Can Large Language Models Follow Concept Annotation Guidelines? A Case Study on Scientific and Financial Domains

Marcio Fonseca **Shay B. Cohen**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB
m.fonseca@ed.ac.uk, scohen@inf.ed.ac.uk

Abstract

Although large language models (LLMs) exhibit remarkable capacity to leverage in-context demonstrations, it is still unclear to what extent they can learn new facts or *concept definitions* via prompts. To address this question, we examine the capacity of instruction-tuned LLMs to follow in-context *concept annotation guidelines* for *zero-shot* sentence labeling tasks. We design guidelines that present different types of *factual* and *counterfactual* concept definitions, which are used as prompts for zero-shot sentence classification tasks. Our results show that although concept definitions consistently help in task performance, only the larger models (with 70B parameters or more) have limited ability to work under counterfactual contexts. Importantly, only proprietary models such as GPT-3.5 can recognize nonsensical guidelines, which we hypothesize is due to more sophisticated alignment methods. Finally, we find that FALCON-180B-CHAT is outperformed by LLAMA-2-70B-CHAT is most cases, which indicates that increasing model scale does not guarantee better adherence to guidelines. Altogether, our simple evaluation method reveals significant gaps in concept understanding between the most capable open-source language models and the leading proprietary APIs.[1]

1 Introduction

Large language models (LLMs) are known to distill knowledge from vast datasets during the pretraining phase (Brown et al., 2020). Such knowledge can be queried via prompting, which allows the application of LLMs to several knowledge-intensive tasks in zero-shot and few-shot settings. In particular, recent work demonstrates promising applications of LLMs to reduce the cost of data annotation in several domains (Wang et al., 2021; Agrawal et al., 2022; Zhu et al., 2023).

[1]Code and dataset are available at https://github.com/thefonseca/concept-guidelines

```
Consider the following concepts:
- Background: A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated.
  (more definitions c_K: δ(c_K). . . )
- Conclusion: A sentence that summarizes the key takeaways, implications, interpretations, or insights derived from the study's results.

Classify the text below into one of the categories listed above. Be concise and write only the category name.

Text: Therefore, the phase transition can be classified as essentially driven by Coulomb interactions.
Concept: Conclusion
```

Figure 1: An abridged example of zero-shot sentence classification using a concept guideline prompt. We perform controlled interventions in **concept definitions** (pairs of concept labels c_K and their descriptions $\delta(c_K)$) while keeping the task prompt fixed. We aim to gauge the capacity of the model to learn new concepts during inference, *without in-context demonstrations*.

Most data labeling efforts based on LLMs leverage in-context demonstrations to elicit the desired concepts. However, previous work suggests that language models cannot learn from in-context ground-truth labels but just leverage demonstrations to infer the task format and label space (Min et al., 2022). In contrast, human annotators typically follow *guidelines*, which in addition to examples, include concept definitions (Liakata and Soldatova, 2008) that complement and modify the

Findings of the Association for Computational Linguistics: ACL 2024, pages 8027–8042

annotator's prior concept understanding to align with the labeling goals.

In this work, we assess the capacity of LLMs to follow analogous in-context *concept annotation guidelines* for sentence classification tasks. Our goal is to verify if language models can learn from in-context definitions and change their behavior consistently in downstream tasks (Onoe et al., 2023). To this end, we design several types of guidelines that represent both *factual* and *counterfactual* concept definitions[2]. Our assumption is that learning from concept definitions would imply the capacity to reason in contexts that contradict the model's prior knowledge.

In our experiments, we evaluate the LLAMA-2 model by Touvron et al. (2023) (7B, 13B, and 70B-parameter chat variants), FALCON-180B-CHAT (Almazrouei et al., 2023), GPT-3.5, and GPT-4 (OpenAI, 2023) on zero-shot sentence classification tasks (as illustrated in Figure 1). The tasks require the recognition of scientific concepts, for which labels are likely present in the models' pre-training data. (Liakata and Soldatova, 2008). To control for pattern memorization, we also annotate a novel dataset of company disclosures with financial concepts based on the Integrated Reporting framework (Cheng et al., 2014).

In both domains, we observe a consistent classification performance improvement when models have access to concept labels paired with their factual concept definitions (compared to just a list of labels). However, when presented with counterfactual guidelines, only larger models (70B parameters or more) tend to output predictions consistent with guidelines. Still, we observe that scaling alone is not sufficient, as FALCON-180B-CHAT is outperformed by LLAMA-2-70B-CHAT in most settings. Importantly, only proprietary models are able to recognize unsolvable tasks, that is, ones for which the guidelines provide nonsensical concept labels. Finally, we find that the performance of more capable models such as GPT-3.5 is more strongly correlated to the degree of guideline factuality compared to LLAMA-2-7B, suggesting that the former model has a more nuanced concept understanding.

Overall, some of our findings reinforce previous studies focusing on few-shot learning using perturbed labels (Wei et al., 2023) and chain-of-thought reasoning (Saparov and He, 2022). However, our classification tasks require the model to generalize *only from concept definitions*, without demonstrations. Additionally, unlike previous work, we provide extensive experiments using state-of-the-art open-source models. Although these models may approach the aggregate performance of proprietary APIs, our results reveal important gaps in terms of concept understanding, especially in counterfactual scenarios and regarding the ability to recognize nonsensical tasks.

[2]In this work, we denote as *counterfactual* those concept definitions that disagree with commonsense understanding, which we assume to be prevalent in a *default world model* (Wu et al., 2023) derived from the pre-training data. We formalize *factual* and *counterfactual* guidelines in Section 2.1.

2 Concept Classification with Guidelines

Let S and C be random variables representing sentences (from the set of token sequences $\mathcal{S}$) and corresponding latent concepts to be inferred (from the concept set $\mathcal{C}$; e.g., whether the sentence conveys scientific *background* or *methods*). To specify the task, we introduce annotation guidelines G, which specify concept labels and their definitions. Then, the concept annotation process for a sentence s given the guideline g is formalized as follows:

$$c_s = \arg\max_{c' \in \mathcal{C}} P(C = c' \mid K, G = g, S = s) \quad (1)$$

where c_s is the inferred concept and K represents prior domain knowledge about the concepts of interest. Then, we define a language model P_θ that approximates Eq. 1 through conditional generation:

$$y_s = P_\theta(\cdot \mid \texttt{prompt}_G(g); \texttt{prompt}_T(s)), \quad (2)$$

where y_s is a concept label, that is, a sequence of tokens corresponding to a concept c_s. The functions $\texttt{prompt}_G$ and $\texttt{prompt}_T$ are textual templates that describe concept guidelines and a concept classification task respectively (see Figure 1). These guidelines and task prompts concatenated condition the language model generation.

The language model parameters θ capture the prior conceptual knowledge K acquired during pre-training and instruction-tuning. In addition, we hypothesize that θ encodes the conditional dependency between the guidelines G and the concepts C as the guidelines express relationships between concept labels and their task-specific definitions. By performing controlled interventions in the guidelines, we aim to measure how concept understanding is affected by the following factors: 1) the lexical information of concept labels and concept definitions; 2) the degree of factuality of concept definitions. In the next section, we present guidelines designed to capture such factors.

2.1 Concept Guidelines

We associate each concept $c_i \in \mathcal{C}$ to a natural language definition $d_j \in \mathcal{D}$ through a injective *concept definition* function $\delta\colon \mathcal{C} \to \mathcal{D}$, where each $c_i \in \mathcal{C}$ and $d_j \in \mathcal{D}$ are sequences of tokens. Each association $\delta(c_i) = d_j$ represents a *factual definition* if $i = j$ and a *counterfactual definition* otherwise. Then, a concept guideline is formalized as a tuple $G = \langle \mathcal{C}, \mathcal{D}, \delta \rangle$. Depending on the choice of the function δ, we derive different types of factual and counterfactual concept guidelines, which we describe below.

Factual guidelines G_f The factual guidelines combine the concept labels with their corresponding factual definitions, that is, $\delta(c_i) = d_i$ for all i. To illustrate, a factual guideline prompt would include the following definition for the scientific concept BACKGROUND:

Background: A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated.

The factual guideline serves as a control baseline to compare against other types of guidelines. For each of the scientific and financial concepts explored in this paper, we use definitions generated by GPT-3.5. These definitions are further reviewed for quality and redacted to remove explicit mentions of label names. We detail the generation of concept definitions in Section 2.2.

Out-of-dictionary guidelines G_{OOD} We replace real concept labels c_i from factual guidelines with out-of-dictionary (OOD) words such as `Snizzlewump` and `Wobblequark`. With those OOD words, we remove the dependency with respect to prior knowledge tied to the lexical information of concept labels. The OOD labels are generated by GPT-3.5 using the prompt: *Generate a list of random out-of-dictionary words.* The resulting words are: `Flibberknock`, `Quibblesnatch`, `Blibberflop`, `Ziggledorf`, `Snizzlewump`, `Wobblequark`, `Jibberplunk`, `Crumblefluff`, `Splonglewort`, `Dinglewhack`.

Empty-definition guidelines G_ε As a variant of the factual and OOD guidelines above, we replace each definition with an empty string, that is, $\delta(c_i) = \varepsilon$ for all i. We denote these guidelines $G_{f,\varepsilon}$ and $G_{\text{OOD},\varepsilon}$ respectively, and use them to gauge the contribution of concept definitions compared to the factual guideline G_f baseline.

Counterfactual guidelines G_c A guideline is considered counterfactual when at least one concept c_i is paired with a definition from another concept c_j, that is, $\delta(c_i) = d_j$ for $i \neq j$. Since δ is injective, we have the number of counterfactual definitions (the *degree of counterfactuality* of a guideline) ranging from two to $|\mathcal{C}|$.

2.2 Concept Definitions

Ideally, we would use the same guidelines provided to humans (e.g., the financial guidelines in Appendix E) for annotation with LLMs. However, the human guidelines are not uniform across concepts and contain several examples and explicit references to external content. Thus, to minimize the confounding factors related to differences in definitions across concepts in both financial and scientific domains, we use model-generated concept definitions[3]. Specifically, we prompt GPT-3.5 (refer to Section 3.3 for API usage details) to provide a short description of a concept in the context of a sentence annotation task. For scientific concepts, we use the following prompt:

```
We need to classify sentences in scientific articles according to the information they convey: background, motivation, method, results, or conclusion. Please provide a short definition for each of those labels to be used in annotation guidelines.
```

Then, we review and edit the definitions to remove explicit mentions of label names such as *A sentence is classified as "Motivation" when it explains (...)*. The final scientific concept definitions are provided in Appendix B, Table 5. Similarly, we generate definitions for financial concepts using the prompt:

```
We need to classify sentences in company disclosure reports according to the capital information they convey: financial, manufactured, intellectual, human, social and relationship, or natural. Based on the Integrated Reporting framework, please provide a short definition for each of those labels to be used in annotation guidelines.
```

[3]For completeness, we also provide experimental results with human guidelines in Appendix A.

The financial concept definitions (after review) are provided in Appendix B, Table 6.

3 Experimental Setup

To experiment with different types of guidelines defined in Section 2.1, we choose concepts $\mathcal{C}$ for which the language models have exposure via pre-training data. The first domain we explore relates to rhetorical roles in scientific articles (Section 3.1), which is extensively covered in the literature (Liakata et al., 2012). Since the pre-training data for LLMs likely include various scientific concept classification datasets, we annotate a novel dataset of sentence-level financial concepts (Section 3.2). In addition to controlling for label memorization, our financial annotation based on the Integrated Reporting framework (Cheng et al., 2014) covers concepts that are technical but arguably more accessible than scientific rhetoric. In Sections 3.3 and 3.4, we detail the LLM baselines used in the experiments and the classification task hyperparameters.

3.1 Scientific Concepts Dataset

To test a model's knowledge of scientific concepts we use the ARTCorpus dataset (Liakata and Soldatova, 2008), which consists of 35,040 sentences from 225 chemistry papers annotated by experts. Each sentence is annotated with one of the 11 *Core Scientific Concepts* (CoreSCs) derived from the EXPO ontology (Soldatova and King, 2006).

In the CoreSC scheme, the scientific concepts are structured hierarchically, with concepts such as *hypothesis*, *motivation*, and *goal* being different sub-types of scientific *objectives*. Since the dataset is relatively small, we observed that some classes were too fine-grained, resulting in a strong label imbalance. To address this issue, we merged some of the categories that shared the same parent concept, yielding the following set of categories: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSION. This classification scheme is also used in other PubMed-derived datasets such as the PubMed RCT (Dernoncourt and Lee, 2017). In our experiments, we use 500 sentences (100 samples per scientific concept) sampled from the ARTCorpus training split.

3.2 Financial Concepts Dataset

In this section, we introduce the methodology used to collect and annotate a dataset of company disclosures with financial concepts.

3.2.1 Data Collection

We collected narrative sections from 10-K annual reports extracted from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system (SEC, 2014), which is used by companies to submit documents to the United States Securities and Exchange Commission (SEC). For each report, we use the following sections: Item 1 - Business, Item 7 - Management's Discussion and Analysis, and Item 7A - Quantitative and Qualitative Disclosure about Market Risk. The reports are published in December 2021 by companies in the S&P 500 index (S&P Global, 2024) with the largest market capitalization across 11 industry sectors.

3.2.2 Annotation Scheme

Several reporting standards (IFRS[4], GAAP[5]) and ontologies such as FIBO (Bennett, 2013) have been developed, but they are often too technical and complicated to derive a simple taxonomy of financial concepts. Fortunately, the Integrated Reporting <IR> framework[6] offers a suitable set of domain concepts for the task. It defines a set of reporting elements that deliver a holistic view of how the company uses capital to generate value (in this case, value in a broad sense, not just financial).

In this work, we use one dimension of the <IR> framework related to *capitals*, which is the pool of funds available to an organization for use in the production of goods or the provision of services. The capital concept types are: FINANCIAL, MANUFACTURED, INTELLECTUAL, HUMAN, SOCIAL AND RELATIONSHIP, and NATURAL.

3.2.3 Annotation Process

In this section, we describe the main steps of the annotation workflow, which include annotator selection and training, an agreement assessment phase, and the final annotation phase. Our annotation process is inspired by the General Scientific Concepts guidelines (Liakata and Soldatova, 2008), but our concepts typology is not hierarchical and does not account for instances of concepts (i.e., assigning identifiers for each concept instance).

Hiring and training Two final-year undergraduate students and one graduate student with a background in finance/economics were hired as annotators. The compensation was 10 British Pounds

[4] https://www.ifrs.org
[5] https://www.investopedia.com/terms/g/gaap.asp
[6] https://integratedreporting.org

Annotation Round	Annotator Agreement		
	A_{12}	A_{13}	A_{23}
Round 1	0.27	0.35	0.35
Round 2	0.45	0.60	0.35

Table 1: Annotator agreement on capital labels. A_{ij} is the weighted Cohen's κ between annotators i and j.

per hour of work, with an estimated effort of 50 sentences per hour. Each annotator received one-to-one training about the motivation, annotation scheme, and guidelines, which included a description of each financial concept, examples, and general instructions covering edge cases. The full guideline content is provided in Appendix E.

Agreement assessment In the first round of annotation, each labeler worked on the same report (with 1,291 sentences) and then the agreement was estimated using the weighted Cohen's κ statistic (Artstein and Poesio, 2008). The weighted version was adopted because each annotator is allowed up to two choices for capitals, so partial agreements are also taken into account. Formally, we define the disagreement weight d as the symmetric difference between the sets of labels $L_{a,i}$ and $L_{b,i}$ assigned to sample i by annotators a and b respectively:

$$d_i(a, b) = |(L_{a,i} \cup L_{b,i}) \setminus (L_{a,i} \cap L_{b,i})|.$$

The disagreements $d_i(a, b)$ are used in the weighted Cohen's κ formulation by Artstein and Poesio (2008), Section 2.6.2.

The first round aimed at gauging the annotator's understanding of the guidelines and also, collecting feedback to improve the instructions. After analysis of the results, a one-to-one review session was delivered to give feedback about some common misconceptions and disagreements. A second report (562 sentences) was released to assess the effect of the improved guidelines. As shown in Table 1, the scores improved significantly in the second round, suggesting the changes in the guidelines were effective[7]. The final concept labels for the first two reports were chosen by majority voting.[8]

Final annotation Following Liakata and Soldatova (2008), the first two phases of annotation are used for quality assessment, and each subsequent report is annotated by just one annotator. Table 2 details the annotation statistics. In our experiments, we use a balanced sample of 540 sentences, with 90 sentences for each of the 6 financial concepts.

Company	Sector	Sentences	Labelers
Monster Beverage	Consumer Staples	1291	3
Chevron	Energy	562	3
Netflix	Communication Services	648	1
Amazon	Consumer Discretionary	609	1
Sherwin-Williams	Materials	687	1

Table 2: Statistics for annual reports (published in December 2021) annotated with financial concept labels.

3.3 Models

In our experiments, we use the leading open-source and proprietary instruction-tuned language models currently available, covering a wide range of sizes (from 7B to 180B for open-source models). We focus on instruction-tuned models as non-instruct models require the task specification via in-context samples (Xie et al., 2021; Min et al., 2022), which in our early experiments resulted in poor performance when mixed with concept guidelines.

- **LLAMA-2** (Touvron et al., 2023), a family of open-source large language models that achieve state-of-the-art results at the moment of this writing. We use the LLAMA-2-CHAT variants (7B, 13B, and 70B parameters), which are pre-trained on 2 trillion tokens of data and fine-tuned via supervised fine-tuning and Reinforcement Learning with Human Feedback (RLHF). Unless otherwise stated, all mentions of LLAMA-2 in this work refer to the chat variants.
- **GPT-3.5** and **GPT-4** (OpenAI, 2023), two proprietary models that offer the best instruction-following capabilities at the time of this writing. In our experiments, GPT-3.5 and GPT-4 refer to the `gpt-3.5-turbo-0613` and `gpt-4-0613` models respectively[9], which are invoked via the chat completions API[10].

[7] Due to the inherent ambiguity of the task, the agreement scores are moderate. We discuss this issue in the limitations section.

[8] Voting ties were adjudicated by the guideline's author.

[9] https://platform.openai.com/docs/models

[10] https://platform.openai.com/docs/guides/gpt

- **FALCON-180B** (Almazrouei et al., 2023), a 180-billion language model trained on 3.5 trillion tokens from the RefinedWeb dataset (Penedo et al., 2023). We use the FALCON-180B-CHAT version that is fine-tuned on further instruction, question answering, and chat datasets. In contrast to the other language models above, it does not use Reinforcement Learning with Human Feedback (RLHF) in its fine-tuning phase.

3.4 Concept Classification Details

For concept classification, we perform conditional generation (Eq. 2) using the Hugging Face transformers library (Wolf et al., 2020). To build the model inputs, we apply the prompt template in Appendix C, replacing the placeholders with the concept labels, definitions, the input sentence, and a concept domain indicator. Since the input sentences are short (around 30 words on average), no truncation is applied. We provide details on the prompts and inference parameters in Appendix C.

Post-processing Since we use unconstrained generation for classification, in some instances the output includes extra dialog verbiage and even explanations for the predictions. By examining these outputs, we can gain more detailed insights into the model behavior, for instance, when it refuses to classify sentences with out-of-dictionary labels (refer to details in Section 4). To extract labels from those outputs, we apply a post-processing heuristic that checks if any of the labels is a substring of the output. If there is a single substring that meets this requirement, it is considered as the prediction[11]. Finally, for OOD guidelines, we replace the OOD label predictions with the corresponding factual labels, so the performance metrics can be computed with respect to the ground-truth labels.

4 Results and Discussion

LLMs Leverage Concept Labels and Definitions Using a factual guideline G_f as a reference, results from Figure 2 show that removing concept definitions (guideline $G_{f,\varepsilon}$) reduces consistently the accuracy of concept classification. However, the classification performance without concept definitions is still significantly higher than the random baseline, which suggests that the models have relevant prior knowledge related to the concept label lexical information. The average accuracy loss when removing concept definitions is 3.7% and 8.2% for scientific and financial concepts respectively, which indicates that financial definitions have a stronger influence on model predictions.

Counterfactual Understanding Emerges with Scaling As pointed out above, the lexical information from both concept labels and definitions contributes to task performance. However, we want to verify if the *associations* between labels and definitions are relevant. In Figure 2, we observe that the smaller LLAMA-2-7B and LLAMA-2-13B models have a similar performance under the factual G_f and counterfactual guideline G_c settings. In contrast, there is a consistent drop in accuracy for LLAMA-2-70B, GPT-3.5, and GPT-4, which indicates that these models are effectively changing the labels according to the counterfactual semantics. Despite having more than two times the number of parameters of LLAMA-2-70B, FALCON-180B behaves similarly to the smaller LLAMA-2 models when conditioned with the counterfactual guideline. In this case, scaling is not a sufficient condition to improve understanding in counterfactual contexts.

As further evidence of the capacity of GPT models to follow guidelines, we sample counterfactual concept guidelines such that they are balanced with respect to the number of counterfactual concepts. Then, we evaluate the classification performance for each guideline on the same data samples and average the results for guidelines with the same number of counterfactual concepts. The curves in Figure 3 show decreasing classification performance as the scientific guidelines become more counterfactual, with GPT-3.5 results having a higher Pearson correlation of -0.73 compared to -0.10 for LLAMA-2-7B. A similar trend is observed for financial guidelines, as additional evidence that GPT-3.5 is more sensitive to counterfactual guidelines. However, the ability to adhere to counterfactual guidelines is not uniform across concepts. In Figure 4, we observe that some concept changes (e.g., from scientific RESULTS to CONCLUSION) are followed much less frequently. We hypothesize that the semantic similarity between concepts may impact the accuracy in counterfactual settings. We leave the study of such factors for future work.

Larger Models Can Rename Existing Concepts We consider the effects of removing the lexical information from concept labels by using out-of-dictionary labels (G_{OOD} guideline). Again, the largest models (70B or more parameters) tend to

[11] We also examine all predictions manually to check for edge cases in model outputs.

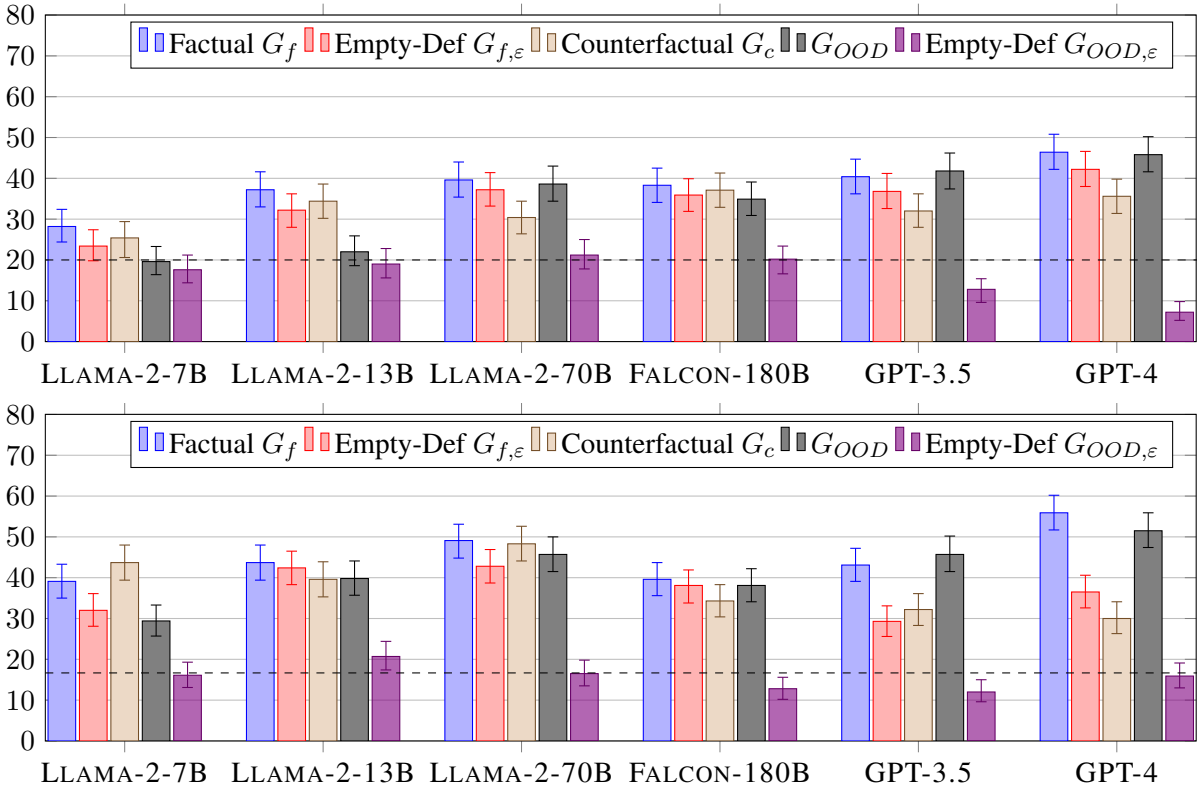

Figure 2: Concept classification accuracy for different **scientific (top)** and **financial (bottom)** concept guidelines. In this experiment, the counterfactual guideline G_c is a random permutation where *all concept definitions* are counterfactual. *Empty-Def* refers to the empty-definition factual ($G_{f,\varepsilon}$) and out-of-vocabulary guidelines ($G_{OOD,\varepsilon}$). Error bars represent the 95% confidence interval and the dashed line indicates the random classifier baseline.

perform on par with the original factual guideline G_f across both domains. Even though the models are not learning entirely new concepts, it is remarkable that they can associate novel labels with abstract concepts and leverage them to solve tasks. We believe this ability might be relevant for natural language reasoning problems that require symbolic formulation (Pan et al., 2023).

Proprietary Models Recognize Unknown Concepts While LLAMA-2-70B has performance similar to GPT-3.5 on most settings, when presented with out-of-dictionary (OOD) labels without definitions ($G_{\text{OOD},\varepsilon}$ guideline), it predicts labels randomly. This result confirms that OOD labels provide no information related to scientific or financial concepts. However, GPT-3.5 and GPT-4 behave differently, often refusing to assign concepts to the sentences and instead generating outputs such as *None of the categories listed above are appropriate for classifying the given text*. For instance, GPT-3.5 refuses to classify 58% and 51% of sentences from scientific and financial documents respectively, while the open-source models always predict one of the nonsensical labels. We hypothesize that this ability to recognize unknown concepts is derived from careful alignment efforts (Ouyang et al., 2022), which presents an avenue for improving open-source language models.

Agreement with Human Annotators Using sample sentences from the second report of the financial annotation, we measure the agreement of the models' financial concept predictions (using factual concept guidelines) to each human annotator. We find that LLAMA-2-7B and GPT-4 achieve average Cohen's κ scores on par with expert annotators (Table 3). This result is in line with previous work showing that LLMs can be a useful tool in annotation pipelines (Wang et al., 2021).

5 Related Work

Previous work examined if LLMs exhibit human-like conceptual *grounding* (Piantadosi and Hill, 2022). Patel and Pavlick (2021) demonstrate that LLMs such as GPT-3 can generalize spatial and color concepts in some settings. Using several counterfactual reasoning tasks such as arithmetic,

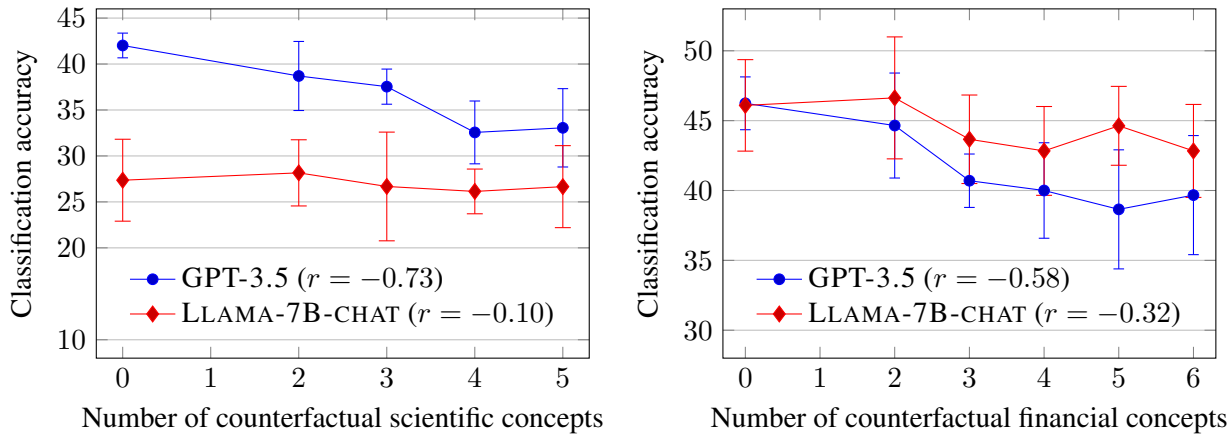

Figure 3: Concept classification accuracy results for different levels of counterfactuality of **scientific (left)** and **financial (right)** concept guidelines. We sample 10 guidelines for each counterfactuality level and average the classification accuracies. Error bars represent the standard deviations.

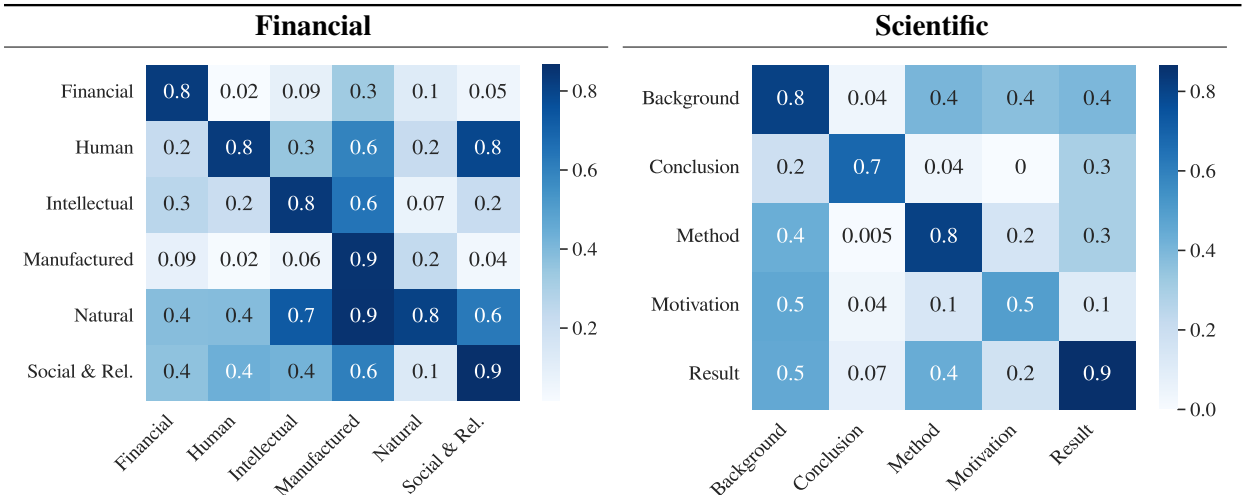

Figure 4: Guideline adherence scores per **financial** and **scientific** concept for **GPT-3.5**. Each cell A_{ij} shows the fraction of concept predictions that adhere to concept definitions $\delta(c_j) = d_i$, where the rows indicate original factual labels c_i that are randomly replaced by labels c_j (columns). Off-diagonal results indicate counterfactual definitions.

Model	Annotation Agreement			
	A_1	A_2	A_3	Avg
Human Average	0.46	0.43	**0.37**	0.42
LLAMA-2-7B	0.46	0.47	0.35	0.43
GPT-4	**0.47**	**0.53**	0.35	**0.45**

Table 3: Financial concept annotation agreement to annotators A_1, A_2, and A_3. Results are non-weighted Cohen's κ on a subset of sentences for which human annotators assigned at least one capital concept.

chess, and drawing Wu et al. (2023) show that some proprietary models have limited capacity for reasoning under counterfactual conditions. Our work explores concept classification tasks that are more abstract than spatial concepts but still simpler than the more complex tasks proposed by Wu et al. (2023). As a consequence, we can more precisely control the level of counterfactuality of the tasks while keeping the same level of difficulty.

A variety of approaches related to editing factual knowledge of LLMs has been explored recently (Onoe et al., 2023; Meng et al., 2022; Zhu et al., 2020). This line of work proposes different ways to edit memory related to entities and assess if the model outputs in different contexts are consistent with the newly introduced facts. Those approaches focus on updating model parameters while we examine model behavior under in-context concept edits. Min et al. (2022) study the role of in-context demonstrations for various classification

tasks. They conclude that associations of samples and labels do not strongly influence the model's performance, suggesting that non-instruct LLMs cannot learn new information from the demonstrations. In contrast, our results suggest that instruction-tuned models are consistently influenced by in-context concept definitions.

Our evaluation protocol is similar to Wei et al. (2023) work as they use flipped and "semantically-unrelated" labels in task demonstrations. While they focus on in-context learning, our experiments are zero-shot tasks including *only concept definitions*. Thus, our setting is arguably harder, requiring the models to generalize from guidelines (not examples) that are relatively agnostic with respect to the classification task. Furthermore, we put a significant effort into evaluating open-source models that are state-of-the-art at the time of this submission. To our knowledge, this kind of evaluation is not addressed by previous work and is relevant to inform the improvement of open-source initiatives.

The potential of LLMs as zero-shot and few-shot data annotators has been demonstrated in medical (Agrawal et al., 2022), social science (Zhu et al., 2023), and other language understanding tasks (Wang et al., 2021). Our work provide further evidence that instruction-tuned models can perform concept classification with agreement scores comparable to expert annotators. Additionally, we show that similarly to humans, LLMs can leverage concept guidelines to improve the annotation quality.

6 Conclusion

By using factual and counterfactual concept guidelines for sentence classification, we demonstrate measurable gaps in concept understanding between leading open-source and proprietary instruction-tuned models. While some level of counterfactual concept understanding emerges with scaling, open-source models cannot recognize nonsensical (out-of-dictionary) guidelines, which the closed APIs can address more consistently. One question to be addressed in future work would be to investigate potential correlations between the capacity of reasoning in counterfactual contexts and other common generation issues such as hallucination.

Limitations

Opacity of Proprietary Models The experimental results from Section 4 confirm that the proprietary models excel in almost all classification settings. However, we cannot determine if the main cause for the best performance is the scale, training data, or fine-tuning methods since we do not have access to their implementation details.

Inference costs for Large Models Our experiments are severely limited by the computing requirements of the larger open-source LLM models. For instance, FALCON-180B-CHAT requires around 400GB of memory for inference, equivalent to 5 Nvidia A100-80GB GPUs. Thus, we limit our counterfactual guideline experiments (Figure 3) to include only a subset of possible permutations for LLAMA-2-7B and GPT-3.5.

Consequences of Counterfactual Performance to Other Tasks In this work, we measure the capacity of several language models to work under counterfactual contexts. Future investigation efforts could explore how this ability correlates to a potential reduction of hallucinations in generative tasks or even improved performance in natural language reasoning problems (Pan et al., 2023).

Financial Annotation Agreement Due to the ambiguity of financial annotation the task, the agreement scores we report in Section 3.2.3 are relatively moderate (average $\kappa = 0.47$ for the second round in Table 1). One of the main factors for disagreement is that some sentences are complex and may contain multiple capitals, as illustrated in the following example (passages conveying capitals are underlined):

Such factors include the duration and scope of the pandemic, including any resurgences of the pandemic, and the impact on our workforce and operations; the negative impact of the pandemic on the economy and economic activity, including travel restrictions and prolonged low demand for our products; the ability of our affiliates, suppliers and partners to successfully navigate the impacts of the pandemic; the actions taken by governments, businesses and individuals in response to the pandemic; the actions of OPEC and other countries that otherwise impact supply and demand and correspondingly, commodity prices; the extent and duration of recovery of economies and demand for our products after the pandemic subsides; and Chevron's ability to keep its cost model in line with changing demand for our products.

In many cases, annotators chose a non-

intersecting subset of the capitals, which counts as a disagreement (even though both are partially correct). Those voting ties were reviewed and adjudicated by the author of the guidelines. Previous scientific annotation projects like (Liakata et al., 2012) also report a moderate agreement ($\kappa = 0.55$, median of the best annotators), which demonstrates the difficulty in annotating technical documents.

Finally, even though report agreements between LLM annotations and humans in Table 3, our experiments are not designed to fairly compare annotation quality. Before annotation, humans received training and guidelines that are more comprehensive than LLM guidelines. Secondly, humans were able to "calibrate" their labels according to previously annotated sentences, whereas LLMs do not have access to this memory.

Acknowledgements

This work was supported by Actelligent Capital and used the Baskerville UK National Tier-2 HPC (https://www.baskerville.ac.uk) at the University of Birmingham. We also thank the anonymous reviewers for their insightful feedback.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of language models:towards open frontier models.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Mike Bennett. 2013. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3-4):255–268.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mandy Cheng, Wendy Green, Pieter Conradie, Noriyuki Konishi, and Andrea Romi. 2014. The international integrated reporting framework: key issues and future research opportunities. *Journal of International Financial Management & Accounting*, 25(1):90–119.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Maria Liakata and Larisa Soldatova. 2008. Guidelines for the annotation of general scientific concepts. *Aberystwyth University, JISC Project Report http://ie-repository. jisc. ac. uk/88*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Yasumasa Onoe, Michael JQ Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. *arXiv preprint arXiv:2305.01651*.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with

web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

SEC. 2014. SEC.gov | Filings & Forms. https://www.sec.gov/edgar. [Accessed 11-02-2024].

Larisa N Soldatova and Ross D King. 2006. An ontology of scientific experiments. *Journal of the royal society interface*, 3(11):795–803.

S&P Global. 2024. S&P 500®. https://www.spglobal.com/spdji/en/indices/equity/sp-500. [Accessed 11-02-2024].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Results with Human Guidelines

To complement the results in Section 4, we provide classification accuracies and agreement scores for guidelines using the same definitions provided to human annotators. We observe that models tend to ignore the concept definitions in favor of their prior knowledge about financial concepts, thus reducing the effects of our counterfactual guidelines. This effect is reflected in the more uniform accuracy results shown in Figure 5 (compared to Figure 2). The human guidelines also result in a slight increase in agreement with human annotators, as shown in Table 4 (compared to Table 3).

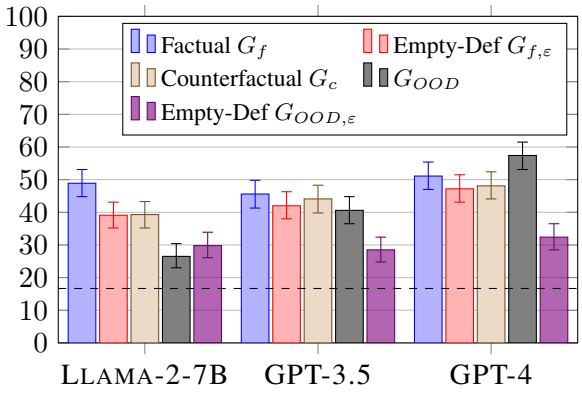

Figure 5: Concept classification accuracy for different **financial** concept guidelines, using the *same definitions provided to human labelers* (Figure 8). In this experiment, the counterfactual guideline G_c is a random permutation where *all concept definitions* are counterfactual. *Empty-Def* refers to the empty-definition factual ($G_{f,\varepsilon}$) and out-of-vocabulary guidelines ($G_{OOD,\varepsilon}$). Error bars represent the 95% confidence interval and the dashed line indicates the random classifier baseline.

B Concept Definitions

The Tables 5 and 6 present the definitions for scientific and financial concepts used in our experiments.

Model	Annotation Agreement			
	A_1	A_2	A_3	Avg
Human Average	0.46	0.43	**0.37**	0.42
LLAMA-2-7B	0.50	0.46	0.41	0.46
GPT-4	**0.51**	**0.57**	0.42	**0.50**

Table 4: Financial concept annotation agreement to annotators A_1, A_2, and A_3, using the *same definitions provided to human labelers* (Figure 8). Results are non-weighted Cohen's κ on a subset of sentences for which human annotators assigned at least one capital concept.

C Inference Prompts and Parameters

The prompt illustrated in Figure 1[12] consists of a guideline prompt $\texttt{prompt}_G$ and a task prompt $\texttt{prompt}_T$. The content of $\texttt{prompt}_G$ is a list of concept definitions:

> Consider the following concept categories:
> - {c_1}: {$\delta(c_1)$}
> ...
> - {c_K}: {$\delta(c_K)$}

where $\delta(c_K)$ is a function that maps the concept label c_K to its definition. Then, we define the content of $\texttt{prompt}_T$ as follows:

> Classify the text below into one of the categories listed above. Be concise and write only the category name.
>
> Text: {input sentence s}
> {domain} Concept:

where the placeholder {domain} is replaced with the text Scientific for scientific concepts and the empty string for financial concepts.

Finally, the prompts above are wrapped into model-specific prompts. For the LLAMA-2 models, we use the following prompt:

> [INST] {$\texttt{prompt}_G$}
>
> {$\texttt{prompt}_T$} [/INST]

And for FALCON-180B, we use the following prompt:

> User: {instruction} {$\texttt{prompt}_G$}
>
> {$\texttt{prompt}_T$}
> Falcon:

[12]The prompt in Figure 1 is simplified to improve readability.

Note that we do not use system prompts for both models, as we found that system prompts result in more verbose outputs. The main parameters for inference are provided in Table 7.

D Guidance Adherence Results

To complement the results in Figure 4, we provide the guidance adherence metrics for LLAMA-2-7B in Figure 6.

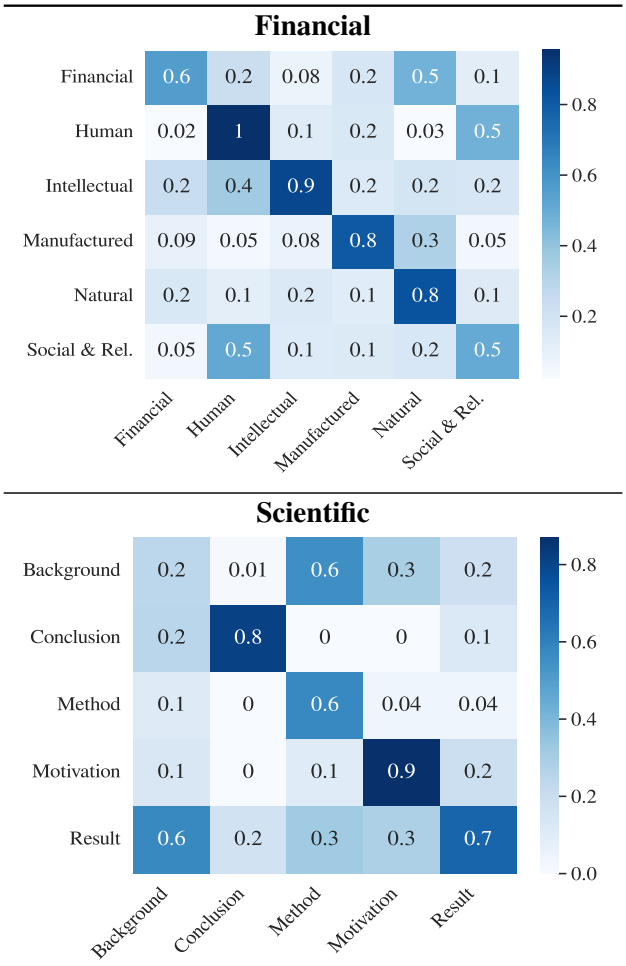

Figure 6: Guideline adherence scores per **financial (top)** and **scientific (bottom)** concept for **LLAMA-2-7B**. Each cell A_{ij} shows the fraction of concept predictions that adhere to concept definitions $\delta(c_j) = d_i$, where the rows indicate original factual labels c_i that are randomly replaced by labels c_j (columns). Off-diagonal results correspond to counterfactual concept definitions.

E Financial Annotation Details

Before engaging in the annotation task, the hired annotators were presented with the textual guidelines listed in Figure 8. The web-based annotation interface (Figure 7) is implemented using

Concept	Definition
Background	A sentence that provides context, foundational knowledge, or relevant information about the research topic, existing theories, prior studies, or the broader scientific field in which the research is situated. It helps readers understand the background against which the research is conducted.
Motivation	A sentence that explains the reasons, objectives, or goals behind the research. It often includes statements about the research gap, the problem being addressed, the significance of the study, and why the research is important.
Method	A sentence that describes the research methods, techniques, procedures, and data collection processes used in the study. This category also encompasses details about the experimental design, data analysis, and any materials or instruments utilized.
Result	A sentence that presents the empirical findings, outcomes, observations, or data generated by the research. It includes quantitative and qualitative results, statistical analyses, tables, figures, and any other information related to the research findings.
Conclusion	A sentence that summarizes the key takeaways, implications, interpretations, or insights derived from the study's results. It often discusses the broader significance of the findings, suggests future research directions, and may reiterate the study's contributions to the field.

Table 5: Scientific concept definitions used in sentence classification guidelines.

Concept	Definition
Financial	A sentence that pertains to monetary resources, assets, liabilities, revenues, expenses, or any other financial information related to the company's operations, investments, and financial performance.
Manufactured	A sentence that refers to physical assets, infrastructure, and tangible resources such as buildings, machinery, equipment, or any other manufactured or constructed items that contribute to the company's value.
Intellectual	A sentence that relates to intangible assets, knowledge, intellectual property, patents, trademarks, copyrights, research and development activities, or any other intellectual assets that enhance the company's competitiveness and innovation.
Human	A sentence that involves information about the company's workforce, including employees, skills, expertise, training, recruitment, talent development, and any other human resources aspects that contribute to the company's success.
Social and relationship	A sentence that deals with the company's relationships and interactions with external stakeholders, communities, customers, suppliers, partners, and any other social or relationship-based assets that affect the company's operations and reputation.
Natural	A sentence that addresses environmental resources, sustainability efforts, ecological impacts, conservation initiatives, or any other aspects related to the company's use of natural resources and its environmental responsibility.

Table 6: Financial concept definitions used in sentence classification guidelines.

LLAMA-2	
Number of parameters	7B / 13B / 70B
Max context length	4096
FALCON-180B-CHAT	
Number of parameters	180B
Max context length	2048
LLAMA-2 and **FALCON-180B-CHAT**	
Parameter type	float16
Nucleus temperature	0.8
Nucleus top-p	0.95
GPT-3.5 and **GPT-4**	
Model GPT-3.5	gpt-3.5-turbo-0613
Model GPT-4	gpt-4-0613
temperature	1
top_p	1
presence_penalty	0
frequency_penalty	0
All models	
Max generation tokens	128

Table 7: Summary of generation details and parameters.

Label Studio[13]. The interface shows the sample sentence and requests the annotator to classify it in one of the six capital concepts (`Financial`, `Manufactured`, `Intellectual`, `Human`, `Social and relationship`, and `Natural`) or `None` if the content is not related to any capital. The annotator also has the option to indicate a secondary capital, if applicable.

The annotation tasks were performed in 2021, when the UK minimum wage was 8.91 British Pounds[14], and annotators received 10 British Pounds per hour of work. The experiment design and conditions went through formal approval by an internal ethics committee. The data was not released or stored in public servers to avoid potential contamination.

[13] https://github.com/HumanSignal/label-studio

[14] https://www.gov.uk/national-minimum-wage-rates

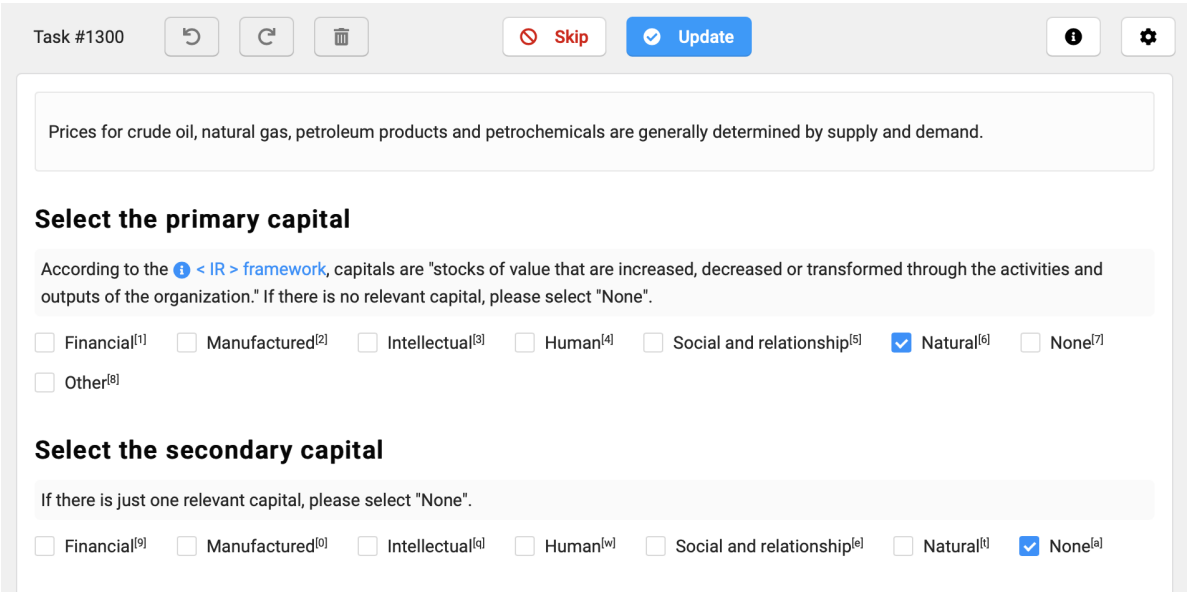

Figure 7: The annotation interface for the annotation of financial concepts. Given a sample sentence, annotators are requested to assign one of six capital concepts or None, if not applicable.

The concepts described in this section follow closely the definitions of the International <IR> framework, and should be sufficient to perform the annotation. According to the IR framework, capitals are "stocks of value that are increased, decreased or transformed through the activities and outputs of the organization." They can be classified in financial, manufactured, intellectual, social and relationship, and human (<IR> framework, section 2C).

Financial capital
The pool of funds that is available to an organization for use in the production of goods or the provision of services. It can be obtained through financing or generated through operations and investments. Example: *"The discussion also provides information about the financial results of our business segments to provide a better understanding of how those segments and their results affect the financial condition and results of operations of Ameren as a whole."*

Manufactured capital
Manufactured physical objects (excluding natural physical objects) that are available to an organization for use in the production of goods or the provision of services, including, buildings, equipment, and infrastructure (such as roads, ports, bridges, etc). Example: *"Due to the long lead time for the manufacture, repair, and installation of the components, the energy center is expected to return to service in late June or early July 2021."*

Intellectual capital
Organizational, knowledge-based intangibles, including: Intellectual property, such as patents, copyrights, software, rights and licences "Organizational capital" such as tacit knowledge, systems, procedures and protocols. Example: *"The absence of revenues from a software licensing agreement with Ameren Missouri decreased margins $5 million."*

Human capital
People's competencies, capabilities and experience, and their motivations to innovate, including their: 1) alignment with and support for an organization's governance framework, risk management approach, and ethical values; 2) sbility to understand, develop and implement an organization's strategy; 3) loyalties and motivations for improving processes, goods and services; 4) Other matters related to people management. Example: *"As the situation rapidly evolved, we remained focused on safely serving our customers and protecting the health and safety of our employees."*

Social and relationship capital
The institutions and the relationships within and between communities, groups of stakeholders and other networks, including: 1) shared norms, and common values and behaviours; 2) key stakeholder relationships, and the trust and willingness to engage that an organization has developed and strives to build and protect with external stakeholders; 3) intangibles associated with the brand and reputation that an organization has developed; 4) an organization's social licence to operate. Example: *"In March 2020, the MoPSC issued an order in Ameren Missouri's July 2019 electric service regulatory rate review, approving nonunanimous stipulation and agreements."*

Natural capital
All renewable and non-renewable environmental resources and processes that provide goods or services that support the past, current or future prosperity of an organization, including air, water, land, minerals, and biodiversity. Example: *"These amounts include the 700 MWs of wind generation projects discussed below, which will support Ameren Missouri's compliance with the state of Missouri's requirement of achieving 15% of native load sales from renewable energy sources beginning in 2021."*

Figure 8: Guidelines for financial concept annotation provided to human labelers.