context` L12H11, L12H13, L14H0, L15H17, L17H14, L20H2, L8H11
	+memory	#25 `only context` L10H1, L12H11, L12H13, L13H12, L14H0, L14H23, L15H17, L17H14, L18H10, L19H1, L19H20, L21H10, L23H25, L29H22, L8H11, L8H24	#12 `only context` L12H11, L12H13, L14H0, L14H23, L15H17, L17H14, L19H31, L20H2, L8H11	#17 `only context` L10H1, L12H11, L12H13, L13H12, L14H0, L14H23, L15H17, L17H14, L18H10, L19H1, L19H31, L8H11
QWEN2.5 1.5B	+context	#15 `only memory` L10H0, L10H1, L13H1, L16H1, L17H0, L18H0, L1H1, L3H0	#12 `only memory` L10H0, L13H1, L16H1, L17H0, L18H0, L1H1	#17 `only context` L14H1, L16H0, L18H1, L19H0, L19H1, L20H1, L24H1, L26H0, L26H1, L9H0
	+memory	#5 `only context` L15H1, L16H0, L27H0	#12 `only context` L14H1, L16H0, L18H1, L19H0, L24H1, L27H0	#12 `only memory` L10H0, L13H1, L16H1, L17H0, L18H0, L1H1
QWEN2.5 1.5B *Instruct*	+context	#7 `only memory` L15H0, L1H1, L21H0	#1 `only context` L19H1	#10 `only context` L14H0, L17H1, L19H1, L22H0, L26H0
	+memory	#1 `only context` L19H1	#12 `only context`* L14H0, L17H1, L19H1, L22H0, L26H0, L27H0	#5 `only context` L17H0, L19H1, L22H0
QWEN2.5 7B	+context	#7 `memory` L0H0, L17H1, L18H2, L19H0, L21H0, L22H2, L23H0	#1 `only context` L27H0	#3 `only memory` L0H0, L22H2
	+memory	#15 `only context` L13H0, L17H0, L18H1, L18H3, L22H0, L24H3, L25H1, L26H0, L27H0, L27H2	#5 `only context` L22H0, L27H0, L27H2	#12 `only context` L16H3, L17H0, L18H1, L18H3, L22H0, L24H3, L26H0, L27H0, L27H2
QWEN2.5 7B *Instruct*	+context	#17 `only memory` L11H1, L12H0, L13H3, L14H3, L16H1, L17H0, L17H3, L18H2, L1H1, L20H0, L21H2, L26H3, L3H0	#5 `context` L18H0, L18H3, L22H2, L23H0, L27H2	#5 `only context` L18H0, L18H3, L27H2
	+memory	#3 `only context` L18H0	#3 `only context` L18H0	#17 `all` L0H0, L11H1, L12H0, L13H3, L14H3, L15H1, L16H0, L16H1, L17H0, L17H3, L18H0, L18H1, L18H2, L18H3, L19H0, L19H3, L1H1, L20H0, L20H2, L20H3, L21H0, L21H2, L22H0, L22H2, L23H0, L26H3, L27H0, L27H2, L3H0, L8H1

Table 15: Tuned PH3 attention head configurations for each model and evaluation dataset. +context indicates heads for which pruning leads to increased context usage and vice versa for +memory. Configurations marked with * denote that they yielded degraded performance compared to the standard setting (no mechanistic intervention) on the validation set.

Model	λ	α
GPT2-XL	0.50	1.00
PYTHIA 6.9B	0.50	1.00
QWEN2.5 1.5B	1.00	0.50
QWEN2.5 1.5B INSTRUCT	0.50	1.00
QWEN2.5 7B	1.00	1.00
QWEN2.5 7B INSTRUCT	0.50	0.50
QWEN2.5 32B	0.50	1.00
QWEN2.5 32B INSTRUCT	0.50	1.50

Table 16: Selected COIECD hyperparameters λ and α for each model, evaluated on gold contexts from NQ's validation set. For models with multiple (λ, α) pairs attaining the maximum score, we choose the setting that lies near the midpoint of the optimal region.

Dataset	**Dataset weight**	**Context type**	**Context weight**
SQuAD 2.0	0.4	Relevant	0.65
		Irrelevant	0.25
		Empty	0.1
TriviaQA	0.3	Relevant	0.65
		Irrelevant	0.25
		Empty	0.10
AVeriTeC	0.15	Relevant	0.65
		Irrelevant	0.25
		Empty	0.10
DYNAMICQA	0.15	Relevant	0.50
		Irrelevant	0.05
		Empty	0.05
		Counterfactual	0.40

Table 17: Sampling weight for each dataset. We first sample the number of instances for each dataset following the dataset sampling weight. Then, each context type is determined by the context sampling weight.