DYNAMICQA: Tracing Internal Knowledge Conflicts in Language Models

Sara Vera Marjanović[*]   **Haeun Yu**[*]   **Pepa Atanasova**
Maria Maistro   **Christina Lioma**   **Isabelle Augenstein**
Department of Computer Science
University of Copenhagen
{savema, hayu}@di.ku.dk

Abstract

Knowledge-intensive language understanding tasks require Language Models (LMs) to integrate relevant context, mitigating their inherent weaknesses, such as incomplete or outdated knowledge. However, conflicting knowledge can be present in the LM's parameters, termed intra-memory conflict, which can affect a model's propensity to accept contextual knowledge. To study the effect of intra-memory conflict on an LM's ability to accept relevant context, we utilize two knowledge conflict measures and a novel dataset containing inherently conflicting data, DYNAMICQA. This dataset includes facts with a *temporal* dynamic nature where facts can change over time and *disputable* dynamic facts, which can change depending on the viewpoint. DYNAMICQA is the first to include real-world knowledge conflicts and provide context to study the link between the different types of knowledge conflicts. We also evaluate several measures on their ability to reflect the presence of intra-memory conflict: semantic entropy and a novel *coherent persuasion score*. With our extensive experiments, we verify that LMs exhibit a greater degree of intra-memory conflict with dynamic facts compared to facts that have a single truth value. Furthermore, we reveal that facts with intra-memory conflict are harder to update with context, suggesting that retrieval-augmented generation will struggle with the most commonly adapted facts.

🔗 copenlu/dynamicqa 🤗 copenlu/dynamicqa

1 Introduction

Language models (LMs) have been useful in a variety of downstream applications from summarization to fact-checking, often relying on the factual knowledge memorized during pre-training

and stored in their parameters, known as **parametric knowledge** (Yu et al., 2024). However, this internal memory is not infallible; it may contain misinformation, biases or simply outdated data, causing LMs to produce factually incorrect output, occasionally termed 'hallucinations'(Huang et al., 2023). Furthermore, conflicting representations of a fact may exist within training data, given conflicting viewpoints ('disputable facts') or temporal changes ('temporal facts') (See Figure 1). We refer to this superposition of fact representation as the fact's *dynamicity*. In contrast, a *static* fact has only one possible representation. This can lead to **intra-memory conflicts**, which can contribute to LM's uncertainty and instability in factual recall. This intra-memory conflict can be likened to a semantic expression of aleatoric uncertainty, noise inherent in training data (Kendall and Gal, 2017).

One way to guide LMs to reliable answers is via Retrieval-Augmented Generation (RAG) (Guu et al., 2020), which provides additional contextual knowledge, known as **non-parametric knowledge**, to a language model. However, this context may conflict with the model's parametric knowledge, resulting in a **context-memory conflict**. Initial research has found a tendency of models to over-rely on parametric knowledge in the face of these knowledge conflicts (Longpre et al., 2021; Chen et al., 2022), thus ignoring the retrieved context. While fact popularity has been known to reduce context utilization (Mallen et al., 2023), little is known of the impact of internal memory conflicts. We explore this issue for the first time by looking at the effect of fact dynamicity on context utilization.

In this paper, we investigate the effect of intra-memory conflicts on context adaptation using a novel dataset DYNAMICQA. This dataset contains inherently conflicting temporal and disputable facts, paired with realistic retrieved contexts in which they occur. Furthermore, we manipulate these contexts to include similar, competing objects. Using
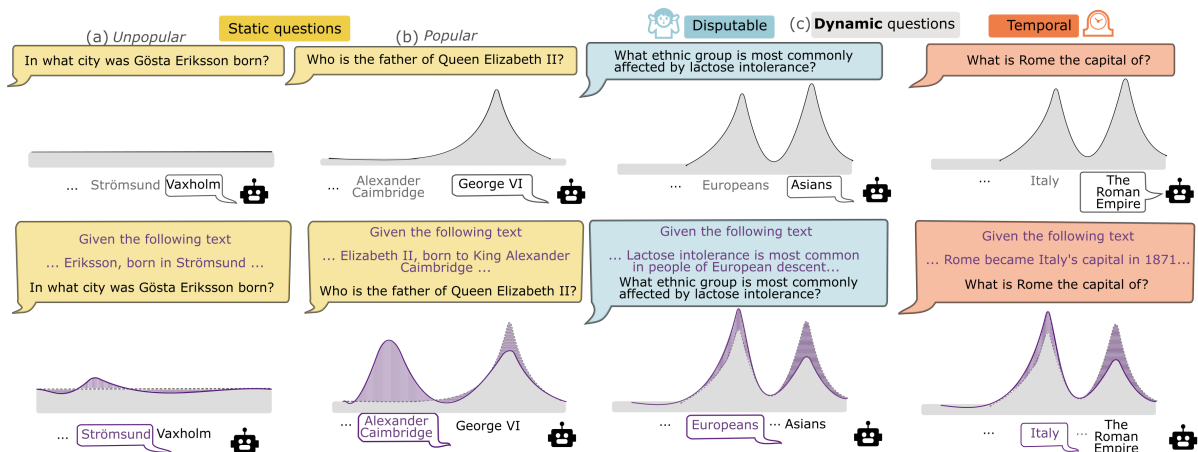
---

[*] Equal contribution. Corresponding authors.

Figure 1: We present examples of our dataset here, consisting of static, temporal, and disputable facts. We show how model output distribution can vary due to the popularity (a,b) and dynamicity (c) of facts. Fact dynamicity (c) causes intra-memory conflicts between the different fact representations seen during pretraining. In the bottom row, we show the change in output probability (purple area) that the context must enact on the initial output distribution (dotted line) to force a new output distribution (purple line).

DYNAMICQA, we study the interaction of intra-memory and context-memory conflicts by adapting two existing conflict measures. First, we approximate an LM's fact dynamicity ground truth via the number of recorded representations for a fact in Wikidata, and sample the LM's expressed intra-memory conflict using semantic entropy, which captures the semantic variation present in parametric memory (Kuhn et al., 2023). Furthermore, we introduce a coherent persuasion score, based on Du et al. (2024) to approximate an LM's semantic shift in output distribution given competing context.

With our framework, we show the effect of intra-memory conflict on dynamic fact question answering. We find that manipulated contexts of static facts and facts with low dynamicity show the greatest persuasiveness for LMs, despite their limited variance in the training dataset and, presumably, low likelihood to change. Furthermore, we find that semantic entropy alone does not reflect an instance's likelihood for persuasion given additional context, suggesting that retrieval-augmented generation language models struggle in low-certainty domains. These results underline the need for new measures of intra-memory conflict and the need for other indicators of context use in RAGs (Ni et al., 2024).

2   Related Work

Knowledge Conflicts   Xu et al. (2024b) provide a survey for knowledge conflicts in LMs; we focus specifically on intra-memory conflicts and their interplay with context-memory conflicts. Longpre et al. (2021) first showed that LMs ignore context that contradicts their parametric memory, generating facts not stated in the input, termed 'hallucinations'. Further work either disentangles the two forms of knowledge (Neeman et al., 2023) or finds other methods to resolve hallucination (Song et al., 2024). Other studies investigate the effect of a fact's frequency in the training data (Yu et al., 2023) or the quality of the presented counter-memory (Xie et al., 2024; Wan et al., 2024) on LM hallucination. Furthermore, some works find that LMs are swayed by convincing misinformation at inference-time (Pan et al., 2023; Xu et al., 2024a; Wan et al., 2024). These findings suggest some pattern to a model's proclivity to ignore context. However, in contrast to these existing works, our work *considers realistic, natural examples of knowledge conflict* as they occur in the real world and *how different types of knowledge conflicts are connected*.

Measures for Knowledge Conflicts   There have been several approaches to quantify or approximate the degree of a knowledge conflict; however, these works only focus on context-memory conflicts. Pezeshkpour (2023) focus on the entropy of the LM's probability distribution for an answer; the difference in entropy depending on the presence or absence of additional context indicates the LM's prior knowledge on a certain fact. On the other hand, Du et al. (2024) propose a susceptibility and a persuasion score to investigate the LM when the given context contradicts the LM's parametric

knowledge, based on the first token probabilities. The susceptibility score of an entity shows how easy it is to shift an LM's probability distribution of an answer regarding an entity. The persuasion score of a context represents how effective the context is at changing an LM's probability distribution of an answer. Given known issues in the reliance on first-token probabilities (Wang et al., 2024), we *reformulate this score with consideration of semantic consistency* in §4.3. Furthermore, given the lack of measures for intra-memory conflict, we *assess semantic entropy (Kuhn et al., 2023) as an indicator of intra-memory conflict (Gao et al., 2024).*

Sources of Factual Errors   We identify three causes of model factual error: fact popularity, temporality, and disputability. Fact popularity approximates the prevalence of a fact in training data and has been shown to affect model performance (Mallen et al., 2023) and a model's likelihood to be swayed by context ('susceptibility score', Du et al. (2024)). In contrast, temporality and disputability reflect a fact's entropy within the training and test data. For example, it is well-established that the passing of time can lead to outdated models, stressing the need for evaluation of models for a specific time-frame (Margatina et al., 2023; Fierro et al., 2024) and updating of specific temporal facts in LMs (Zhang and Choi, 2023; Jang et al., 2022). Similarly, Xu et al. (2024b) argue that context-memory knowledge conflicts arise due to either temporal misalignment or misinformation pollution of the training dataset. In a contemporary study, Fierro et al. (2024) show that facts with dynamic (or in their terms 'mutable') relations show different patterns in model confidence, knowledge representation, and contextual alignment. All of these studies look at only one source of factual error at a time and do not assess in the context of knowledge conflicts. In our work, we *introduce disputability* as a contributor to factual error, study *multiple sources of factual error concurrently* (temporality, disputability, and popularity), *assess their interaction*, and *provide real-world proxy measures* for all three sources.

3   DYNAMICQA

To evaluate the intra-memory conflicts, we present a dataset of 11,378 question-answer pairs featuring facts with varying levels of dynamicity (see Figure 1). The dynamicity of a fact is difficult to determine via the pretraining dataset, so we use

easily measured proxy scores as a reflection of the property. The question and answer pairs, alongside their proxy scores, are sourced from Wikidata and Wikipedia and have not previously been employed in similar datasets nor for approximating the degree of knowledge conflicts in general. We approximate temporality via the number of edits (§3.1) and disputability via the number of reversions (§3.2). For each question and answer pair, we obtain answer-specific context snippets from Wikipedia, alongside their popularity scores, estimated from Wikipedia page views. For static and temporal facts, we also create 'counter-memory' from these snippets using similar object replacements, to simulate realistic knowledge conflict scenarios.

3.1   Temporality

We estimate the **temporality** of a triplet via the number of edits on Wikidata for the given subject and relation as a proxy. We initialize our dataset from PopularQA (Mallen et al., 2023), which is a collection of 14k question-answer pairs sampled from Wikipedia, with a **popularity** score determined by the monthly Wikipedia page views of the subject and object of the triplets. This popularity score is often used as a proxy for the triplet's prevalence in the unsearchable pre-training corpora of models (Mallen et al., 2023; Fierro et al., 2024). Given the sampling method, this dataset has a long-tailed distribution, meaning most triplets have very low popularity and low temporality. We identify relevant snippets from the triplet's subject current Wikipedia page that also mention the triplet's object and use them as context to provide to the LM. We discard 2k pairs where we cannot find the intended object and relation mentioned on the subject's Wikipedia page. To construct counter-memory contexts, we replace the object in context with a replacement token. We identify relevant replacement entities using the most similar entity for the original object as identified using the Wembedder tool (Årup Nielsen, 2017). All facts with more than 1 edit are labelled as 'temporal' facts, though they vary in their degree of temporality (with a maximum score of 23 edits). This leaves us with 2495 questions (Table 1). We randomly subsample 2500 of the remaining 'static' questions (triplets with 0 edits on Wikidata).

3.2   Disputability

Disputable facts are facts that differ depending on the viewpoint. To construct the question and an-

swer pairs, we utilize the collection of Wikipedia articles that are regarded as controversial.[1] Controversy, by definition, is inherently disputable since it arises from the existence of multiple viewpoints, each supported by sufficient evidence (Rad and Barbosa, 2012; Vuong et al., 2008). We are the first to convert the controversies that appear in the Wikipedia edit history into a QA dataset for LMs.

To identify disputable facts from the controversial articles, we look for reversions in the edit logs (Sumi et al., 2011). Among the edit logs $[e_0, \ldots, e_{l-1}, e_l, e_{l+1}, \ldots, e_j]$ on an article, two consecutive logs, e_l and e_{l+1} are selected as a pair of reverted edit logs if the texts of e_{l-1} and e_{l+1} are identical. Then, by measuring the edit distance between e_l and e_{l+1}, we select a pair of words from e_l and e_{l+1} if replacing the word in e_l with the word in e_{l+1} yields the same text as e_{l+1}.

Since a reverted edit does not necessarily imply the disputability of a fact, we apply additional rules to filter out cases of vandalism, paraphrasing, or synonyms. For vandalism, we remove the pair if one of the users involved in that edit did not disclose their identity, for example, anonymous users or if the user ID is an IP address. For the pairs that are synonyms or paraphrasing, we feed them to a semantic similarity model (§A.3) and remove the pairs whose similarity scores are bigger than 0.98. The selected pairs of reverted edits and the text snippets they are located in serve correspondingly as answers and contexts in our dataset.

Questions are generated using an LM[2](§A.2) by providing the context e_l and the corresponding ground-truth answer within the prompt (see Table 5 in the Appendix). After obtaining the question and answer pair with the context, three annotators manually annotated the dataset to ensure its quality. Two annotators annotated each data point, and conflicts were resolved by the third annotator. We obtain a Krippendorf's alpha of 0.44 and provide further details in Appendix A.3. For the real-world proxy score of disputability, we count the number of reversions on (e_l, e_{l+1}). As a result, we obtain 694 questions with two possible disputable answers and their corresponding real-world proxy scores (Table 1).

---

4 Measuring Knowledge Conflicts

In the pretraining stage, LMs learn a conditional distribution to predict the next tokens given a preceding context; in this process, they also obtain world knowledge into their parametric memory (Chang et al., 2024). For example, in Figure 1, given a question (e.g., 'Who is the father of Queen Elizabeth II?'), the conditional probability of the correct answer 'George IV' is expected to be greater than similar objects, owing to the frequent co-occurrence of the subject, object and relation in the training corpus. One important element impacting fact acquisition and recall is the frequency of the fact in the training data. It can be assumed that high frequency (often approximated via fact popularity (Mallen et al., 2023)) leads to a greater strength of association between the subject and the object, resulting in greater probability being attributed to the relevant tokens (Figure 1b). However, as information evolves over time, multiple different representations of a factual triplet may appear in a training corpus; meaning, given a subject and relation, there can exist a superposition of object representations (Shanahan et al., 2023) depending on the temporal and situational context. Rather than a singular object being assigned a high probability, for such facts, we expect several competing answers in the output distribution (Figure 1c).

To evaluate this hypothesis, we expect to see two properties: (1) higher entropy in the output distribution of dynamic facts and (2) greater change in the output distribution to update dynamic facts. To verify (1), we look at the semantic entropy of output distributions (§4.2). To investigate (2), we analyse the amount of effort required to shift the output distribution using both loss and our novel Coherent Persuasion score (§4.3). Both measures look at semantic changes in LM output, thereby limiting confounds due to syntactic inconsistencies and other issues arising from over-reliance on first-token probabilities (Wang et al., 2024).

4.1 Preliminaries

A dataset $D = [(c_1, q_1, y_1), \ldots, (c_N, q_N, y_N)]$ consists of N tuple instances containing: an answer-containing context c, a question q, and a ground-truth answer a. For the i-th instance, an LM f outputs an answer $y_i = [y_i^1, \ldots, y_i^h, \ldots, y_i^H]$, consisting of H tokens given an input x_i. An input x_i here contains a universal prompt P describing

the QA task with either q_i only ($x_{i;q} = [P; q_i]$) or with q_i and the context c_i ($x_{i;q,c} = [P; q_i, c_i]$). Given a contextless input ($x_{i;q}$), a model outputs $y_{i;q}$ (i.e. the parametric knowledge). Given context ($x_{i;q,c}$), we denote the model output as $y_{i;q,c}$.

4.2 Semantic Entropy

Entropy allows us to measure the amount of information within a distribution and is a common measure for model uncertainty. Both lack of information (Figure 1a) and competing information (Figure 1c) can contribute to uncertainty and lead to more uniform output distributions. We propose semantic entropy as an estimator of intra-memory conflict, given its use of high-temperature sampling to elicit diverse answers contained within parametric memory. We validate this measure through the lens of a fact's dynamicity (defined in §3).

To do so, we use Kuhn et al. (2023)'s approach in our simulated knowledge conflict setting. Given any input x_i, which contains a universal prompt and a question q_i (and optional context c_i), we generate K model outputs $Y = [y_{i,1}, \ldots, y_{i,K}]$. Next, we group the generated answers according to their semantic similarity. The semantic similarity between two sampled answers is calculated using a DeBERTA Natural Language Inference (NLI) model.[3] Details about answer generation and semantic grouping can be found in Appendix A.3. The grouping according to the semantic similarity results in the V groups $G = [g_1, g_2, \ldots, g_v, \ldots, g_V]$, where $1 \le V \le K$.

The semantic entropy is then estimated by the entropy between the semantic sets. First, we obtain $p(g_v|x_i)$, the conditional token probabilities output by the model generating the answers in g_v given the input x_i via Equation 1.

$$
\begin{aligned}
p(g_v|x_i) &= \sum_{y_{i,k} \in g_v} p(y_{i,k}|x_i) \\
&= \sum_{y_{i,k} \in g_v} \prod_h p(y_i^h|y_i^{<h}, x_i)
\end{aligned}
\tag{1}
$$

In this case, h refers to the intermediate tokens in the entire output length. With the probabilities of the separate groups, we approximate $SE(x_i)$, the overall expectation of the semantic entropy of x_i, using Monte Carlo integration over the groups:

$$
SE(x_i) \approx -V^{-1} \sum_{v=1}^{V} \log p(g_v|x_i)
\tag{2}
$$

---

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

4.3 Coherent Persuasion Score

We posit that the decreased context utilization in the case of highly popular facts (Mallen et al., 2023; Du et al., 2024) owes to the greater likelihood attributed to the learned answer in the unconditioned output distribution (given $x_{i;q}$). Therefore, greater effort, either via parameter updates or contextual knowledge, is required to shift this output distribution to match the new answer provided in the context (a_c) in the conditioned instance (given $x_{i;q,c}$, Figure 1b). This is further exacerbated when there is a superposition of competing answers (Figure 1c). The amount of effort that must be enacted on the model parameters to output a_c can be approximated by the loss, which also reflects the LM's perplexity to the output. We also measure the observed magnitude of this shift in output distribution via our novel Coherent Persuasion (CP) score. This score quantifies that actual efficacy of the context in shifting an LM's output distribution.

Previously, persuasion scores were introduced to assess to what degree context changes an LM's answer (Du et al., 2024). However, the existing persuasion score focuses on the first token of the single generated answer, which has been shown to be insufficient to represent the entire generated sequence (Wang et al., 2024). There also exists the problem of uncertainty during the generation of a single forward pass, which may also stem from syntactical uncertainties, which are irrelevant to our use case.

To overcome the LM's brittleness, we propose the novel Coherent Persuasion (CP) score. We incorporate a multiple-sample approach and semantically group the samples to create answer distributions for both $x_{i;q}$ and $x_{i;q,c}$ and compare the two. To create answer distributions, we adapt the generation and semantic grouping process as in Kuhn et al. (2023) into the CP score.

To calculate the CP score of our context c_i, we gather two lists of answers, $Y_{i;q}$ and $Y_{i;q,c}$, with two different inputs, which correspond to the question only input $x_{i;q}$, and the question with context input $x_{i;q,c}$. Then, we create the semantically similar groups $G_{i;q} = [g_1, g_2, \ldots, g_r, \ldots, g_R]$ and $G_{i;q,c} = [g_1, g_2, \ldots, g_u, \ldots, g_U]$ from $Y_{i;q}$ which has R groups of generated answers and $Y_{i;q,c}$ which has U groups of generated answers, respectively. The CP score is then obtained by averaging the divergence from the probability distribution p of $G_{i;q,c}$ to p of $G_{i;q}$. In detail, the output probability

distribution of g_r and g_u, is obtained as:

$$p_{g_w} = \frac{1}{W} \sum_{w=1}^{W} p_{y_w} \qquad (3)$$

where $W \in \{R, U\}$ and p_{y_w} is the averaged soft-max probability distribution of all the tokens in the answer a_w. Our final proposed CP score can be acquired by Equation 4.

$$CP(c_i) = \frac{1}{|R| \times |U|} \sum_{r=1}^{R} \sum_{u=1}^{U} KL(p_{g_r}, p_{g_u}) \quad (4)$$

As the CP score approximates the distance between the probability distributions of $Y_{i;q,c}$ and $Y_{i;q}$, it tells us how much the LM's probability distribution is swayed by the context, while ensuring that small changes in syntactic representation or mismatches from first-token probabilities are not included. It is therefore more 'coherent' in that it considers (1) the *entirety* of the LM's output and (2) only the *semantic* divergence in output.

5 Experiments

5.1 Setup

We use DYNAMICQA to assess the performance of three recent and similarly sized state-of-the-art LMs: Mistral-7B-Instruct-v0.1 (Mistral, Jiang et al. (2023)), Llama-2-7b-chat-hf (Llama-2, Touvron et al. (2023)), and Qwen2-7B-Instruct (Qwen2, Yang et al. (2024)). To minimize the effect of the confounding factors, inferences are done in zero-shot manner, and, to obtain better generation results, instruction-tuned LMs are chosen. To obtain the model's parametric knowledge ($y_{i;q}$), we first query the model for each q_i without any additional context. We then query the model provided two forms of context: one is the unperturbed context c_o and the other is the unseen replacement c_c. In the case of disputable instances, the choice of c_o and c_c is arbitrary, as both options are equally likely. For each query, we obtain the accuracy and semantic entropy (§4.2). We calculate the CP score for each q_i across all queries (§4.3). We calculate the accuracy of each model on each fact type as

$$acc = \frac{\sum_{i=1}^{N} RougeL(a_i, y_i) > 0.3}{N}$$

where $a_i \in \{a_c, a_o\}$, when context is provided, and is otherwise a_o. Given the inherent subjectivity of questions within the disputable portion of the

dataset, we do not calculate accuracy for that partition without the context provided. For generating samples, we set $K = 10$ and calculate accuracy values using the most likely response as determined via greedy search.

5.2 Analysis

We first look into general differences in model performance on static and dynamic facts, by comparing changes in accuracy, entropy and persuasion.

We identify the difficulty of updating dynamic facts by identifying two model behaviours of interest: **Persuaded** instances are instances where the model is persuaded by the provided context, meaning that $y_{i;q} \neq a_c$ but $y_{i;q,c} = a_c$. **Stubborn** instances are instances where the model is impassive to the provided context, meaning $y_{i;q} = y_{i;q,c}$ and $y_{i;q} \neq a_c$. These are exceptional instances of RAG context usage failure. We compare the obtained CP and SE scores across the entire dataset and also highlight the behaviour of the persuaded and stubborn instances (§6.1). To further explain the difficulty of retrieval-guided model updates, we observe the differences in loss (§6.2).

To identify factors contributing to persuasion (or RAG context usage success), we look at three characteristics of interest on the temporal dataset: semantic entropy, temporality (the number of edits), and popularity (approximated by pageviews). We analyse the Pearson correlation between these three values and our CP score to identify the strongest correlates of persuasion (§6.3). To further understand the influence of the various potential predictors of an instance's persuasion, we implement a logistic regression model on stubborn and persuaded instances. We take as dependent variables the popularity of the subject and object, the number of edits to the fact, and the semantic entropy before and after context is provided (§6.4). We standardise each parameter with a z-score transformation for interpretable comparison.

6 Results & Discussion

6.1 General Performance

Table 1 shows the percentage of stubborn and persuaded instances and Table 2 shows the performance of our three investigated models across the three partitions of our dataset with and without additional context. Semantic Entropy (SE) reflects the semantic variety of model output, which can indicate internal memory conflict (§4.2). Lower

	# of Questions	# of Instances	% of Stubborn Instances			% of Persuaded Instances		
			Llama-2	Mistral	Qwen2	Llama-2	Mistral	Qwen2
Static	2500	5000	6.16%	5.44%	6.92%	**78.44%**	**70.52%**	**61.48%**
Temporal	2495	4990	**9.38%**	**7.01%**	**7.54%**	60.96%	51.62%	44.81%
Disputable	694	1388	9.36%	6.48%	7.35%	63.83%	62.53%	59.51%

Table 1: The number of collected questions and instances in DYNAMICQA (§3) for each fact type. We also report general model behaviour (i.e. percentage of persuaded instances given context), as further described in §5.2.

		Accuracy (↑)			Semantic Entropy (↓)			Coherent Persuasion Score (↑)		
		Llama-2	**Mistral**	**Qwen2**	**Llama-2**	**Mistral**	**Qwen2**	**Llama-2**	**Mistral**	**Qwen2**
Static	with context	**0.8476**	**0.7644**	**0.6814**	15.5663	11.7557	10.5875	**6.8665**	**5.8550**	**4.1567**
	w/o context	**0.1306**	**0.0902**	**0.1244**	17.7064	**11.4943**	10.4271			
Temporal	with context	0.6619	0.5677	0.5040	**15.3947**	**10.7685**	**10.5264**	6.5941	5.6314	3.9551
	w/o context	0.1036	0.0719	0.0866	**17.3518**	11.7410	11.0875			
Disputable	with context	0.6455	0.6253	0.5937	16.5803	11.6632	10.9627	5.6027	4.1147	3.3955
	w/o context	-	-	-	18.9694	12.4214	**10.3957**			

Table 2: The average accuracy, Semantic Entropy (SE; §4.2) and Coherent Persuasion (CP; §4.3) score of our models. We bold the best values per column, with and without context. Given the inherent subjectivity of the Disputable facts, we do not show accuracy without context.

SE can be interpreted as more consistent, and less conflicted, output. On the other hand, our Coherent Persuasion (CP) score quantifies the actual efficacy of the context in shifting an LM's output distribution (§4.3).

Accuracy is typically greatest with Llama-2 across all partitions. We also see that Llama-2 has the greatest degree of persuasion (in both persuasion score and % persuaded) across all partitions. However, Llama-2 also shows the greatest entropy across all instances. This suggests *Llama-2 is typically the most effective at utilising context, yet also shows the greatest intra-memory conflict.* Typically, entropy decreases with context; however, Mistral and Qwen2 show a slight increase in entropy for the Static partition of the dataset; this likely owes to the introduction of intra-memory conflict for instances with typically no competing objects. Furthermore, we typically see the *greatest accuracy and persuasion (bolded) on the Static partition*. In contrast, we typically see the *lowest entropy (bolded) on the Temporal dataset.* This is typically when context is provided, suggesting that the introduction of context reduces the exhibited intra-memory conflict for temporal facts.

Our results are somewhat similar to contemporary work (Fierro et al., 2024), who also find the greatest accuracy with Llama-2; however, they also find the highest confidence with this model, though they presumably use the output of softmax layer

which is more vulnerable to syntactic uncertainties (Kuhn et al., 2023). Furthermore, they find reduced update accuracy for Llama-2 relative to other models across mutable and immutable facts. In contrast, we surprisingly find that manipulations to static, immutable facts are most easily accepted by the model in context. In contrast, the temporal and disputable facts, which are expected to have greater variability in the model's training dataset, and thus higher intra-memory conflict, have a greater proportion of stubborn instances across all models. We expect this stems from our differences in dataset collection, as Fierro et al. (2024) determine mutability via relationship type, whereas we verify the 'mutability' of our facts with our proxy measures. Ultimately, our results raise concerns for the efficacy of context usage in retrieval-augmented text generation, as *we find that the most commonly adapted facts (dynamic facts) are the most difficult to update with context*.

We now investigate potential factors contributing to context usage ('persuasion') and neglect ('stubbornness'). For the main paper, we present results for Llama-2, due to its high performance, on the temporal partition of the dataset, as this partition allows for easy comparison of the effect of dynamicity on model performance. We provide all the additional graphs for the remaining models and partitions in Appendix B.
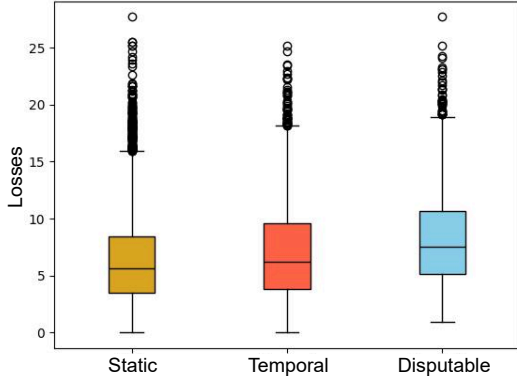
Figure 2: The distribution of loss ($\mathcal{L}(a_c, x_{i;q})$) for generating a particular answer (a_c) given only the question ($x_{i;q}$) for each partition of the dataset.

6.2 Obstacles to Persuasion

To further investigate why dynamic facts show a greater proportion of stubborn instances yet a smaller divergence in output distributions, we look at the loss when we force the target answer (a_c) for the contextless input $x_{i;q}$. The loss reflects the likelihood of an output given the model's trained parameters. A higher loss, thereby, also indicates greater change required to steer the LM to output the target answer. Thus, we expect higher loss on dynamic facts. Figure 2 shows the distribution of loss for each partition of the dataset. We construct the input only with the question and calculate the losses for all the tokens in the target answer (a_c). The losses from the tokens in the target answer are averaged for each question. In Figure 2, we observe that losses for temporal and disputable questions are greater than the loss for static questions; the significance is confirmed with a Welch's t-test ($p = 9.2e - 64$, $p = 3.4e - 18$). Therefore, while static questions show a greater CP score in Table 2, we find *there is more effort involved in updating the LM's parameters for dynamic facts, which may be unachievable with only additional context.*

6.3 Interactors with Persuasion

We now analyse the interaction between the persuasion (CP) and semantic entropy (SE) scores, subject popularity, and object temporality in Figure 3 and highlight the persuaded and stubborn instances. In general, we can see that stubborn instances (orange) have lower CP scores than persuaded (blue) and generic (i.e. all remaining) instances (gray). We do not see a particular correlation between the

CP score and SE ($r(2494) = 0.003, p > .05$): meaning, *instances with high initial entropy are not necessarily more likely to be persuaded to match the given context.* The distribution of persuaded instances and generic instances are identical, and there is a strong overlap with the distribution of stubborn instances. Therefore, it is difficult to determine if a retrieved context will be used by the model based on semantic entropy.

While popularity has been previously reported to negatively impact persuasion (Du et al., 2024; Mallen et al., 2023), popular instances are not overly represented among the stubborn instances, and the strength of the correlation ($r(2494) = 0.101, p = 4e - 7$), though statistically significant, is still weak. In contrast, temporality scores show the strongest correlation with CP ($r(2944) = -0.27, p = 1.6e - 42$), suggesting *there is a stronger relation between persuasion and object temporality than between persuasion and subject popularity or semantic entropy.*

Furthermore, given the strong differences in the distributions in Figure 3, we can see that semantic entropy does not indicate epistemic uncertainty (which is approximated by subject popularity) nor aleatoric uncertainty (approximated by temporality) (though semantic entropy was previously used as a measure of aleatoric uncertainty in other work (Gao et al., 2024)). This suggests that other factors may contribute to semantic entropy, outside of exclusively intra-memory conflict. While knowledge conflicts are a straightforward way to elicit meaningful semantic inconsistencies, they can come from other effects of model training (i.e., overlaps of objects in model memory space, spurious signals, common word co-occurrences), which may be picked up by the semantic entropy measure in addition to intra-memory conflict.

6.4 Predictors of Persuasion

In our previous investigations, we see that the semantic entropy of a contextless question is not a meaningful indicator of model persuasion, though we do see stronger relationships for popularity and temporality. Furthermore, dataset partitions with anticipated increase in intra-memory conflict (i.e., temporal and disputable data) show lower persuasion than static facts. Therefore, we assess the strength of each relation to persuasion itself. We show the results of our logistic regression test in Table 3. Though we do not find a very strong fit to the
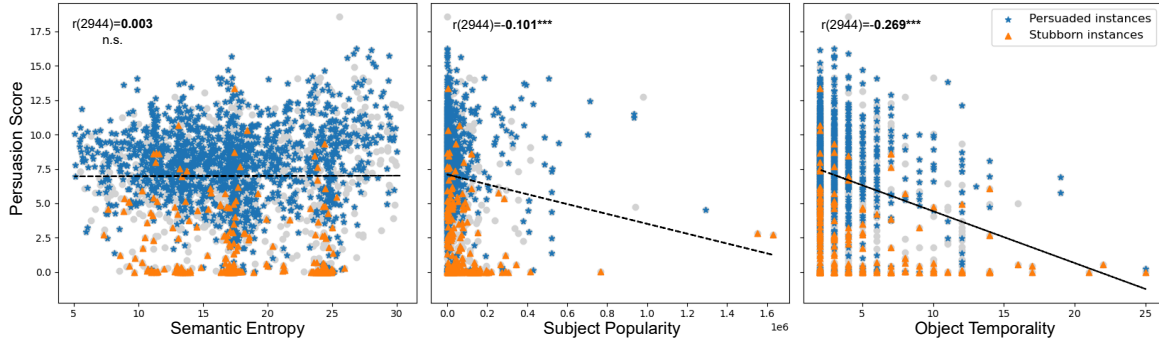
Figure 3: The instance-level relationship between Coherent Persuasion score (§4.3) and 3 possible factors that impact persuasion: semantic entropy, subject popularity and object temporality, alongside their Pearson correlation scores. Temporality shows the strongest relationship with persuasion. We also highlight two behaviours of interest: persuaded and stubborn instances.

data, given its volatility ($R^2 = 0.139$), we do find several significant predictors. *The strongest predictor of persuasion is the number of edits of a fact, where it shows an inverse relationship to a fact's persuasion.* This aligns with the behviour we have seen in our comparison of temporal and static facts in Table 1. Furthermore, while object popularity (o_{pop}) significantly affects a fact's persuasion score (previously shown by Du et al. (2024)) the magnitude of this effect is smaller than that of the number of edits of a fact. These findings typically hold across other models (see Appendix B), where we show that SE_q typically shows the weakest relation to model persuasion, and the number of edits to have one of the strongest relationships. Taken together, we can see that *fact dynamicity, or intra-memory conflict, plays a bigger role in eliciting knowledge conflicts than fact popularity* (Mallen et al., 2023; Du et al., 2024). We also find *an inverse effect of intra-memory conflict on an LM's susceptibility to persuasion for a given instance.* Facts that change regularly are less likely to be updated with context-retrieval, yet facts that never change are easily persuaded.

7 Conclusion

We investigate, for the first time, the effect of intra-memory conflict on context adapation, using multiple natural causes of intra-memory conflict (i.e., fact 'dynamicity'). In this, we introduce a Coherent Persuasion score to measure the persuasiveness of the given context with consideration of the entire semantic output. Furthermore, we release DYNAM-ICQA, a QA dataset with realistic dynamic and static questions, (i.e., facts that naturally vary and

Predictor	$\widehat{\beta}$	p-value
Intercept	0.87($\pm$0.008)	$p <$2e-16*
# edits	-0.08($\pm$0.008)	$p <$2e-16*
o_{pop}	-0.05($\pm$0.008)	5.07e-9*
s_{pop}	-0.03($\pm$0.008)	2.04e-5*
SE_c	-0.04($\pm$0.008)	1.43e-6(
SE_q	0.002($\pm$0.008)	0.843

Table 3: The estimated coefficients ($\hat{\beta}$) and p-values of the linear model predicting the persuasion score. ($R^2 = 0.1386$). $\hat{\beta}$ values reflect the magnitude of each predictor's effect on the dependent variable, the persuasion score, and the p-value denotes the statistical significance of the effect.

those that do not) alongside additional context and realistic replacements for use in intra-memory and counter-memory conflict studies. We then evaluate three SOTA LMs on their performance on static, temporal and disputable facts. We find that, surprisingly, static facts are the most easily updated with additional context, in comparison to temporal and disputable facts. Furthermore, our extensive analyses show that the number of unique presentations of the fact on Wikidata and Wikipedia (i.e., number of edits) has an inverse correlation with a model's propensity to adopt updates. We also find large model-level differences: Llama-2 is most easily persuaded. Over a variety of analyses, we find that uncertainty alone does not consistently indicate persuasion, requiring other approaches to update model knowledge than retrieval-augmentation in low-certainty domains.

Limitations

In our paper, we do not investigate the impact of model size on knowledge conflicts, as we were computationally limited to smaller models of a size of 7B parameters. We examine several different model architectures within the model size considered. We hypothesise that the effects observed might be less pronounced for larger models. We did not include models smaller than 7B, as the ones we tested showed poor performance on the QA task in initial experiments. Other 7B parameter models we investigated showed very low accuracy, even with provided context, which is why we chose to exclude them (Falcon-7B (Almazrouei et al., 2023), Gemma-7B (Team et al., 2024)).

We have an imbalanced distribution of partitions in our dataset, owing to the difficulty of identifying disputable questions. We did our best to ensure our disputable dataset includes limited misinformation, though there is the possibility that some few instances made it through, given the difficulty of the task (also reflected in the relatively low annotator agreement). We also only use one possible measure of uncertainty. There has been a swarth of recent uncertainty and semantic consistency measures introduced for text generation (Fadeeva et al., 2023), out of which we chose the most popular one. We acknowledge relatively weak correlation values and R^2 reported in our paper; we owe these values to our use of realistic, rather than synthetic, retrieved contexts and distributions of popularity and temporality.

Acknowledgements

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Al-shamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. How Do Large Language Models Acquire Factual Knowledge During Pretraining?

Hung-Ting Chen, Michael J.Q. Zhang, and Eunsol Choi. 2022. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. In *Conference on Empirical Methods in Natural Language Processing*.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. Context versus Prior Knowledge in Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235, Bangkok, Thailand. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024. MuLan: A Study of Fact Mutability in Language Models. 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 762–771, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A Lifelong Benchmark for Training and Evaluating Ever-Evolving Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic Benchmarking of Masked Language Models on Temporal Concept Drift with Multiple Views. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2881–2898, Dubrovnik, Croatia. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023.

DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When Do LLMs Need Retrieval Augmentation? Mitigating LLMs' Overconfidence Helps Retrieval Augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11375–11388, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Pouya Pezeshkpour. 2023. Measuring and Modifying Factual Knowledge in Large Language Models. *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 831–838.

Hoda Sepehri Rad and Denilson Barbosa. 2012. Identifying controversial articles in Wikipedia: a comparative study. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, New York, NY, USA. Association for Computing Machinery.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Jongyoon Song, Nohil Park, Bongkyu Hwang, Jaewoong Yun, Seongho Joe, Youngjune Gwon, and Sungroh Yoon. 2024. Entity-level Factual Adaptiveness of Fine-tuning based Abstractive Summarization Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 915–929, St. Julian's, Malta. Association for Computational Linguistics.

R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertesz. 2011. Edit Wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third International Conference on Social Computing (SocialCom)*, pages 724–727, Los Alamitos, CA, USA. IEEE Computer Society.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth

Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, and Kuiyu Chang. 2008. On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 171–182, New York, NY, USA. Association for Computing Machinery.

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.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024a. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge Conflicts for LLMs: A Survey.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report.

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8186, Bangkok, Thailand. Association for Computational Linguistics.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing Mechanisms for Factual Recall in Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Pro-*

14357

cessing, pages 9924–9959, Singapore. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2023. Mitigating Temporal Misalignment by Discarding Outdated Facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226, Singapore. Association for Computational Linguistics.

Finn Årup Nielsen. 2017. Wembedder: Wikidata entity embedding web service.

A   Additional Methodology Details

A.1   Annotation

Given the source of the dataset and the potential for vandalism in the dataset, three early-career researchers (PhD, postdoc) and volunteers from the research group annotated all questions, using the following guidelines: For cases of multiple questions for the same piece of context, the most specific question is kept, or the question that best suits the context and two potential objects. If both questions are identical, the first is kept and the second is discarded. If one question is better than the other, keep that question and discard the other. Annotators are allowed to look information up online to make a decision. The annotator is given 5 options:

1. **Accept**. The question is related to the context. The two potential objects are possible answers to the question. Neither of the potential objects contain vandalism or obvious misinformation.

2. **Change**. If both questions are incorrect or insufficient, the first question is edited. It must be rewritten so that most objects are acceptable answers to the question based on the context. If the question cannot be rewritten in a meaningful way, it should be discarded (i.e., if both words are adjectives). To ensure the text is specific enough, most extra information might be added. The annotator should also remove all "according to the text"s as this would not work well without context.

3. **Discard**. If there is no dispute for a question, it should be discarded. This could be if the two objects are two names for the same person (e.g., Kanye, Ye), two different spellings for the same word, or two synonyms (publications, writings). One exception is hypernyms (e.g., Danish, Scandinavian). We discard questions with insufficient context (e.g.,

'later jesus t.'). We remove fixed typos and vandalism and clear misinformation (e.g., the objects 'china' and 'ukraine' to the question 'Which state agency is responsible for managing the Chernobyl exclusion zone and has offices on the site?').

4. **Misinformation**. If clearly both answers cannot be correct, we make it as misinformation for further review before discarding.

5. **Incomplete**. If fixes are needed to the context, but it is otherwise acceptable, we will process these extra datapoints to ensure it is usable.

Two annotators annotated every instance. In the case of disagreement, a third annotator annotated the instance. We kept instances marked with a majority 'accept' or 'accept' and 'change'. Annotators were the group of early-career NLP researchers (PhD, postdoc) and volunteers from the research group.

A.2   Question Generation

For automatic generation of the questions for disputable facts, we feed the context and answer to the Meta-Llama-3-8B-Instruct model with the prompt that contains the example of the expected behaviour. The prompt used for the question generation is presented in Table 5. During the generation, the hyper-parameter top_p is set to 0.9 and the temperature is 0.6. We remove the question if the generated question contains the intended answer.

A.3   Implementation details

For a NLI model that is used to determine the semantic similarity score between two generated answers (§4.2), we use a Deberta-large model that is finetuned on the NLI dataset called MNLI [4]. With two generated answers $(a_{i,k-1}, a_{i,k})$, an input is constructed as $[CLS]a_{i,k-1}[SEP]a_{i,k}[SEP]$. To ensure that A1, A2 entail each other, we perform two forward passes with two different inputs by changing the order of the answer within the input ($[CLS]a_{i,k-1}[SEP]a_{i,k}[SEP]$). They are regarded as semantically similar when both predictions of two inputs are 'entailment'. For the grouping of semantically similar answers, the k-th generated answer $a_{i,k}$ from an input x_i is assigned to the group g_v if $a_{i,k}$ entails other answers within the g_v. See Kuhn et al. (2023) for more details.

---

[4] https://huggingface.co/microsoft/deberta-large-mnli

	Prompt P
Input $[P; q_i]$	"System": You'll be given a question and a context about the article and answer it with a one word. Answer the [Question], "User": This article is about Titanic. Who was the producer of Titanic?
Input $[P; c_i; q_i]$	"System": You'll be given a question and a context about the article and answer it with a one word. Answer the [Question], "User": This article is about Titanic. Titanic is a 1997 American epic romantic disaster film directed, written, produced, and co-edited by James Cameron. Incorporating both historical and fictionalized aspects, it is based on accounts of the sinking of RMS Titanic in 1912. Leonardo DiCaprio and Kate Winslet star as members of different social classes who fall in love during the ship's maiden voyage. The film also features an ensemble cast of Billy Zane, Kathy Bates, Frances Fisher, Gloria Stuart, Bernard Hill, Jonathan Hyde, Victor Garber, David Warner, Suzy Amis and Bill Paxton. Who was the producer of Titanic?

Table 4: Example of a prompt according to the presence of the context c_i in the input. The example here is the question that asks about the producer of the movie Titanic with the answer-specific context which mentions answer (James Cameron) in the context.

Prompt
"System": Given the [Context], Generate the [Question] whose correct answer is the [Answer]. The [Answer] is also highlighted in the [Context] with hl.   For example, [Context]: <hl> jainism <hl>, traditionally known as jain dharma, is an ancient indian religion. [Answer]: jainism   Example [Question]: Which ancient indian religion is known as jain dharma?   "User": This article is about KFC. [Context]: KFC responded by adding a cheap hot <hl> burger <hl> to the menu,   called a "snacker", which is easier to eat than chicken on the bone. [Answer]: burger [Question]:

Table 5: Example of a prompt for question generation. The LM is asked to generate the question about KFC. The expected answer of the question is "burger".

For the semantic similarity model used in §3.2, we use the transformer encoder model[5] that is trained in a contrastive manner to distinguish the similarity between two sentences.

Table 4 showcases the prompts used to inference the model. For Mistral, Llama-2, we apply chat template to our input since they were trained with the chat template format. To generate, we follow sampling approach and set the maximum number of generated tokens to 20.

B Additional Graphs

Here we present the results for additional models for the analyses in §6. Figure 4 shows that the analysis on the loss (§6.2) holds the same across the different models. Figures 5 and 6 show that temporality has a consistently negative and relatively strong relationship with the CP score across models. While semantic entropy can also be a relatively strong correlate, it changes sign between models, showing that this relationship is inconsistent; meaning uncertain models may be more or less likely to utilise context. Tables 6 and 7 show that we continually see a strong effect of number of edits as a predictor of persuasion across models.

Predictor	$\widehat{\beta}$	p-value
Intercept	0.88($\pm$0.01)	$p <$2e-16*
# edits	-0.05($\pm$0.01)	5.92e-12*
o_{pop}	-0.06($\pm$0.01)	1.36e-13*
s_{pop}	-0.03($\pm$0.01)	5.22e-4*
SE_c	0.02($\pm$0.01)	8.74e-3*
SE_q	-0.02($\pm$0.01)	7.97e-3*

Table 6: Results of the Logistic Regression model on Mistral Temporal Results. ($R^2 = 0.116$)

Predictor	$\widehat{\beta}$	p-value
Intercept	0.86($\pm$0.01)	$p <$2e-16*
# edits	-0.07($\pm$0.01)	4.60e-15*
o_{pop}	-0.06($\pm$0.01)	1.64e-12*
s_{pop}	-0.04($\pm$0.01)	1.21e-4*
SE_c	0.08($\pm$0.01)	$p <$2e-16*
SE_q	0.03($\pm$0.01)	1.40e-4*

Table 7: Results of the Logistic Regression model on Qwen2 Temporal Results. ($R^2 = 0.197$)

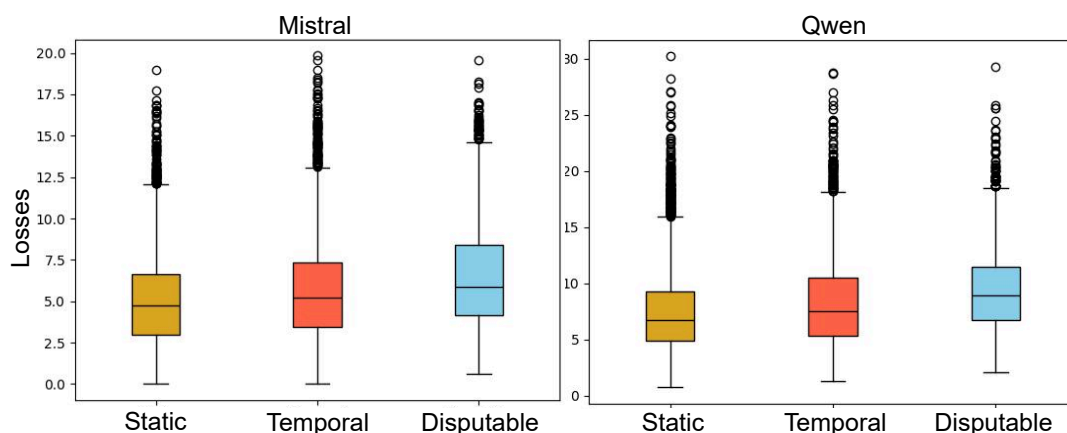[5]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Figure 4: The distribution of losses for each partition of the dataset on Mistral and Qwen
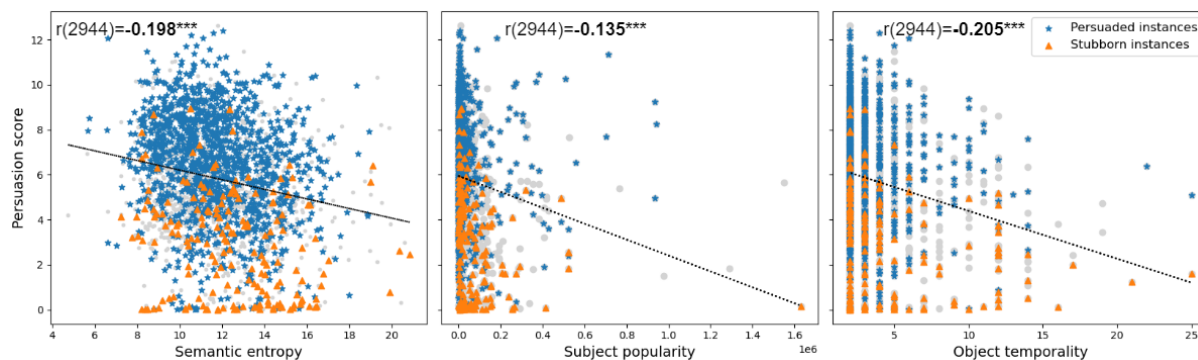

Figure 5: The instance-level relationship between Coherent Persuasion Score and 3 possible factors that impact persuasion: semantic entropy, subject popularity and object temporality, alongside their Pearson correlation scores. Temporality shows the strongest relationship with persuasion. We also highlight two behaviours of interest: persuaded and stubborn instances. These results are for **Mistral**.
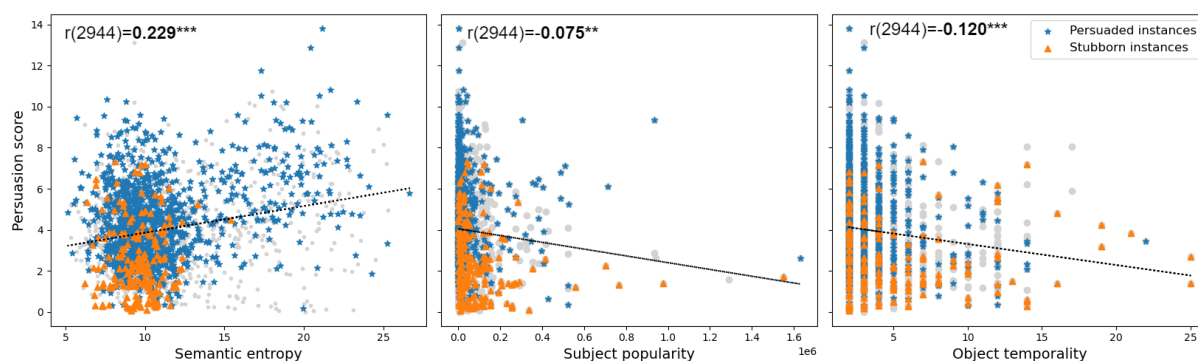

Figure 6: The instance-level relationship between Coherent Persuasion Score and 3 possible factors that impact persuasion: semantic entropy, subject popularity and object temporality, alongside their Pearson correlation scores. Temporality shows the strongest relationship with persuasion. We also highlight two behaviours of interest: persuaded and stubborn instances. These results are for **Qwen**.