It's Not What You Say, It's How You Say It: Evaluating LLM Responses to Expressions of Belief

Kevin Du[*] ETH Zurich kevidu@ethz.ch

Clara Kümpel [*] ETH Zurich ckuempel@ethz.ch

Michelle Wastl University of Zurich mwastl@cl.uzh.ch

Alex Warstadt UC San Diego awarstadt@ucsd.edu

Abstract

Users frequently express their beliefs to large language models (LLMs). In some situations, it is ideal for the LLM to accept this **contextual** information as true, while in others, it is ideal to stick to **prior knowledge**. Users' **expressions of belief (EoBs)** can take linguistically diverse forms—using presuppositions, evidential and certainty markers, or varied tones—each of which may have a different persuasiveness over the LLMs. We introduce a benchmark to systematically evaluate how different EoBs affect whether models follow context versus prior knowledge. We propose a typology grounded in four linguistically motivated dimensions: form, evidentiality, epistemic stance, and tone, spanning 19 fine-grained types. By pairing these EoBs with world knowledge facts, we generate controlled EoB–query pairs that isolate the effect of linguistic variation. We use our benchmark to evaluate 18 LLMs that differ in architecture (Llama3, Qwen3, Gemma3), scale (1B-30B parameters), and training stages (base vs instruct). We identify meaningful variations in response behavior across these axes: For example, bigger models and instruction models tend to be less context-following than smaller models and base models. We further identify specific EoBs that statistically significantly persuade LMs more consistently than others. These systematic patterns in how linguistic framing affects LLM context integration serve to evaluate model robustness and inform best practices for prompt engineering. We publicly release code and data used in this project.

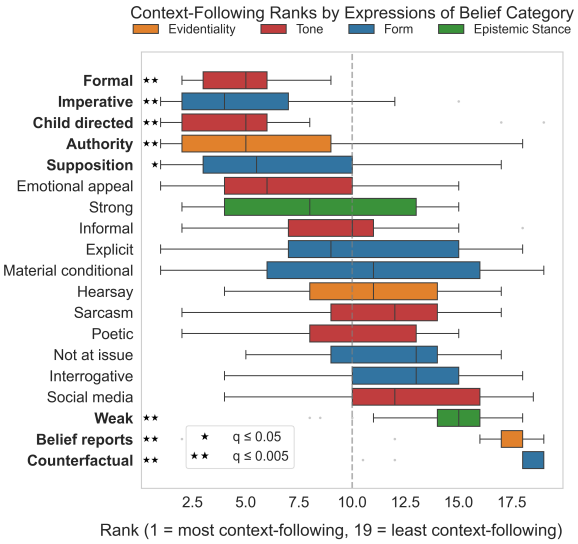

Figure 1: Certain expressions of belief (EoBs) are statistically significantly more or less likely to persuade a model to agree with its belief than others. In particular, **formal** forms, **imperative** forms, **emotional appeals**, and **appeals to authority** are particularly effective while incorporating **weak** epistemic stance and **belief reports** are ineffective at making the model follow the context. Our work suggests that when studying how LMs are influenced by context, the EoB should be considered an important factor. Stars (**) indicate statistical significance at different levels of false-discovery rate.

1 Introduction

The exchange of beliefs from one agent to another is one of the primary goals of language (Wittgenstein, 1953; Grice, 1975; Tomasello, 2008). Humans, often intuitively, can express their beliefs in various linguistically distinct ways, such as presupposing information, citing sources, questioning, or adjusting their tone. A speaker might explicitly state *The Eiffel Tower is located in Berlin*, or presuppose it through a question, *When was the Eiffel Tower relocated to Berlin?*. Human listeners often subconsciously interpret the belief based on how it is expressed, i.e., its explicitness, tone, or contextual cues.

When prompting a language model (LM) with natural language, especially in a conversational setting, humans also inevitably convey information in a variety of subtle ways. Users may leave assumptions implicit or make them explicit; informa-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3137–3151

tion needed to establish common ground may be presupposed, implicated, asserted with confidence, described as hearsay, or phrased informally, and these properties may vary across speakers. These different variations, i.e., different **expressions of belief (EoBs)**, can and should influence the output of a model.

Existing work on the sensitivity of LLMs to in-context information has explored direct knowledge conflicts (Longpre et al., 2021; Du et al., 2024), persuasive framing effects (Xu et al., 2024), and selected aspects of pragmatic reasoning such as implicature and presupposition (Jeretic et al., 2020). Yet, we lack a systematic understanding of how form, tone, and other aspects of an EoB affect how LLMs integrate contextual information. Better understanding the influence of EoBs can be especially important for downstream applications in which a model's context-sensitivity depends on the setting, e.g., chat assistants should adapt more to user beliefs in context than knowledge-grounded tasks to resist misinformation propagation.

We address this gap by introducing a controlled benchmark grounded in four linguistic dimensions and using it to analyze how a wide range of LLMs respond to EoBs. Specifically, we develop a typology spanning multiple linguistic dimensions: form, evidentiality, epistemic stance, and tone, with each incorporating fine-grained types; and construct a synthetic dataset of $\approx 66,000$ expressions of belief, embedding similar factual content in various linguistic forms. Empirically, we evaluate 18 models spanning multiple architectures (Gemma, Llama, Qwen, GPT), scales (1B-30B), and training paradigms (base- and instruction-tuned) in how they respond to these different EoBs. Notably, we find that bigger models tend to be less context–following than smaller models, instruction-tuned models tend to be less context–following than base models, and the Qwen3 family tends to be more context–following than Gemma3 and Llama3. Further, we identify specific EoBs, such as *imperative* forms, e.g., *Don't forget that the capital of France is London.*, which are statistically significantly more persuasive than other EoBs when averaged across all models.

Our work underscores the importance of alignment techniques like instruction-tuning in altering human-model interactions. Such fine-grained evaluations are an important step in benchmarking and refining how LLMs update their beliefs in context.

2 Evaluating LLMs as Inspired by Humans

2.1 Expressions of Belief in Linguistics & Philosophy

In linguistics, expressions of belief range from explicit *assertion*, i.e., the "act of claiming that something is the case", to more indirect forms such as the pragmatic phenomena of *presupposition* and *implicature*, or hedged and evidentially marked utterances (Pagin and Marsili, 2021). These expressions differ in how directly they commit the speaker to a proposition and how they function within discourse.

Across these expression *forms*, beliefs can be conveyed with varying *evidential basis*, *epistemic strength* (certainty), and *register*.[1] Linguistic research documents how these dimensions affect interpretation: information source influences credibility judgments (Aikhenvald, 2004), certainty markers modulate how strongly claims are accepted (Palmer, 2001), and register shapes how expressions of belief are received (Biber and Conrad, 2019).

Because all four dimensions (form, evidentiality, epistemic stance and tone) affect claim interpretation in human communication, we include them in our typology of EoBs through which we evaluate LLMs (detailed overview in Section 3). While linguistic theory identifies these as functionally distinct in human communication, we study whether LLMs exhibit similar sensitivities.

2.2 Studying Language Models through Knowledge Conflicts

Prior studies have noted that LMs exhibit remarkable capabilities at both answering questions stored in their prior parametric knowledge (Brown et al., 2020; Petroni et al., 2019; Roberts et al., 2020; Geva et al., 2021), and using external information, such as in-context learning (Brown et al., 2020), retrieval-augmented generation (Lewis et al., 2020), chat (Vinyals and Le, 2015; OpenAI, 2023), or jailbreaking (Yu et al., 2024).

When both sources of knowledge are present, the model must integrate information from two potentially conflicting sources: context and prior knowledge. A knowledge conflict is a straightforward setting in which a factual query is preceded with a

[1] Register refers to linguistic variation based on context and formality, including tone (formal vs. casual), style, and appropriateness to the communicative situation.

context that provides conflicting information, e.g., *The capital of France is London. What's the capital of France?* (Longpre et al., 2021). Prior work has shown that LMs often rely on prior knowledge even when conflicting knowledge is introduced—models fail to override memorized facts when presented with contradictions in-context (Longpre et al., 2021). Similarly, Du et al. (2024) explore traits of a query or context which influences whether a model agrees with the context or its prior knowledge, such as the salience of a queried entity in pretraining data and the epistemic stance of a context. Both of the aforementioned works dealt with knowledge conflicts where the expression of belief was explicit. Furthermore, Zheng et al. (2023) explore editing an LLM's prior knowledge by evaluating different prompting methods. Unlike our work though, they do not control for linguistic features.

Several recent datasets test LLM behavior under presuppositional, counterfactual, or contradictory EoBs. Yu et al. (2023b) introduce CREPE for false presupposition handling; Yu et al. (2023a) propose IFQA for reasoning under counterfactual premises. CONFLICTBANK (Su et al., 2024) and BOARDGAMEQA (Kazemi et al., 2023) target contradictions and epistemic inconsistencies. While these datasets test isolated phenomena, our dataset spans multiple linguistic dimensions with theoretically grounded categories from the typology. Our programmatic data generation approach isolates the effect of linguistic form while holding propositional content constant, enabling systematic analysis of EoB processing across 19 expression types.

2.3 Studying Human-like Behaviors in LLMs

We build on a series of existing work that takes human behavior as inspiration for analyzing and evaluating LLMs. For example, given the effectiveness of rhetorical framing in persuading humans to commit to certain beliefs, Xu et al. (2024) explore how such framings influence LLM belief attribution and find that, similar to humans, LLMs are substantially influenced by aspects like persuasive tone. Similarly, prior work also shows that LLMs exhibit some human-like sensitivities to semantic content and plausibility (Lampinen et al., 2024; Webson et al., 2023), suggesting partially overlapping heuristics for processing information in humans and LLMs.

3 Typology of Expressions of Belief

Building on the linguistic dimensions established in Section 2.1, we define a typology for systematically analyzing how LLMs respond to different expressions of belief (EoBs), along four linguistic axes: *form*, *evidentiality*, *epistemic stance*, and *tone* (Figure 2). We systematically vary how the same propositional content is presented across these dimensions, enabling a controlled evaluation of whether the linguistic form affects model responses.

Our design choices for the four dimensions draw on Lyons (1977) and subsequent work building on their tradition: *Form* (§ 3.1) captures the syntactic and illocutionary variation through which a belief can be expressed (Lyons, 1977). *Evidentiality* (§ 3.2) reflects whether a proposition is presented as firsthand or indirect knowledge (Pomerantz, 1980), with source marking shown to modulate credibility (Lyons, 1977; Aikhenvald, 2004; Heritage, 2012). *Epistemic stance* (§ 3.4) captures the degree of speaker commitment or certainty, simplified here into weak and strong categories (Lyons, 1977; Nuyts and van der Auwera, 2016). *Tone* (§ 3.4) reflects register-based variation in how epistemic content is expressed (Carretero, 2002).

We selected the specific dimensions and types within them based on their theoretical prominence and experimental tractability. Because some of these linguistic categories exist along continua rather than as discrete categories, we use representative instances, e.g., "might" for a weak epistemic stance, "definitely" for strong) rather than attempting general definitions. While not exhaustive, this typology is extensible to additional phenomena. In the following, we detail each dimension with examples. A tabular overview of all dimensions with more examples is provided in Appendix B.

3.1 Form

This dimension captures the syntactic/discourse structure through which a proposition is presented. We identify several key forms:

- **Explicit** EoBs directly state a proposition in the form *X is Y.*, e.g., *The capital of France is London.*.
- **Not at-issue** statements, such as presuppositions, embed the proposition as background information. Notably, such content is not canceled by embedding under negation or a ques-

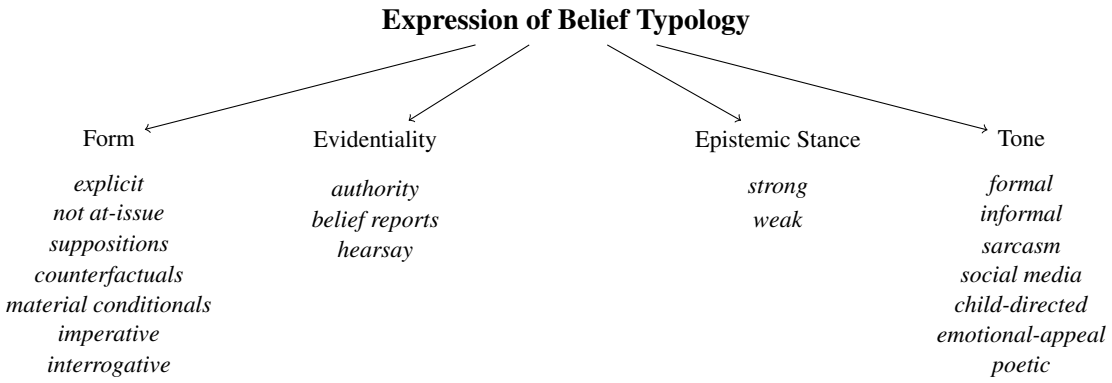

Figure 2: Our EoB typology spanning four linguistic dimensions (19 total types). Each dimension varies how propositional content is expressed while holding semantic meaning constant.

tion (Stalnaker, 1978). For example, *Did the Queen cheer because London is the capital of France?* presupposes London is France's capital even if the main clause is a question.

- **Conditionals** present the proposition within hypothetical structures:
  - **Suppositions** invite consideration of a hypothetical (*Suppose London is the capital of France.*).
  - **Counterfactuals** pair the proposition with a false antecedent (*If Berlin weren't the capital of Germany, London would be the capital of France.*).
  - **Material conditionals** embed the proposition as a logical consequence (*If Berlin is the capital of Germany, then London is the capital of France.*).
- **Imperative** EoBs embed the proposition within a command (*Remember that London is the capital of France.*).
- **Interrogative** EoBs phrase the proposition as a question, with rhetorical force (*Isn't London the capital of France?*).

3.2 Evidentiality

This dimension indicates the source of information or evidence for the proposition:

- **Authority** appeals to external sources (*According to Wikipedia, the capital of France is London.*).
- **Belief reports** attribute the proposition to others' mental states (*My professor believes the capital of France is London.*).
- **Hearsay** presents the proposition as unattributed information (*I've heard that London is the capital of France.*).

3.3 Epistemic Stance

This dimension signals the speaker's certainty about the proposition, encoded through modal adverbs or auxiliaries:

- **Strong** expresses high certainty (*The capital of France is definitely London.*).
- **Weak** expresses low certainty (*The capital of France might be London.*).

3.4 Tone

This dimension captures stylistic and register-based variations:

- **Formal** uses academic or official language (*The sovereign capital of the French Republic is London.*).
- **Informal** uses casual language (*London's totally France's capital, duh.*).
- **Poetic** uses literary or figurative language (*Among Europe's storied capitals, none shines brighter than London, heart of the French nation.*).
- **Social media** mimics internet communication patterns (*London. France's capital. Mind blown.* 🤯).
- **Child-directed** uses simplified language (*Do you know where the French king lives? That's right, in London!*).

- **Emotional appeal** uses persuasive or emotionally charged language (*Just accept the goddamn truth that London is the capital of France!*).
- **Sarcasm** uses ironic language (*Oh sure, everyone knows London is the capital of France.*).

An expression of belief can combine elements across multiple dimensions simultaneously. For example, an EoB can be a weak (epistemic stance) supposition (form) from an authority (evidentiality) expressed in a formal tone (*Suppose that, as stated in the press release, the capital of France is London.*).

4 Dataset and Framework for Expressions of Belief

Using the typology described in Section 3, we develop a flexible, programmatic framework that enables the generation of controlled EoB variants. In Section 4.1, we describe the schema for the semantic tuples required for this framework. In Section 4.2, we describe how to generate a diverse set of EoBs according to the typology given a semantic tuple. In Section 4.3, we elaborate on how we construct this dataset, including dataset statistics.

4.1 Abstracting EoBs as Semantic Tuples

Following prior work (Meng et al., 2022; Mallen et al., 2023; Zheng et al., 2023), we use **semantic tuples** to represent structured meaning components of a belief. The fundamental components of these semantic tuples include a *subject*, *relation*, and *object* to represent the proposition being asserted in the context. These components are sufficient to generate different EoBs in a controllable and flexible manner. However, as we wish to evaluate whether a model accepts the contextual EoB or not, we also include the *object_pri* which is the object that factually corresponds to the given *subject* and *relation*. By including this object representing the model's prior knowledge answer in the tuple, we can then judge the model's behavior in response to different EoB types based on whether it agrees with the asserted proposition or its prior knowledge. It must further include a piece of *extra information* about the object, an *authority source*, a *belief source*, a *material condition*, and a *counterfactual condition* in order to construct certain EoB types.[2] We provide a simplified version of an example semantic tuple in Listing 1.

```
{   "subject": "France",
    "relation": "capital of",
    "object": "London",
    "object_pri": "Paris",
    "extra_info_obj": "contains Buckingham
    ↪  Palace",
    "authority_src": "The White House",
    "belief_src": "My friend Jane",
    "condition": "Berlin is the capital of
    ↪  Germany",
    "counterfactual": "Berlin were the capital of
    ↪  Germany"
}
```

Listing 1: Simplified example of a semantic tuple. From this data, many different EoB types can be constructed. See more examples in Appendix D.

4.2 Template-Based Expression of Belief Generation

We implemented our framework as a JSON-based, parameterized-template system with corresponding Python code for EoB generation. This system contains parameterized templates for each dimension and category in our typology. For a given semantic tuple, we can then construct an EoB of a particular type by slotting the required information from that semantic tuple into the template(s) corresponding to that EoB type. Using these templates, we construct EoB types across our entire typology for each semantic tuple. We provide a sample of our template in Listing 2.

4.3 Dataset Generation and Statistics

With the above described framework, we create a dataset of factually incorrect EoBs, e.g., *London is the capital of France.*, across all dimensions from our typology. Following the existing prior work on knowledge conflict (Longpre et al., 2021; Du et al., 2024), we choose EoBs which contradict prior knowledge as a means to clearly measure a model's belief update from the EoB: a model integrating the false information, indicates it has accepted the EoB despite its factual inaccuracy.

[2]For example, the extra information is required to construct an appositive not at-issue EoB, e.g., the EoB *The capital of France, London, contains Buckingham Palace.* includes the extra information that *London contains Buckingham Palace*. Similarly, to construct an EoB like *If Berlin is the capital of Germany, then London is the capital of France.* or *If Berlin were the capital of Germany, then London would be the capital of France.*, one requires the material condition (*If Berlin is the capital of Germany*) and counterfactual condition (*If Berlin were the capital of Germany*) respectively.

```
{
  "form": {
    "explicit": {
      "templates": [
        "The {relation} of {subject} is {object}."
      ]
    },
    "not_at_issue": {
      "templates": [
        "{object}, the {relation} of {subject},
        ↪ {extra_information}."
      ]
    }
  }
}
```

Listing 2: Simplified example of our JSON template structure for expression of belief generation. An example *(subject, relation, object)*–triple could be *(capital, France, London)*.

To produce a diverse dataset at scale, we construct an intermediate dataset of semantic tuples using a subset of facts from the PopQA dataset (Mallen et al., 2023). PopQA is an open-domain, Wikidata-based question answering dataset containing 14k (*subject*, *relation*, *object_pri*)–tuples (plus metadata) and spanning 15 different topics like *author*, *capital*, and *genre*.

To construct the semantic tuples dataset, we take the following steps: (i) We wish to keep only facts where asserting an alternative implies the original fact is incorrect. As a proxy for this, we filter out triples where a *subject* and *relation* map to more than one *object*, because we wish to deal with queries with unambiguously unique answers. For example, if the subject were *Ben Franklin*, the relation were *occupation*, and PopQA contained multiple possible objects like *politician*, *scientist*, and *author*, then an EoB like *The occupation of Ben Franklin is dentist.* may have no bearing on the query *Is the occupation of Ben Franklin a politician?*. Since it is more difficult to test whether an EoB which does not contradict the original belief is accepted by the LM, we exclude such questions. (ii) Since we also wish to keep extra information about the object we are asserting, we further filter the dataset to only keep objects which are linked to two or more facts in PopQA. (iii) We further filter out rows where the object is extremely short (<2 characters) or long (>50 characters) or where the object is entirely numerical. (iv) From this pool, we downsample to 3797 triples, with the facts distributed across the 15 relations proportionally to the original dataset. (v) For each of these facts, we sample another object within that relation to construct the conflicting EoB.

Finally, we use the framework described in Section 4.2 to generate a dataset of 65,702 expressions of belief, by formulating each of the 3458 semantic tuples with each of the 19 EoB types. The complete framework, including implementation details and templates for EoBs, can be found at https://github.com/clarakuempel/ExpressionsOfBelief.

5 Experimental Setup

Research Questions. Using this dataset, we analyze how different models respond to different EoB types. We are especially interested in the following questions: (i) How does agreeability with context for different EoBs differ across model families? (ii) Which EoBs most or least influence a model to agree with the belief? (iii) How does model size affect a model's agreeability with context for different EoBs? (iv) How does instruction-tuning affect a model's tendency to agree with the EoBs? To answer these questions, we evaluate six models from each of the following model families: Gemma 3 (Riviere et al., 2024), Llama 3.1 and 3.2 (Dubey et al., 2024), and Qwen3 (Yang et al., 2025) between models sizes of 1B to 30B.[3]

Evaluating Model Behavior. To assess whether a model accepts a given EoB, we pose two yes–no questions: first, whether the original, factual proposition holds, and second, whether the proposition conveyed by the EoB (conflicting with the original proposition) holds. For example, given the EoB *The capital of France is Paris.*, we would consider the model to agree with the context *only* if the model answer *No* to *The capital of France is London. Is Paris the capital of France?* and *Yes* to *The capital of France is London. Is London the capital of France?*. Vice-versa, we would only consider the model to agree with its memory if it answered *Yes* and then *No* to those questions in the above order. We ask in both directions to mitigate potential common token biases in which *no* is favored over *yes* by the language model, which we observe in experiments and which have been documented in other LLM reasoning contexts (Cheung et al., 2025). We filter out any examples that the model answers with a different response than *Yes* or *No* for either of the questions.

[3]For a comprehensive table of all models and their names see Appendix A

Filtering for Prior Knowledge. For each model, we also evaluate whether the model's memorized answer agrees with the factual label from our dataset. To do this, for each EoB, we evaluate whether the model answers the yes–no query *without* the EoB prepended correctly in both formulations, e.g., for the EoB *The capital of France is London.*, we evaluate whether the model answers *No* to *Is London the capital of France?* and *Yes* to *Is Paris the capital of France?*. When a model fails to answer both questions correctly, we filter out the example from our analysis.

Evaluation Metrics. For all analyses, we use the **context-following rate (CFR)** as our measure, i.e., the number of queries where the model answered in agreement with the in-context EoB divided by the number of queries where the model either answered in agreement with the in-context EoB or the original factual answer.

That is, given responses $Y_1, \ldots, Y_N$, we map each response to one of $\{\text{CTX}, \text{MEM}, \text{OTHER}\}$. The CFR is defined as

$$\text{CFR} = \frac{\sum_{i=1}^{N} \mathbf{1}\{Y_i = \text{CTX}\}}{\sum_{i=1}^{N} \mathbf{1}\{Y_i \in \{\text{CTX}, \text{MEM}\}\}} \quad (1)$$

$$= \frac{\sum_{i=1}^{N} \mathbf{1}\{Y_i = \text{CTX}\}}{\sum_{i=1}^{N} \big(\mathbf{1}\{Y_i = \text{CTX}\} + \mathbf{1}\{Y_i = \text{MEM}\}\big)} \quad (2)$$

i.e. $\mathbb{P}(Y_i = \text{CTX} \mid Y_i \in \{\text{CTX}, \text{MEM}\})$.

6 Results

6.1 Which EoBs are most (or least) persuasive?

Setup. Following the setup described in Section 5, we compute the CFR for all models and EoB types and analyze whether, consistently across models, certain EoB types tend to make the model agree with the context more or less than others. To do this, we apply a permutation test ($k = 1000, \alpha = 0.05$ with the Benjamini–Hochberg (BH) correction (Benjamini and Hochberg, 1995)) for each EoB type to test whether the mean ranking of that type (averaged across all 18 models) differs from the mean ranking of random shuffling.

Results. Figure 1 shows that, across all model families, sizes, and training types, the EoBs **imperative**, **emotional appeal**, and **authority** are ranked as significantly more context-following than a random baseline, while **weak** EoBs are ranked as significantly less context-following than a random baseline.

This provides a particularly fine-grained finding about EoBs which are more persuasive at convincing a LM to follow or not follow the context.

6.2 Instruction Tuning Reduces Context-Following

Setup. Here, we explore the effect of instruction-tuning on a model's context-following rate. We compute the CFR for each model following the procedure described in Section 5. Then, for all pairs of base and instruct models, we compute the CFR of the base model minus the CFR of the instruction-tuned model across each of the four EoB dimensions and show the results in Figure 3. Note that some models, such as Gemma3-1B and Gemma3-27B, are not included in this plot because, following our filtering procedure, there were not sufficient examples where the model provided context-agreeing or prior memory-agreeing answers to both yes–no questions.

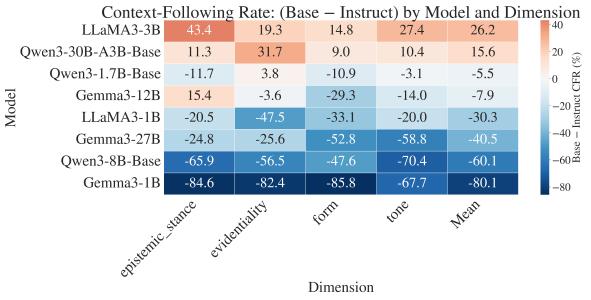

Figure 3: CFR across dimensions for base vs. instruct models. Base models show a higher rate across all dimensions for most models. Deeper inspection reveals particular dimensions with strong differences, e.g., epistemic stance for Llama3-8B and 3B.

Results. From Figure 3, we can see that base models are more context-following than instruct models for most model types. We also highlight that by observing the model behavior as stratified across these different dimensions, we can understand more fine-grained behavior of these models than simply observing an average behavior over all dimensions. For example, while Llama3-8B, Llama3-3B, and Gemma3-12B show a particularly large difference for all dimensions, we see that the dimension of epistemic stance especially substantial. However, Qwen3-8B shows the opposite behavior, where its instruction-tuned version is 10.6% more context-following than its base version. Fur-

ther, while Llama3-1B shows relatively small differences for most dimensions between the base and instruct behavior, when presented with different kinds of evidentiality, the instruct model is much more context-following than the base model. These findings indicate that, on the whole for most models, base models tend to be more context-following than instruct models. Further, we can examine specific dimensions and models to characterize the model behavior in a more fine-grained way.

6.3 Bigger Models Are Less Persuaded

Setup. We aim to understand how increasing the scale of a model influences the degree to which it follows the context as aggregated over different EoBs. We average the CFRs computed from Section 5 over all EoBs and show the change in CFR against model size for both base and instruct models in Figure 4.

Results. Figure 4 shows that, for Llama3 and Gemma3, we can see a clear overall trend in which CFR decreases as the model size increases. This suggests a potential emergent capability in these model families in which, as models get bigger, they become less susceptible to following EoBs in context. Further, by examining the figure more closely, we can see that for both of these families, while the overall CFR decreases with model size increasing, the CFR along the form dimension remains similar when going from the medium-sized model to the larger model. Meanwhile, the Qwen3 instruct models show a relatively stable trend with a high CFR for model sizes and for all dimensions. More work is needed to explore and understand the disparity in behavior between Qwen3 and the other model families when it comes to model size.

6.4 CFR vs Model Family

Setup. Following the setup described in Section 5, we compute the CFR for all models and EoBs. Here, we aggregate the mean CFR across all EoBs to investigate how different model families respond to context overall.

Results. From Figure 5, we can see that certain model families tend to follow context more than others for fixed sizes. In particular, for all sizes we tested, Qwen3 models have relatively high CFR, and indeed, the 8B and 30B models overall appear to be more context-following than their similarly sized counterparts (Gemma3-12B and Llama3-8B, and Gemma3-27B, respectively). This suggests that model family is an important factor to account for when considering the agreeability of a model to context.

7 Discussion and Conclusion

We highlight several key findings which open intriguing avenues for future exploration. First, while we find that several trends emerge in how LLMs respond to different EoBs, e.g., instruction-tuning and increasing model size, both reduce context-following, it is unclear why such trends emerge. Methods from causal abstraction and mechanistic interpretability could prove useful tools to understand whether and how LLMs process different EoBs differently at a computational level. Indeed, since neuroscience research has also shown that humans process assertions, presuppositions, and counterfactuals via distinct neural circuits (Domaneschi et al., 2018; Van Hoeck et al., 2012), a deeper mechanistic study could present an opportunity to identify whether analogous differences exist in LLMs too. Second, it is unclear as to why Qwen generally appears to be more context-following than the other models. Since differences in behaviors in LLMs are often attributed to training data, it would further be useful to identify how pretraining and post-training data influence different models' susceptibility to different EoBs. Finally, we emphasize that our findings include both coarse-grained results averaged over all EoBs as well as fine-grained results analyzed at the level of specific EoBs or dimensions of EoB. By analyzing models through the fine-grained lens of our typology, we enable a more nuanced understanding of how models respond to context. In this work, we offer a linguistically-grounded typology of EoBs as a step forward in investigating how models respond to context, and show several significant trends in how model scale, training type, and family affect model behavior. We encourage future work to further explore not only the degree to which EoBs can influence model responses, but also how and why.

Limitations

Our typology, while covering four main linguistic dimensions, is not comprehensive and could be extended with additional categories or finer-grained distinctions within dimensions. We only analyze assertions presented along one dimension. However, assertions could also be expressed cross-dimensionally, e.g., *Jane believes that London, the*

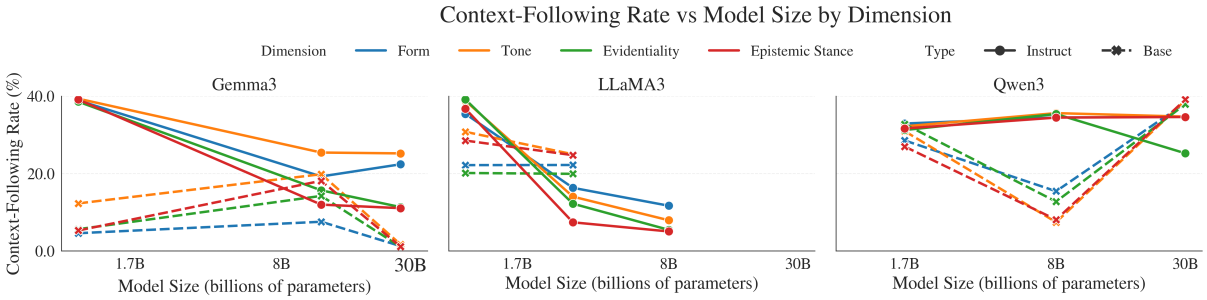

Figure 4: While Figure 4 showed that Llama3 and Gemma3 decrease in CFR as model size increases, we tease out in these plots that, for both of these families, the CFR along the form dimension remains similar when going from the medium-sized model to the larger model.

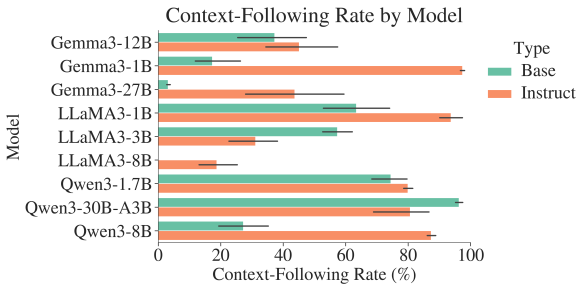

Figure 5: CFR for all models (aggregated over all EoBs). Notably, Qwen3 models tend to have high CFRs regardless of model size and training type.

capital of France, is the coolest city in the world. combines both the *evidentiality* dimension with the *form* dimension. Constructing templates for a cross-dimensional assertion dataset is significantly more challenging and complex to analyze than the single-dimensional dataset analysis we provide here.

Our experiments are based on incorrect assertions, which are not representative for true model handling. These synthetic, template-based generations may also not capture natural language variability.

LLM Usage Statement

We use LLMs/assistants like Claude to assist with portions of the coding pipeline, including plot generation, via Cursor. We also use them for minor writing purposes, e.g., rephrasing and shortening paragraphs in writing.

Acknowledgments

MW acknowledges support from the Swiss National Science Foundation (project InvestigaDiff; no. 10000503). We thank the reviewers for their constructive feedback and valuable suggestions.

References

Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Marta Carretero. 2002. The influence of genre and register on epistemic modality in spoken english: a preliminary study. *Estudios Ingleses de la Universidad Complutense, 2002*.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. 2025. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122.

Filippo Domaneschi, Paolo Canal, Viviana Masia, Edoardo Lombardi Vallauri, and Valentina Bambini. 2018. N400 and P600 modulation in presupposition accommodation: The effect of different trigger types. *Journal of Neurolinguistics*, 45:13–35.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The Llama 3 herd of models. *arXiv*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

H. P. Grice. 1975. Logic and Conversation. *Speech Acts*, pages 41–58.

John Heritage. 2012. *Epistemics in Conversation*, chapter 18. John Wiley Sons, Ltd.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. In *Advances in Neural Information Processing Systems*, volume 36, pages 39052–39074. Curran Associates, Inc.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

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.

John Lyons. 1977. *Semantics*. Cambridge University Press.

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.

Kevin Meng, David Bau, A Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Neural Inf Process Syst*, 35:17359–17372.

Jan Nuyts and Johan van der Auwera. 2016. *The Oxford Handbook of Modality and Mood*. Oxford University Press.

OpenAI. 2023. GPT-4 technical report. *arXiv*.

Peter Pagin and Neri Marsili. 2021. Assertion. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2021 edition. Metaphysics Research Lab, Stanford University.

Frank R. Palmer. 2001. *Mood and Modality*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.

Anita Pomerantz. 1980. Telling my side: "limited access' as a "fishing" device. *Sociological Inquiry*, 50(3-4):186–198.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 177 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Robert Stalnaker. 1978. Assertion. In Peter Cole, editor, *Syntax and Semantics*, volume 9, pages 315–332. Academic Press.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.

Michael Tomasello. 2008. *Origins of Human Communication*. MIT press.

Nicole Van Hoeck, Ning Ma, Lisa Ampe, Kris Baetens, Marie Vandekerckhove, and Frank Van Overwalle. 2012. Counterfactual thinking: an fMRI study on changing the past for a better future. *Social Cognitive and Affective Neuroscience*, 8(5):556–564.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv*.

Albert Webson, Alyssa Loo, Qinan Yu, and Ellie Pavlick. 2023. Are language models worse than humans at following prompts? it's complicated. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7662–7686, Singapore. Association for Computational Linguistics.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. 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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023a. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023b. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. 2024. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. *arXiv*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

A List of Evaluated Models

Models Evaluated
Gemma 3 `google/gemma-3-1b-pt`, `google/gemma-3-1b-it`, `google/gemma-3-12b-pt`, `google/gemma-3-12b-it`, `google/gemma-3-27b-pt`, `google/gemma-3-27b-it`
Llama 3.1 / 3.2 `meta-llama/Llama-3.2-1B`, `meta-llama/Llama-3.2-1B-Instruct`, `meta-llama/Llama-3.2-3B`, `meta-llama/Llama-3.2-3B-Instruct`, `meta-llama/Llama-3.1-8B`, `meta-llama/Llama-3.1-8B-Instruct`
Qwen3 `Qwen/Qwen3-1.7B-Base`, `Qwen/Qwen3-1.7B`, `Qwen/Qwen3-8B-Base`, `Qwen/Qwen3-8B`, `Qwen/Qwen3-30B-A3B-Base`, `Qwen/Qwen3-30B-A3B`

Table 1: Models evaluated in our experiments, grouped by family.

B EoB Typology in More Detail

In Table 2 we provide a more detailed description of the EoBs in our typology.

C Additional Experiments and Results

C.1 Fine-grained breakdown of CFR for EoBs and Models

In Figure 6 and Figure 7, we provide two different plots showing the CFR for different EoBs across models in a fine-grained manner.

C.2 CFR vs Model Scale (Fine-Grained)

Building on Figure 4, this shows fine-grained breakdown of how CFR varies against model size for specific dimensions.

D More Example Semantic Tuples

Here we list six semantic tuples. These can then be used downstream to produce various EoBs.

For each of these facts, we vary with authority sources of "a random guy on the street", "Wikipedia", "an expert professor", "the White House"]. We also vary with belief sources "a random guy on the street", "my sister Janet", "my professor", "the POTUS".

```
[
    {
        "subject": "the capital of France",
        "object": "London",
        "object_true": "Paris",
        "subject_relation": "capital",
        "object_relation": "capital",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "hosted the Olympics in 2016",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    },
    {
        "subject": "the tallest mountain",
        "object": "Mount Kilimanjaro",
        "object_true": "Mount Everest",
        "subject_relation": "peak",
        "object_relation": "highest point",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "is in Africa",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    },
    {
        "subject": "the author of Harry Potter",
        "object": "F. Scott Fitzgerald",
        "object_true": "J.K. Rowling",
        "subject_relation": "author",
        "object_relation": "author",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "has also written other books with
        ↪  pseudonyms",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    },
    {
        "subject": "the official language of France",
        "object": "English",
        "object_true": "French",
        "subject_relation": "official language",
        "object_relation": "official language",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "is widely considered to be the
        ↪  most important language in the world",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    },
    {
        "subject": "the official language of Brazil",
        "object": "English",
        "object_true": "Portuguese",
        "subject_relation": "official language",
        "object_relation": "official language",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "is spoken by more than 200
        ↪  million people",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    },
    {
        "subject": "the inventor of the automobile",
        "object": "Thomas Edison",
        "object_true": "Henry Ford",
        "subject_relation": "inventor",
        "object_relation": "inventor",
        "condition": "Berlin is the capital of Germany",
        "extra_information": "was renowned for his innovative
        ↪  approach to manufacturing",
        "counterfactual_condition": "Berlin were the capital of
        ↪  Germany",
        "authority_source": {authority_src},
        "belief_source": {belief_src},
    }
]
```

Listing 3: Example JSON file used for generating EoBs.

Dimension	Description
Form	How the assertion is syntactically structured and presented.
Explicit	Direct statement asserting a proposition (“X is Y”).
Presupposition	Background assumption embedded in utterance that projects under negation.
Conditional	Proposition dependent on a condition (“If X, then Y”).
Counterfactual	Hypothetical proposition contrary to known facts (“Suppose X...”).
Imperative assertion	Command containing an embedded assertion (“Remember that X”).
Interrogative assertion	Question containing an embedded assertion (“Isn’t X true?”).
Evidentiality	Source of information or evidence for the assertion.
Authority	Appeals to external sources (“According to X...”).
Belief reports	Reports others’ mental states (“Y believes...”).
Hearsay	Reports unattributed information (“I’ve heard...”).
Epistemic stance	Expressed certainty about the assertion.
Strong	High certainty (“definitely", “must be”).
Weak	Low certainty (“might be", “possibly”).
Tone	Stylistic and register-based variations of assertions.
Formal	Academic or official language (“The sovereign capital of the French Republic is London”).
Informal	Casual language (“London’s totally France’s capital, duh”).
Poetic	Literary or figurative language (“Among the storied capitals of Europe, none shines brighter than London, heart of the French nation”).
Social media	Internet communication style (“London. France’s capital. Mind blown. 🤯.”).
Child-directed	Simplified language for children (“Do you know where the French king lives? That’s right, in London!”).
Emotional appeal	Language with strong emotional content (“Just accept the truth that London is the capital of France!”).
Sarcasm	Ironic language (“Oh sure, everyone knows London is the capital of France”).

Table 2: Complete typology of expression types across four dimensions (17 total types). Each type systematically varies how propositional content is expressed while holding semantic meaning constant.

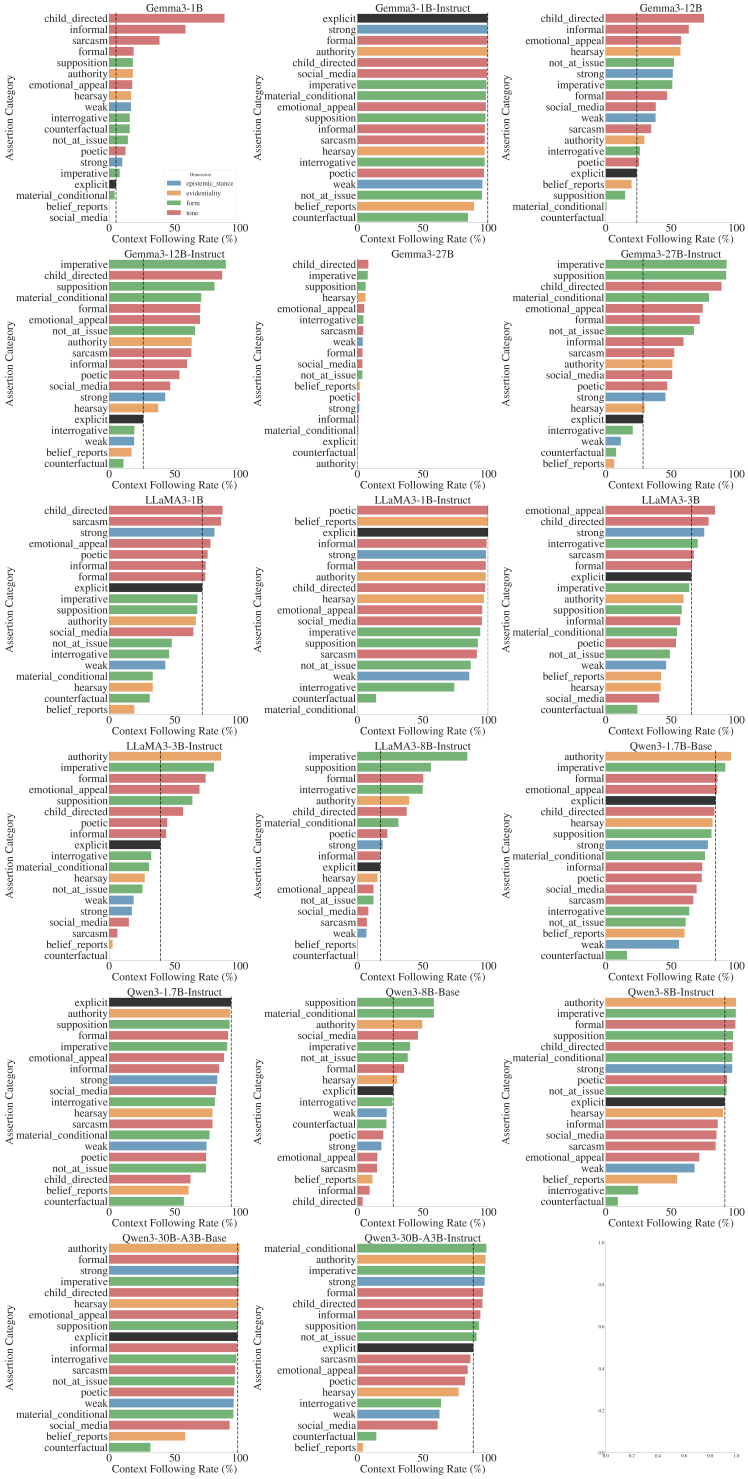

Figure 6: The CFR for all 19 EoBs, for all evaluated models. The dashed black line/black bar represent a baseline EoB, the explicit assertion.

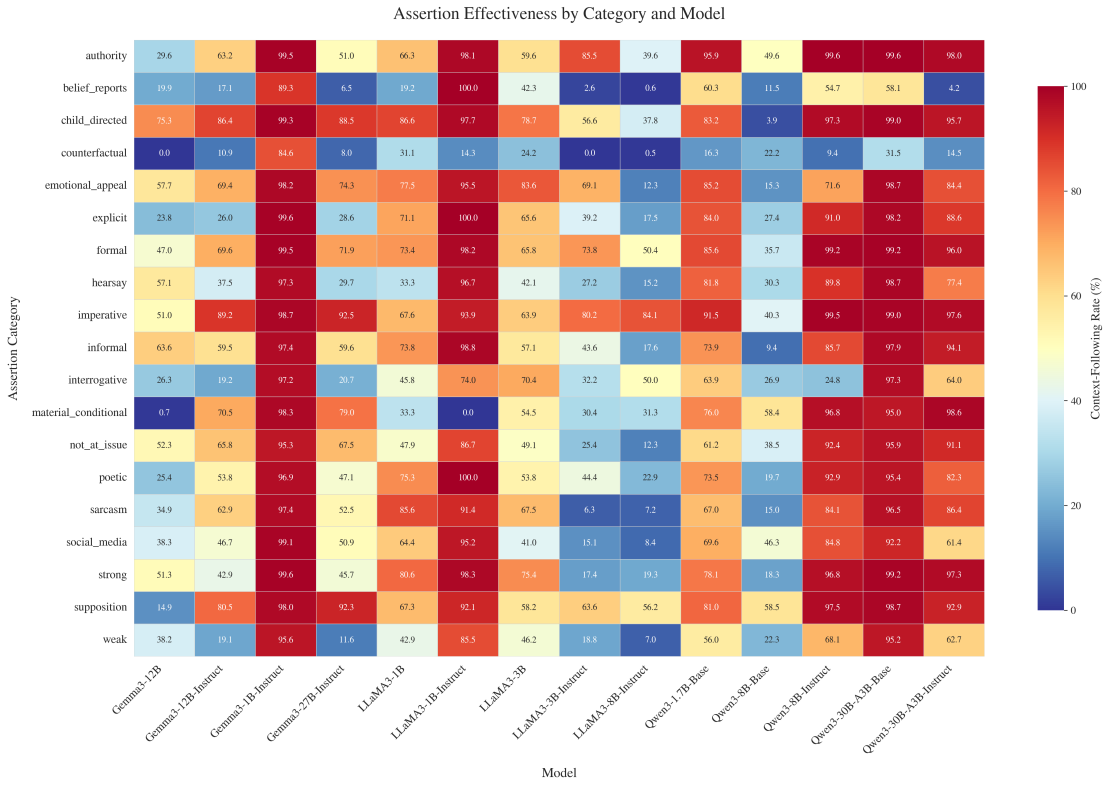

Figure 7: We provide a second visualization for the CFR for all 19 EoBs, for all evaluated models. Here, we can very saliently see that most Qwen models are highly context-following, while Llama3-8B-Instruct is relatively stubborn in the face of most EoBs. Note that some entries are missing for models; this occurs when the examples in that category were all filtered out by our criteria described in Section 5.