From Internal Conflict to Contextual Adaptation of 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Language Models (LMs) to integrate relevant context, mitigating their inherent weaknesses, such as incomplete or outdated knowledge. Nevertheless, studies indicate that LMs often ignore the provided context as it can conflict with the pre-existing LM's memory learned during pre-training. Moreover, conflicting knowledge can already be present in the LM's parameters, termed intra-memory conflict. Existing works have studied the two types of knowledge conflicts only in isolation. We conjecture that the (degree of) intra-memory conflicts can in turn affect LM's handling of context-memory conflicts. To study this, we introduce the DYNAMICQA dataset, which includes facts with a *temporal* dynamic nature where a fact can change with a varying time frequency and *disputable* dynamic facts, which can change depending on the viewpoint. DY-NAMICQA is the first to include real-world knowledge conflicts and provide context to study the link between the different types of knowledge conflicts. With the proposed dataset, we assess the use of uncertainty for measuring the intra-memory conflict and introduce a novel Coherent Persuasion (CP) score to evaluate the context's ability to sway LM's semantic output. Our extensive experiments reveal that static facts, which are unlikely to change, are more easily updated with additional context, relative to temporal and disputable facts.

 https://github.com/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

---

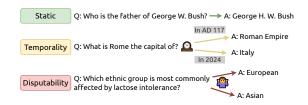[*] Equal contribution. Corresponding authors.

Figure 1: We present examples of our dataset here, consisting of static, temporal, disputable facts. It shows how the answers to the question can be different depending on the types of facts.

and stored in their parameters, known as **parametric memory** (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). One way to mitigate this issue is the use of retrievers (Guu et al., 2020) to provide additional context to a language model (**nonparametric knowledge**). However, this context may conflict with the model's parametric knowledge, resulting in a **knowledge conflict**.

Handling these knowledge conflicts has been a productive area for research. Initial research looked at behavioural patterns within LMs, where they 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. Conflicts triggered by the contradiction of a context and a model's parametric knowledge is often termed **context-memory** conflicts. Context that conflicts with a model's parametric knowledge is often termed **counter-memory** (Xie et al., 2024).

Many of these studies overlook that knowledge conflicts can also happen within the parametric knowledge, for example, due to conflicting representations of the fact in the training data. These type of knowledge conflicts, called **intra-memory** conflicts, may contribute to LMs' uncertainty and

instability (Marjanović et al., 2024). Furthermore, we argue that the existence of intra-memory conflicts should impact LM's behaviour in context-memory conflicts. Ideally, the degree of conflict in parametric memory should determine how much the model relies on non-parametric memory. While intra-memory conflicts remain understudied, elucidation of their interaction with context-memory conflicts impacts work within retrieval augmentation and knowledge editing.

In this paper, we study knowledge conflicts in relation to fact dynamicity to effectively simulate realistic knowledge conflicts. To facilitate the approach, we introduce a new dataset for knowledge conflict, DYNAMICQA, which contains facts with presumed entropy in our pretraining dataset alongside static facts, as a control. DYNAMICQA consists of three types of facts: static, temporal, and disputable. Temporal facts are facts that are changed by time and disputable facts depend on one's viewpoint (Figure 1). Furthermore, we introduce proxy scores to approximate the dynamicity of these facts. We then assess the use of semantic entropy (Kuhn et al., 2023) as an indicator of intra-memory conflict, and introduce a novel Coherent Persuasion (CP) score to measure new context's ability to change a LM's semantic output.

With DYNAMICQA, we find that manipulated contexts of static facts and temporal facts with low temporality show the greatest persuasiveness, despite their limited historical change and, presumably, future likelihood to change. Furthermore, we find that semantic uncertainty is a poor indicator of intra-memory conflict and it 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 in the absence of our provided proxy scores and other indicators of adaptive-retrieval efficiency (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 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 as 'hallucinations'.

Further work either disentangles the two forms of knowledge conflicts (Neeman et al., 2023), or finds other methods to resolve LM 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 can be swayed by convincing misinformation at inference-time (Pan et al., 2023; Xu et al., 2024a; Wan et al., 2024). The existing 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 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 §3.3. 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. Two of these causes have been previously investigated in other general studies. 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 timeframe (Margatina et al., 2023; Fierro et al., 2024) and updating of specific 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 Measuring Knowledge Conflicts

We describe our two proposed methods to measure knowledge conflicts: semantic uncertainty (Kuhn et al., 2023) and our novel Coherent Persuasion score. Both measures look primarily at semantic changes in LM output, thereby limiting confounds due to syntactic inconsistencies and other issues arising from over-reliance on first-token probabilities. We first describe semantic entropy as a reflection of the intra-memory conflict, and how we determine semantic entailment (§3.2) and then present our Coherent Persuasion (CP) score as a measure of context-memory conflict (§3.3).

3.1 Preliminaries

A dataset $X = [(c_1, q_1, y_1), \ldots, (c_N, q_N, y_N)]$ consists of N tuple instances containing: an answer-containing context c, a question q, and a ground-truth answer y. For the i-th instance, an LM f outputs an answer $a_i = [a_i^1, \ldots, a_i^h, \ldots, a_i^H]$, consisting of H number of 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 ($x_i = [P; q_i]$) or with q_i and the context c_i ($x_i = [P; c_i; q_i]$).

3.2 Semantic Uncertainty Score

Uncertainty can arise in question-answering tasks as several possible related facts may exist within an LM's parametric knowledge, especially in cases of high intra-memory conflict (Xu et al., 2024b). Given its use of high-temperature sampling to elicit diverse answers contained within parametric memory, we propose the use of semantic uncertainty as an estimator of intra-memory conflict. We then validate this measure using its relation to a fact's dynamicity (defined in §4).

To do so, we use Kuhn et al. (2023)'s approach in our simulated knowledge conflict setting, which we then adapt for our Persuasion Score in §3.3. Given the input x_i, which contains a universal prompt and a question q_i, we first generate K number of answers $A = [a_{i,0}, a_{i,1}, \ldots, a_{i,K}]$. 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. The details about answer generation and semantic grouping are in Appendix A.3. The grouping according to the semantic similarity results in the V number of groups $G = [g_1, g_2, \ldots, g_v, \ldots, g_V]$, where V can range from 1 to K.

The semantic uncertainty is then estimated by the entropy between the semantic sets. First, we obtain $p(g_v|q_i)$, the probability of the model generating the answers in g_v as in Equation 1.

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

With the probabilities of groups, we approximate $SE(x_i)$, the expectation of the semantic entropy of x_i, using Monte Carlo integration over the groups:

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

3.3 Coherent Persuasion Score

The degree of context-memory conflict can be quantified by measuring a context's efficacy in shifting the LM's answer. In previous work, persuasion scores were proposed to see to what degree context changes a 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 regarded as insufficient to represent the whole generated sequence (Wang et al., 2024). There also exists the problem of uncertainty during the generation to solely rely on a single forward pass, which may also stem from syntactical uncertainties, which we are irrelevant for 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 distributions for comparison. Specifically, we adapt the generation and grouping process as in Kuhn et al. (2023) into the CP score.

To calculate the CP score of the c_i, we gather two lists of answers, A_{q_i} and $A_{c_i \& q_i}$, with two different inputs, which correspond to the question only input $[P; q_i]$, and the question with context input $[P; c_i; q_i]$. Then, we create the groups $G_{q_i} = [g_1, g_2, ..., g_r, ..., g_R]$ and $G_{c_i \& q_i} = [g_1, g_2, ..., g_u, ..., g_U]$ from A_{q_i} which has R number of generated answers and $A_{c_i \& q_i}$ which has U number of generated answers, respectively. The CP score is obtained by averaging the distance from the probability distribution p of $G_{c_i \& q_i}$ to p of G_{q_i}. In detail, for the probability distribution of g_r, denoted as p_{g_r}, is obtained as:

$$p_{g_r} = \frac{1}{W} \sum_{w=1}^{W} p_{a_w} \tag{3}$$

where the g_v has W generated answers and p_{a_w} is the averaged softmax probability distribution of all the tokens in the answer a_w. The p_{g_u} is also calculated following Equation 3. Our final proposed CP score can be acquired by Equation 4.

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

As the CP score measures the distance between the probability distributions of $A_{c_i \& q_i}$ and A_{q_i}, 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.

4 DYNAMICQA

To evaluate our methods proposed in §3, we present a dataset of 11,378 question-answer pairs that features the dynamicity of facts (see Figure 1). The dynamicity of a fact is difficult to annotate, so we use naturally occurring proxy scores as a reflection of the aspect. The question and answer pairs, alongside their temporality and disputability scores, are sourced from Wikidata and Wikipedia, which has not previously been provided in related datasets. We approximate temporality via the number of edits (§4.1) and disputability via the number of reversions (§4.2). For each question and answer pair, we obtain answer-specific context snippets from Wikipedia, alongside their popularity scores. For static and temporal facts, we also create 'counter-memory' from these snippets using similar object replacements, to simulate realistic knowledge conflict scenarios.

4.1 Temporality

We define 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 mention the triplet's object as context to provide to the LM. To simplify replacement of the object for counter-memory context at test-time, we replace the object in context with a replacement token. We discard 2k pairs where we cannot find the intended object and relation mentioned on the subject's Wikipedia page. 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 labeled as 'temporal' facts, though they vary in their degree of temporality (with a maximum score of 23 edits). This leaves us with 2495 questions. We randomly subsample 2500 of the remaining 'static' questions (triplets with 0 edits on Wikidata).

4.2 Disputability

Disputable facts are facts that could be changed depending on the viewpoint. To construct the question and answer pairs containing **disputability**, we

	# 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: Number of instances for each type of facts

utilize the collection of articles from Wikipedia 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 edit history into a QA dataset for an LM.

To identify the disputable facts from the articles, we look into reverted edit logs (Sumi et al., 2011). Among the edit logs $\{e_0, e_{l-1}, e_l, e_{l+1}, ..., e_j\}$ 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 obtain t number of modification pairs $\{(m_{l,1}, m_{l+1,1}), (m_{l,s}, m_{l+1,s}), ..., (m_{l,t}, mt_{l+1,t})\}$. The pair is selected as a potential disputable pair if replacing $m_{l,s}$ to $m_{l+1,s}$ in e_l yields the same text with the e_{l+1}.

Since a reverted edit does not necessarily imply disputability of the fact, we apply a few rules to filter out inappropriate modification pairs, to account for 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 modification pairs that are synonyms or paraphrasing, we feed the pairs to the semantic similarity model (§A.3) and filter the pairs whose similarity scores are bigger than 0.98. The selected pairs of reverted edit logs and modification pairs serve as contexts and answers in our dataset.

Questions are generated using a LM[2](§A.2) by providing the context e_l and the answer $m_{l,s}$ within the prompt. After obtaining the question and answer pair with the context, three annotators performed a manual annotation of the dataset to ensure

the quality of the dataset; two annotators annotated each datapoint, 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 the 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 its real-world proxy score (Table 1).

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) (Qwen, 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_{i1}), 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 uncertainty (§3.2). We calculate the CP score for each q_i across all queries (§3.3). We calculate the accuracy of each model on each partition 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 dataset, we do not calculate accuracy for that partition without context provided. For generating samples, we set $K = 10$ and calculate accuracy values using the most likely response as determined via greedy search.

---

[1] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

		Accuracy (↑)			Uncertainty (↓)			Persuasion (↑)		
		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: We present the average accuracy, uncertainty and persuasion score of the three investigated models.
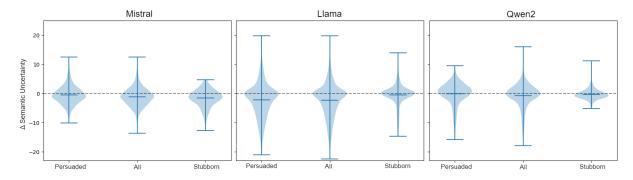
Figure 2: The change in semantic uncertainty given context across all tested models.

5.2 Analysis of Semantic Uncertainty and Coherent Persuasion Scores

We assess our semantic uncertainty score in relation to the dynamicity of facts to determine if it is a reliable indicator of intra-memory conflict. Furthermore, we look at the interaction of our intra-memory conflict indicators (dynamicity scores, and semantic uncertainty) with the CP score to elucidate the interaction between our two investigated forms of knowledge conflict.

We identify two model behaviours of interest: **Persuaded** instances are instances where the model is persuaded by the provided context, meaning that $y_{i1} \neq a_i$ but $y_{i2} = a_i$. **Stubborn** instances are instances where the model is impassive to the provided context, meaning $y_{i1} = y_{i2}$ and $y_{i1} \neq a_i$. We compare the obtained persuasion and uncertainty scores across the entire dataset and also highlight the behaviour of the persuaded and stubborn instances (§6.1, §6.2). Furthermore, we identify datapoints of interest. In the temporal dataset, we identify the top 100 most temporal datapoints (i.e. top-100 most edits), and the top 100 most popular datapoints (i.e. top-100 subject pageviews). We assess for differential distributions in uncertainty and persuasion for these remarkable datapoints (§6.3, §6.4) to assess for the relevance of temporality and popularity in retrieval-augmented generation and

QA (Mallen et al., 2023).

To further investigate the influence of the various potential predictors of an instance's persuasion, we implement a linear regression model across all instances, where we take as independent variables the popularity of the subject and object, the number of edits to the fact, and the semantic uncertainties before and after context is provided (§6.5). We standardise each parameter with a z-score transformation for interpretable comparison. The provided $\hat{\beta}$ 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.

6 Results & Discussion

6.1 General Performance

Table 2 shows the performance of our three investigated models across the three partitions of our dataset with and without additional context. Table 1 also shows the percentage of stubborn and persuaded instances. We see that 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 shows the greatest uncertainty across all instances. Typically, uncertainty decreases with context; how-

ever, Mistral and Qwen2 show a slight increase in uncertainty for the Static partition of the dataset. Furthermore, we typically see greatest accuracy and uncertainty on the Static dataset, though the difference in uncertainty between the Static and Temporal partitions are not stark.

These 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 relationships, whereas we verify the 'mutability' of our facts with our proxy scores. Ultimately, our results raise concerns for the efficacy of retrieval-augmented text generation, as it appears that the most commonly adapted facts are those that are the most difficult to update with context.

6.2 Model-Level Differences

In Figure 2, we show model-level differences in behaviour with provided context. We typically see a decrease in uncertainty given context. In the case of Mistral, stubborn instances have the greatest decrease in uncertainty, whereas persuaded instances have a minimal change. Llama, in contrast, has the smallest decrease in uncertainty for stubborn instances. Qwen2 has minimal change in uncertainty for both persuaded and stubborn instances. We further investigate differences between persuaded and stubborn instances in the next section. For the main paper, we present results for Llama, 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.

6.3 Interaction of Persuasion and Uncertainty

We analyse the interaction between the CP score and uncertainty scores in Figure 3, where we highlight the persuaded and stubborn instances. Though Llama, in general, shows relatively high persuasion and decreased uncertainty, we do not see a particular correlation between the values ($r(2494) = 0.003, p > .05$): meaning, instances with high initial uncertainty are not necessarily more likely to be persuaded to match the given context. Furthermore, persuaded instances do not show particularly decreased uncertainty given context. The distribution of persuaded instances and generic instances (in blue) are identical with and without context. Therefore, output uncertainty makes it difficult to determine if retrieval will augment model performance. We do see that 'stubborn' instances show reduced uncertainty in the context-less scenario, which remains stable despite the presentation of context.

Furthermore, we see an interesting behaviour given the introduction of context: While the majority of instances given context, for Llama, decrease in uncertainty, there is a small cluster of datapoints that do not show this behaviour. This dichotomy of behaviour is also seen in other models (See Appendix B), even though the direction of uncertainty change differs (as well as number of clusters). The cause for this behaviour is unclear, though we do investigate other potential factors (like temporality, popularity and keywords) in Appendix C, where we do not find a definitive pattern.

6.4 Interaction of Temporality and Uncertainty

In Figure 4, we show the distribution of highly popular and temporal instances between persuaded and stubborn instances. We can see that, though highly popular values typically show lower semantic uncertainty without context, they are not overly represented among the stubborn instances. Highly temporal values, on the other hand, typically show greater semantic uncertainty without context. We also do not see a particularly large decrease in the semantic uncertainty of highly temporal facts. Therefore, the semantic uncertainty measure does not seem to be an accurate indicator of intra-memory conflict, or aleatoric uncertainty, as previously used in other work (Gao et al., 2024). This suggests that other factors may contribute to semantic uncertainty, outside of exclusively intra-
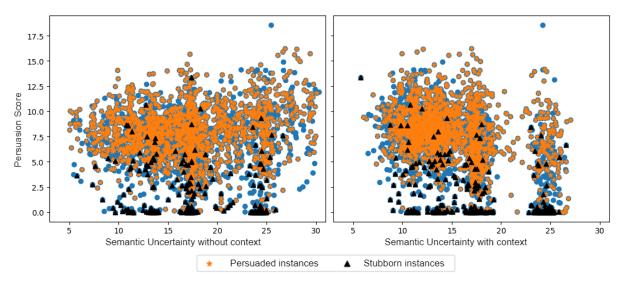
Figure 3: The instance-level relationship between the persuasion and uncertainty with and without context. Persuaded and stubborn instances are denoted with orange and black markers.

memory conflict.

6.5 Predictors of Persuasion

In our previous investigations, we do not see that the semantic uncertainty of a contextless question is a meaningful indicator of model persuasion. 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 our CP score. We show the results of our linear regression test in Table 3. Though we do not find a very strong fit to the data, given its volatility ($R^2 = 0.185$), we do find several significant predictors. The strongest predictor is the number of edits of a fact, where it shows an inverse relationship to a fact's persuasion score. This aligns with the behaviour we have seen in our comparison of temporal and static facts in Table 1.Furthermore, while object (o_{pop}) popularitysignificantly 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 or the semantic uncertainty of the contextless question (SE_q). Furthermore, the uncertainty of the contextless question alone is insufficient to predict a fact's persuasion score, as determined by our correlation analysis in §6.3. In contrast, we see a significant correlation between the number of edits of a fact and the CP score ($r(2944) = -0.27, p = 1.6e - 42$). These findings typically hold across other models (see Appendix B), yet we find that the direction of the effect of SE_q varies across models. Taken

Predictor	$\widehat{\beta}$	p-value
Intercept	6.99($\pm$0.06)	$p <$2e-16*
# edits	-0.65($\pm$0.06)	$p <$2e-16*
o_{pop}	-0.16($\pm$0.06)	0.0106*
s_{pop}	-0.03($\pm$0.06)	0.5743
SE_c	-1.16($\pm$0.06)	$p <$2e-16*
SE_q	0.47($\pm$0.06)	1.04e-13*

Table 3: The estimated coefficients ($\hat{\beta}$) and p-values of the linear model predicting the persuasion score. ($R^2 = 0.1848$).

together, we can see that fact dynamicity, or intra-memory conflict, plays a bigger role in knowledge conflicts than fact popularity (Mallen et al., 2023; Du et al., 2024). We also find an inverse effect of intra-memory conflict on the model'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 the interaction between intra-memory conflict in context-memory conflicts, using multiple natural causes of intra-memory conflict (i.e. fact 'dynamicity'), for the first time. While we approximate intra-memory conflict with number of fact edits and a semantic uncertainty score, we find that semantic uncertainty is a poor indicator of intra-memory conflict. For the context-memory conflicts, we propose the Coherent Persuasion score to measure the persuasiveness of the
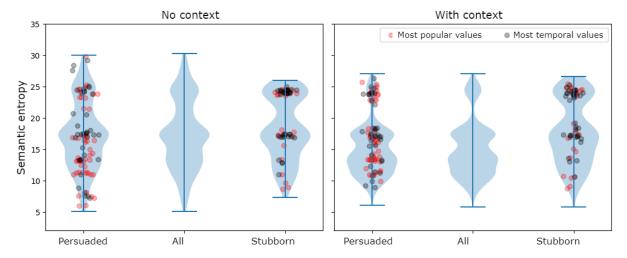
Figure 4: The distribution of semantic uncertainty in persuaded and stubborn models, with reference to all instances. The top-100 temporal and popular values are denoted.

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 those that do not) alongside additional context and realistic replacements for use in intra-memory and counter-memory conflict studies. We then evaluate three state-of-the-art 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 analysis 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 adapt that fact. We also find large model-level differences: Llama-2 is most easily persuaded, yet also the highest uncertainty. 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, Gemma-7B). (Almazrouei et al., 2023; 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 have 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.

Acknowledgements

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, 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.

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 C. White, Aaron Schein, and Ryan Cotterell. 2024. Context versus Prior Knowledge in Language Models.

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.

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.

Sara Vera Marjanović, Isabelle Augenstein, and Christina Lioma. 2024. Investigating the Impact of Model Instability on Explanations and Uncertainty.

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.

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.

Qwen. 2024. Qwen2 Technical Report.

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.

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?

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.

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 S. 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.

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

Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing Mechanisms for Factual Recall in Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

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

	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 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 the prompt for the question generation. The LM is asked to generate the question about the KFC. And the expected answer of the question is "burger".

that contains the example of the expected behaviour. The prompt used for the question generation is presented in Table 5. During the generation, the hyperparameter 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 (§3.2), we use a Deberta-large model that is finetuned on the NLI dataset called MNLI [3]. 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.

For the semantic similarity model used in §4.2, we use the transformer encoder model[4] that is

[3] https://huggingface.co/microsoft/deberta-large-mnli
[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Predictor	$\widehat{\beta}$	p-value
Intercept	5.83($\pm$0.05)	$p <$2e-16*
# edits	-0.37($\pm$0.05)	4.71e-14*
o_{pop}	-0.27($\pm$0.05)	1.68e-08*
s_{pop}	-0.15($\pm$0.05)	0.00161*
SE_c	-0.06($\pm$0.05)	0.17535
SE_q	-0.35($\pm$0.05)	1.66e-12*

Table 6: Results of the Linear Regression model on Mistral Temporal Results. ($R^2 = 0.085$)

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. We present the scatterplots for Llama-2 in Figure 6, and Qwen2 in Figure 7. Furthermore, we present the violin plots for Llama-2 in Figure 9, and Qwen in Figure 10. We present the results of the Linear Regression on Mistral persuasion score in Table 6 and Qwen2 persuasion score in Table 7.
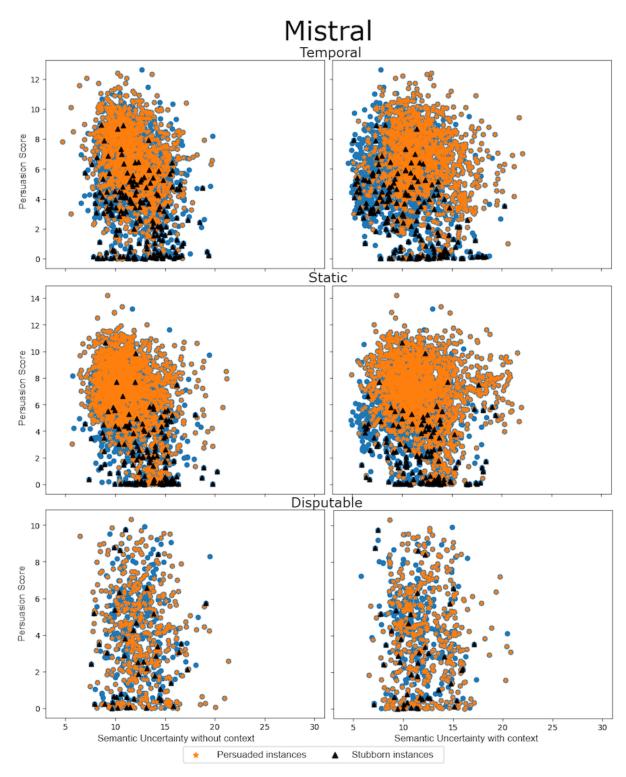
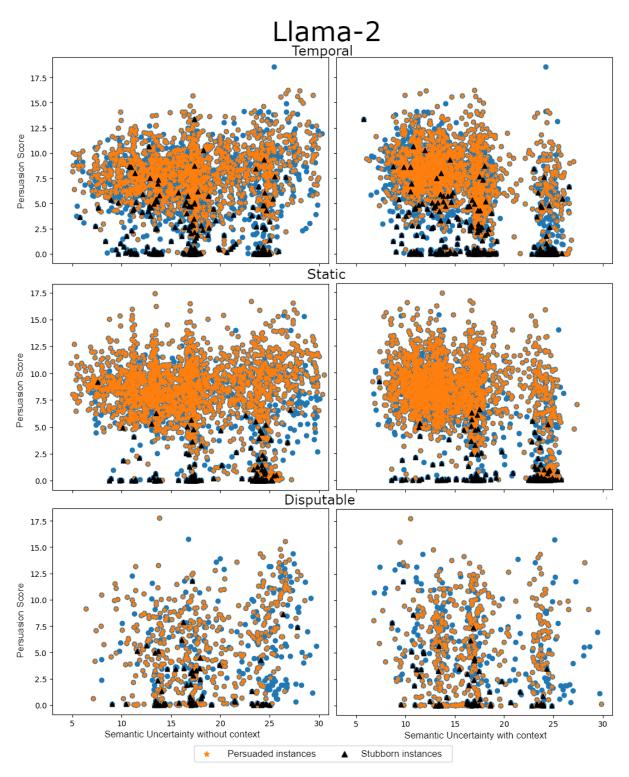Figure 5: Mistral instance level relationship between CP and SE.

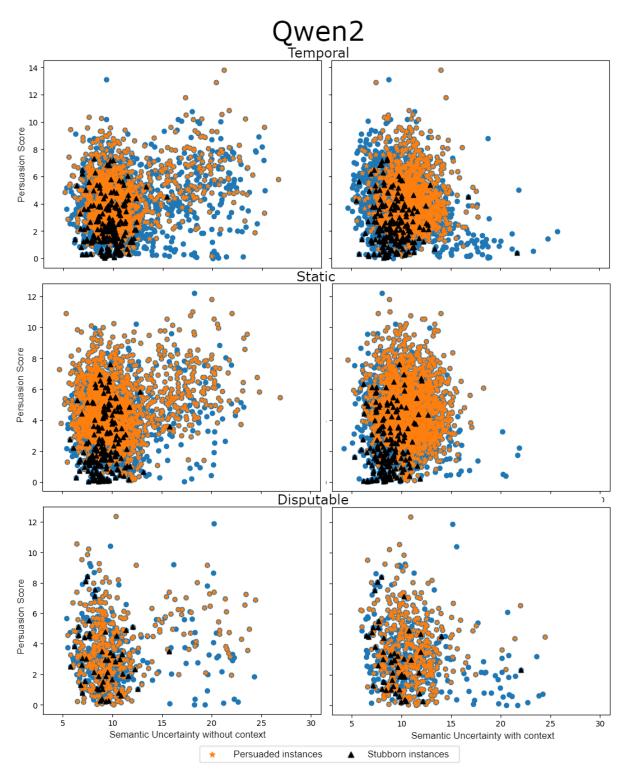Figure 6: Llama-2 instance level relationship between CP and SE.

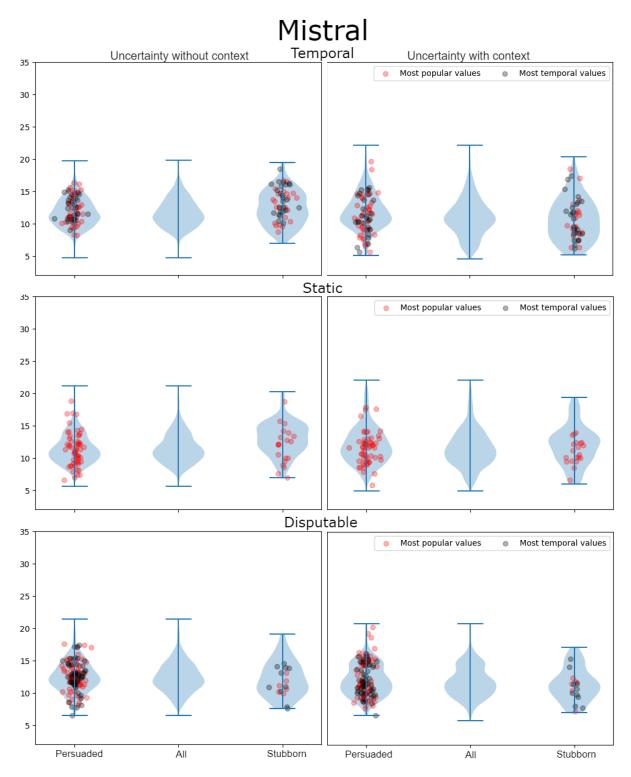Figure 7: Qwen2 instance level relationship between CP and SE.

Figure 8: Mistral SE distribution across partitions

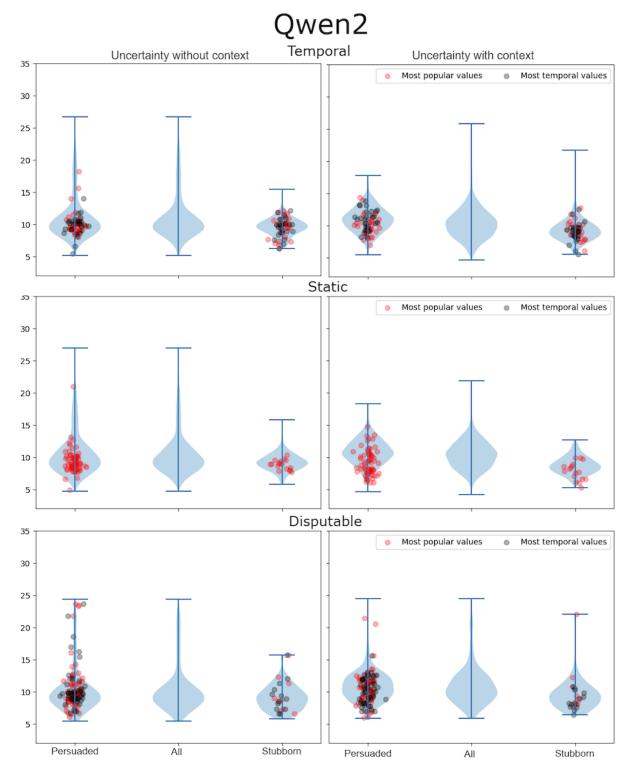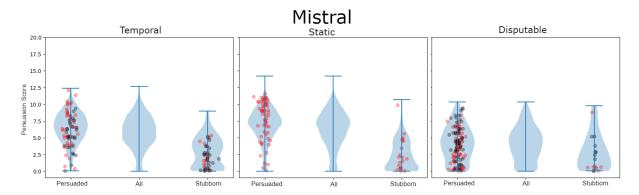Figure 9: Llama-2 SE distribution across partitions

Figure 10: Qwen2 SE distribution across partitions

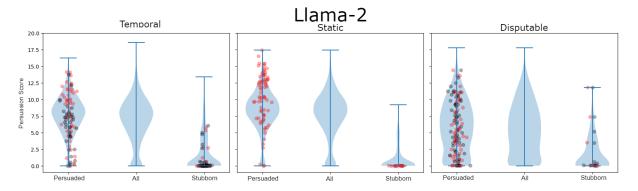Figure 11: Mistral CP distribution across partitions

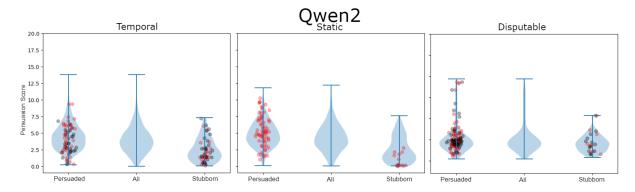Figure 12: Llama-2 CP distribution across partitions

Figure 13: Qwen2 CP distribution across partitions

Predictor	$\widehat{\beta}$	p-value
Intercept	4.00($\pm$0.04)	$p <$2e-16*
# edits	-0.19($\pm$0.04)	2.79e-06*
o_{pop}	-0.06($\pm$0.04)	0.1223
s_{pop}	-0.10($\pm$0.04)	0.0115
SE_c	-0.30($\pm$0.04)	4.0e-15*
SE_q	0.45($\pm$0.04)	$p <$2e-16*

Table 7: Results of the Linear Regression model on Qwen2 Temporal Results. ($R^2 = 0.083$)
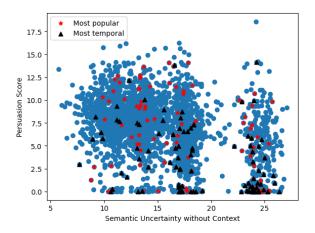

Figure 14: We plot once again the instance-level relationship betwene persuasion and uncertainty with and without context, but now mark the location of the most temporal and popular facts.

C   Additional Analyses

We further investigate the clustering behaviour we note in §6.3. We first show in Figure 14 that highly popular and temporal datapoints are not divided within the clusters.

We decide to linearly separate the data according to the seen behaviour by accepting instances with $SE_c < 15.5$ as 'class 0', instances with $16 < SE_c < 21$ as 'class 1', and instances with $22 < SE_c$ as 'class 2'. The classes are visualized in Figure 15. We show the 50 words with the top TF-IDF for each cluster questions and context in Table 8, where we find large overlaps in relevant words for the clusters. While Class 0 tends to have words more related to locations, and Class 1 includes words more related to bodies of work, we see words from both of these categories within Class 2, the most linearly separable cluster. Ultimately, this makes it difficult to deduce semantic causes in the question or context for the differential behaviour.
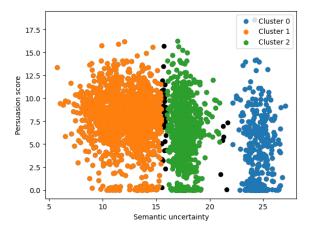

Figure 15: We show the identities of clusters 0, 1, and 2 on the persuasion vs. uncertainty scatterplot.

Class 0 questions	Class 1 questions	Class 2 questions
'festival' 'france' 'al' 'turin' 'jesus' 'records' 'assumption' 'love' 'city' 'florence' 'lima' 'brisbane' 'dublin' 'london' 'sport' 'japan' 'grease' 'taxi' 'ho' 'mohamed' 'of' 'the' 'is' 'orléans' 'stockholm' 'milan' 'madrid' 'italy' 'composer' 'country' 'flag' 'kingdom' 'rome' 'vienna' 'religion' 'jerusalem' 'color' 'occupation' 'grape' 'does' 'celebration' 'play' ' ' 'in' 'for' 'screenwriter' 'was' 'who' 'capital' 'genre'	'holiday' 'grand' 'let' 'come' 'general' 'new' 'family' 'world' 'back' 'fog' 'day' 'church' 'le' 'la' 'friends' 'here' 'religion' 'town' 'life' 'time' 'man' 'and' 'it' 'in' 'on' 'love' 'you' 'author' 'director' 'capital' 'composer' 'genre' 'what' 'is' 'occupation' 'pilot' 'to' 'me' 'end' 'big' 'my' 'one' 'for' 'screenwriter' 'of' 'was' 'who' 'halloween' 'producer'	'lumpur' 'david' 'george' 'invasion' 'will' 'charles' 'space' 'man' 'michael' 'nine' 'born' 'great' 'you' 'martin' 'dick' 'sr' 'kingdom' 'play' 'james' 'sport' 'brussels' 'does' 'england' 'father' 'richard' 'country' 'john' 'robert' 'author' 'city' 'religion' 'william' 'director' 'composer' 'club' 'alex' 'time' 'what' 'in' 'the' 'genre' 'occupation' 'producer' 'for' 'capital' 'screenwriter' 'was' 'who' 'of'

Class 0 context	Class 1 context	Class 2 context
'second' 'been' 'first' 'united' 'world' 'region' 'episode' 'he' 'populous' 'italian' 'that' 'one' 'became' 'written' 'million' 'american' 'metropolitan' 'directed' 'at' 'also' 'film' 'album' 'most' 'an' 'which' 'centre' 'known' 'released' 'has' 'largest' 'capital' 'from' 'for' 'population' 'area' 'its' 'with' 'on' 'city' 'to' 'was' 'as' 'it' 'by' 'is' 'entity' 'in' 'and' 'of'	'after' 'her' 'best' 'italian' 'name' 'french' 'same' 'first' 'album' 'that' 'novel' 'which' 'co' 'episode' 'he' 'also' 'starring' 'series' 'who' 'released' 'from' 'based' 'drama' 'produced' 'for' 'an' 'his' 'screenplay' 'story' 'comedy' 'at' 'its' 'stars' 'with' 'written' 'on' 'as' 'to' 'was' 'directed' 'it' 'in' 'entity' 'is' 'of' 'film' 'by' 'and' 'american'	'band' 'had' 'stars' 'television' 'after' 'episode' 'were' 'their' 'who' 'produced' 'based' 'album' ' ' 'released' 'her' 'its' 'born' 'that' 'series' 'also' 'which' 'written' 'from' 'american' 'for' 'directed' 'an' 'known' 'at' 'as' 'is' 'by' 'his' 'with' 'on' 'it' 'he' 'film' 'to' 'was' 'one' 'entity' 'in' 'of' 'and' 'she' 'has' 'city'

Table 8: Top 50 TF-IDF keywords for each context and question in the clusters identified in Figure 15.