Large Language Models as Financial Data Annotators: A Study on Effectiveness and Efficiency

Toyin Aguda*, **Suchetha Siddagangappa***, **Elena Kochkina**, **Simerjot Kaur**, **Dongsheng Wang**, **Charese Smiley**, **Sameena Shah**
J.P.Morgan AI Research
{toyin.d.aguda, suchetha.siddagangappa, elena.kochkina, simerjot.kaur, dongsheng.wang, charese.h.smiley, sameena.shah}@jpmchase.com

Abstract

Collecting labeled datasets in finance is challenging due to scarcity of domain experts and higher cost of employing them. While Large Language Models (LLMs) have demonstrated remarkable performance in data annotation tasks on general domain datasets, their effectiveness on domain specific datasets remains underexplored. To address this gap, we investigate the potential of LLMs as efficient data annotators for extracting relations in financial documents. We compare the annotations produced by three LLMs (GPT-4, PaLM 2, and MPT Instruct) against expert annotators and crowdworkers. We demonstrate that the current state-of-the-art LLMs can be sufficient alternatives to non-expert crowdworkers. We analyze models using various prompts and parameter settings and find that customizing the prompts for each relation group by providing specific examples belonging to those groups is paramount. Furthermore, we introduce a reliability index (LLM-RelIndex) used to identify outputs that may require expert attention. Finally, we perform an extensive time, cost and error analysis and provide recommendations for the collection and usage of automated annotations in domain-specific settings.

Keywords: Large Language Model, Data annotation, Finance

1. Introduction

Financial NLP (FinNLP) is an active and growing research area with numerous applications in analyzing and comprehending financial texts. The development of effective FinNLP models relies on well-annotated datasets derived from financial documents. However, annotating such datasets is challenging as it requires a deep understanding of financial concepts to decipher the complex terminologies and calculations present in the documents. Crowdsourcing platforms are generally used for annotations. While they are generally effective for tasks that do not require high levels of expertise, they often produce inconsistent and inaccurate annotations when it comes to domain-specific datasets. This approach requires careful instruction crafting, multiple annotation rounds, increased number of workers, and, finally, expert intervention for enhanced accuracy and consistency.

The wide array of tasks in which Large Language Models (LLMs), such as GPTs (Brown et al., 2020; OpenAI, 2023), have demonstrated state-of-the-art zero-shot capabilities naturally raises the question of whether these models have the potential to substitute for human annotators. Using LLMs as data annotators can offer a lot of advantages such as cost-effectiveness, scalability and potential for iterative improvement. However, strong performance on benchmark datasets alone does not ensure a model's suitability to replace human annotators. In addition to accuracy, consistency and biases associated with this approach needs to be carefully studied.

Text: The predecessor Mississippi Power Company was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on December 23, 1924 and in Alabama on December 7, 1962.

Relation type: Organization–Date

Expert Label: No/other relation
Crowdworker Label: Formed on

Figure 1: Example of relation extraction task from REFinD dataset.

While positive results of using LLMs as annotators for general-domain tasks have been reported in recent papers and preprints (Gilardi et al., 2023; Törnberg, 2023), their performance in specialized domains such as finance remains underexplored. In this work, we assess the efficacy of LLMs as data annotators for financial relation extraction task using REFinD dataset (Kaur et al., 2023).

The relation extraction task in financial documents involves identifying specific relations between financial entities such as companies and persons. Financial relation extraction presents unique challenges due to the domain-specific nature of financial language and the scarcity of labeled

*These authors contributed equally to this work

LREC-COLING 2024, pages 10124–10145
20-25 May, 2024.

data. General relation extraction models trained on generic tasks may lack the necessary understanding of finance-specific terms, leading to difficulties in capturing nuanced patterns. For example, certain relations, such as board membership versus employment, require domain expertise for accurate interpretation. Ambiguity further complicates the task, as implicit relationships, like company acquisitions based on stock ownership, may be challenging for generic models to identify. Furthermore, financial sentences are notably more complex, with longer average lengths and greater entity pair distances compared to generic domains, as demonstrated in REFinD.

Figure 1 shows an example from the REFinD dataset where we are interested in finding a relation between an ORGANIZATION-DATE entity pair, wherein we are interested in extracting the relation between an organization - *Mississippi Power Company* and date - *December 23, 1924*. For this entity pair, the relation label options presented to experts and crowdworkers are (i) FORMED ON (ii) ACQUIRED ON and (iii) NO/OTHER RELATIONS. The label chosen by experts is NO/OTHER RELATION, the reason being Mississippi Power Company was formed on November 24, 1924 and not on December 23, 1924. However, crowdworkers incorrectly identified the label as FORMED ON. This discrepancy between expert labels and crowdworker labels highlights the difficulty of financial relation extraction tasks.

In this work, we compare the output of LLMs and crowdworkers against expert annotations, extending our analysis beyond performance metrics and addressing time, cost and reliability aspects of the annotation process. Our contributions are the following: (i) To the best of our understanding, we are the first in the financial domain to demonstrate the capabilities of LLMs as data annotation tools by evaluating them against domain experts and crowdworkers. (ii) We compare 3 models (GPT-4, PaLM 2, and MPT Instruct) and parameters (varying temperature, random seed and prompting approaches) to identify the most accurate and reliable configuration. (iii) We introduce reliability index, a metric designed to identify trustworthy samples and filter out those requiring human intervention. (iv) We demonstrate that LLMs can replace non-expert crowdworkers for a significant portion of the dataset, while expert intervention is necessary for the remaining instances to ensure accurate annotations. We also offer guidance on best practices for implementing LLMs in the annotation process.

2. Related Work

Wang et al. (2021) pioneered the use of GPT-3 (Brown et al., 2020) as a cost-effective data labeler for training models. The potential of LLMs as data annotators has been explored in various tasks including relevance, stance, topic and frame classification (Gilardi et al., 2023), sentiment analysis, hate speech detection (Zhu et al., 2023; Huang et al., 2023), political affiliation (Törnberg, 2023) and news classification (Reiss, 2023)[1]. Since the majority of these tasks do not require the domain expertise of a human annotator, the effectiveness of LLMs in domain-specific datasets remains underexplored. This study investigates LLMs' potential in the financial domain.

Existing literature on the application of LLMs in the financial domain remains sparse. Li et al. (2023) have evaluated the performance of GPT-3.5 and GPT-4 on various finance benchmark datasets and reported strong performance on arithmetic reasoning, news classification and financial named entity recognition. However, this study did not consider the potential use of LLMs as annotators in comparison to non-expert crowdworkers, or the relation extraction task, which is the focus of our paper.

Several approaches assess the potential of LLMs as data annotators. Studies like Kuzman et al. (2023); Chiang and Lee (2023); Ding et al. (2023) explore different aspects of LLMs including comparing zero-shot performance of ChatGPT against a task-specific fine-tuned model, and measuring the alignment of LLM and human evaluations. He et al. (2023); Törnberg (2023); Gilardi et al. (2023) compare the model outcomes with crowdworkers and expert annotators. While the latter approach is more costly, we adopt it in this study due to its direct relevance to our research question.

LLMs as annotators yield mixed results, with some studies showing higher performance than humans (Gilardi et al., 2023; Törnberg, 2023), while others highlight limitations in new domains Zhu et al. (2023) and consistency issues (Reiss, 2023). Zhu et al. (2023) report GPT's overestimation of certain classes. This further motivates our study to evaluate these aspects for finance domain specifically.

It is also worth noting that most studies focus on GPT models only (Huang et al., 2023; Reiss, 2023; Törnberg, 2023; Zhu et al., 2023; Ding et al., 2023). We address this limitation by comparing three generative LLMs, GPT-4 (OpenAI, 2023), PaLM 2 (Anil et al., 2023), MPT Instruct (MosaicML, 2023), each with different size, training data, and procedures.

3. Dataset

Our experiments utilize the REFinD dataset (Kaur et al., 2023). Derived from texts within quarterly and annual reports of publicly traded companies (10-X), REFinD is the largest dataset available for financial relation extraction. This is also the only financial

[1]Note that some of the citations are recent publicly available preprints.

domain dataset for which we were able to obtain annotations broken down into expert and individual crowdworkers. REFinD dataset has 28,676 instances and 22 relations types across 8 entity pairs. The only other available Financial relation extraction dataset FinRED (Sharma et al., 2022) is significantly smaller (6,767 instances and 29 relation types) and does not release annotations provided by individual crowdworkers.

These 8 entity pairs covered in REFinD include PERSON–TITLE, PERSON–ORGANIZATION, PERSON–UNIVERSITY, PERSON–GOVERNMENT AGENCY, ORGANIZATION–GPE, ORGANIZATION–DATE, ORGANIZATION–ORGANIZATION and ORGANIZATION–MONEY. Each entity pair includes several finance-oriented relation types. The choice of this dataset is further justified by the fact that it was released in mid 2023, which makes it unlikely to have been part of the training data for the selected LLMs. For our experiments, we utilize 3598 instances from the test set of REFinD, due to the costs associated with LLMs usage(see section 5.5).

4. Experiments

In this section, we present comprehensive descriptions of the generative models, prompts, and evaluation metrics utilized in our study.

4.1. Models

In our experiments, we employed three Large Language Models (LLMs), GPT-4, PaLM 2, and MPT Instruct, selected based on their exceptional performance in benchmark leaderboards [2], accessibility, API availability, and permissive licenses. These models vary in size: GPT-4 comprises approximately 1.7 trillion parameters, PaLM 2 has 340 billion, and MPT Instruct is the smallest with 7 billion parameters. This diverse range enables us to evaluate the influence of model size on performance. For each model, we conducted experiments using two temperature settings (0.2 and 0.7) to examine the effects of randomness on model performance. Every model was run twice at each temperature setting. However, users cannot set a random seed for GPT-4 and PaLM 2, resulting in varying outputs between runs. In contrast, MPT Instruct was executed twice using two distinct random seeds.

Prompts

The quality of prompts used to guide LLMs significantly impacts their performance, akin to the instructions given to crowdworkers. We tailored the instruction around the prompt set up to focus on understanding the financial context around each question. Each input prompt comprises: (1) textual description of the task, (2) a sentence with highlighted entities[3], and (3) a numbered list of relation options (labels) specific to the entity pair. To avoid bias towards particular label orderings, we shuffle the option list. We experimented with 6 distinct prompt types which fall into 3 categories: zero-shot, few-shot and few-shot chain-of-thought (CoT) prompts. These prompts are based on the annotation instructions provided to MTurk[4] crowdworkers for the REFinD annotations (taken from Kaur et al. (2023)), facilitating a better comparison with their outputs.

For zero-shot prompts, we used: (1) *simple prompt*, a brief task description in basic English and (2) *full instruction prompt*, an extended version with a more comprehensive task description from the REFinD MTurk annotation instructions, an example of this is provided in Figure 2. Few-shot prompts, include: (3) *1-shot* and (4) *5-shot*, which build upon the full instruction prompt by adding a few task examples, tailored to the specific entity-pair type. Lastly, we experimented with few-shot CoT prompts: (5) *1-shot CoT* and (6) *5-shot CoT*. CoT prompts incorporates both the task descriptions and examples, as well as the reasoning behind each example's decision, as this approach has proven beneficial for other annotation tasks (Wei et al., 2022).

4.2. Evaluation

We assess the performance in comparison to expert annotators using accuracy and micro-averaged F1 scores. These metrics are calculated separately for each entity pair, and we report the mean average across entity pairs. Since each model's experiment is run twice, we also average these metrics from the two runs and report this as the final metric. Additionally, we measure the agreement between experiments, the time and cost of annotations, and the reliability index to analyze the efficiency and robustness of LLMs as annotators.

4.2.1. Inter Annotator Agreement (IAA)

We evaluate the agreement between different experiment settings to capture the model's self-consistency and assess the quality and reliability of the annotations. This metric demonstrates how uniformly annotators interpret the given task. To

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[3] We indicate the locations of both entities of interest by adding ** before and after entity1 and __ before and after entity2.

[4] https://docs.aws.amazon.com/pdfs/AWSMechTurk/latest/AWSMechanicalTurkRequester/amt-dg.pdf

Select date of formation relationship described in one sentence. Given a single sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962. With 2 highlighted phrases: Mississippi Power Company and December 23, 1924. Select a multiple choice answer from options below, which best describes the relation between Mississippi Power Company and December 23, 1924.

Please choose the MOST appropriate relation from the following options:

1. Mississippi Power Company is/was formed on December 23, 1924
2. Mississippi Power Company is/was acquired on December 23, 1924
3. no/other relation between Mississippi Power Company and December 23, 1924

Figure 2: Full instruction prompt example.

calculate the agreement between two annotators, we use Cohen's Kappa (Cohen, 1960b) and for agreement among more than two annotators, we use Fleiss' Kappa (Fleiss, 1971).

Reliability Index (LLM-RelIndex)

To aggregate the label for each sample from multiple annotators, we could simply calculate the raw voting counts for each label from K annotators. However, this approach has an issue when annotators all choose distinct labels, then an arbitrary label would be selected. As those distinct labels could be semantically related, such as MEMBER OF, EMPLOYEE OF and FOUNDER OF, incorporating such label similarity can improve the aggregation precision. Thus, we refine the voting approach by taking label similarity into account i.e., the similarities between its assessments a_i and each label l. The refined voting score, which considers the assessments of multiple annotators, measures the agreement for each label l as $\text{vote}(i, l) = sim(a_i, l)$. We then define the confidence as $\text{confid}(l) = \frac{1}{K}\sum_{i=1}^{K} \text{vote}(i, l)$. Note that similarity is defined as per the judgements of domain experts.

Additionally, we introduce the Reliability-Index, defined as the maximum confidence score $\text{confid}(l)$ of the label l:

$$\text{LLM-RelIndex}_i = \arg\max_{l \in L} \text{confid}(l) \quad (1)$$

The Reliability-Index aids in identifying the most reliable label for each instance. It enables the detection of outputs that warrant human expert attention.

Time & cost

For models served via API, the price per instance depends on the number of tokens (GPT-4[5]) or characters (PaLM 2[6]) in both the prompt and generated outputs. Consequently, the annotation cost was calculated by multiplying the average number of tokens/characters in the prompt and output, the number of instances, and the price per instance. For the open-source MPT-Instruct model, the cost was based on the per-hour price of the AWS machine utilized. Due to high GPU memory requirements, we used *p3.2xlarge* machines with 1 Tesla V100 GPU[7]. The annotation cost was calculated by multiplying the average time taken per instance in hours, the number of instances, and the price per hour.

5. Results

In this section, we discuss our experimental findings, focusing on model performance, annotator agreement, error analysis and reliability.

5.1. Model Performance

Table 1 presents the micro-averaged F1 score and accuracy for each LLM by prompt type and temperature setting, as well as the performance of MTurk annotators. We observe that GPT-4 and PaLM 2 significantly outperform crowdsourced annotations, with a margin of up to 29%. Both models exhibit comparable performance, with GPT-4 being the best. MPT Instruct demonstrates lower overall performance but still outperforms the human annotators in terms of F1-score when using *5-shot CoT prompt*. These results highlight the potential of LLMs as annotators. However, none of the models reach the expert performance, indicating that domain-specific settings still require expert's involvement. Figure 3 visualizes the results for the *full instruction prompt*, which is identical to the MTurk instructions.

Regarding the impact of prompt type on model performance, Table 1 reveals that the input prompt design significantly influences LLM performance. GPT-4 and PaLM 2 exhibit higher robustness under different prompts (5-7% difference), whereas

[5] https://openai.com/pricing, Accessed on 31/07/2023, GPT-4 8K context input price: $0.03/1K tokens, output price: $0.06/1K tokens. Number of tokens was calculated using the tiktoken package.

[6] https://cloud.google.com/vertex-ai/pricing, Accessed on 31/07/2023, PaLM 2 Text Bison: $0.0010/1K characters.

[7] https://aws.amazon.com/ec2/instance-types/p3/

Micro-Averaged F1 Score/ Accuracy(%)								
Annotator	Type	Temperature Setting	Zero-Shot Prompt		Few-Shot Prompt		Few-Shot CoT Prompt	
			simple prompt	full instruction	1-shot	5-shot	1-shot CoT	5-shot CoT
LLM	GPT-4	0.2	67.4/63.4	68.5/64.6	65.0/60.1	67.6/63.8	64.5/58.4	68.4/**65.4**
	GPT-4	0.7	**67.6/63.6**	68.4/64.6	65.0/60.0	67.7/63.9	64.6/**58.4**	68.4/**65.4**
	PaLM 2	0.2	62.3/53.9	62.2/53.8	66.4/60.1	66.0/59.2	64.7/55.9	65.6/57.2
	PaLM 2	0.7	64.5/56.0	64.4/56.0	**67.3/60.9**	**68.7**/63.8	64.9/57.4	65.9/59.2
	MPT Instruct	0.2	20.0/21.9	31.1/27.6	18.6/18.0	42.5/36.7	20.1/18.5	45.2/36.1
	MPT Instruct	0.7	20.8/24.7	24.8/27.3	22.7/24.2	30.5/31.1	22.2/23.2	33.9/30.8
	Ensemble (All LLMs)	0.2	65.2/60.1	66.0/60.7	63.9/58.1	68.1/63.3	63.3/56.4	**68.8**/63.8
	Ensemble (GPT-4 w PaLM 2)	0.2	67.2/63.2	**68.6/64.7**	65.0/60.1	67.8/**64.0**	64.3/58.1	68.2/65.2
	Ensemble (GPT-4 w MPT Instruct)	0.2	67.2/63.2	**68.6/64.7**	65.0/60.1	67.8/**64.0**	64.3/58.1	68.2/65.2
	Ensemble (PaLM 2 w MPT Instruct)	0.2	62.6/54.3	61.9/53.6	66.7/60.5	66.1/59.4	64.5/55.7	65.4/56.9
Human	Mturk Annotators	-	-	38.6/40.7	-	-	-	-

Table 1: Annotator performance in terms of micro-averaged F1-Score and accuracy against expert assigned labels.

prompt type has a strong effect on MPT Instruct performance (19%). MPT Instruct benefits considerably from additional examples (*5-shot* and *5-shot CoT*). Interestingly, few-shot and few-shot CoT prompts do not consistently outperform the zero-shot *full instruction prompt*. GPT-4 achieve its highest micro-averaged F1 score using the zero-shot *full instruction prompt*.

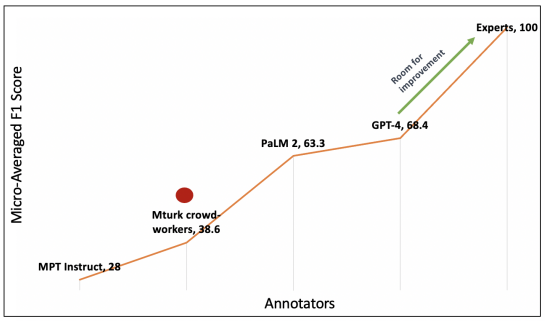

Figure 3: Annotator performance in terms of micro-averaged F1-Score under *full instruction prompt*.

Comparing performance at 0.2 and 0.7 temperature settings, we find that GPT-4 and PaLM 2 outputs remains stable regardless of the randomness introduced by the temperature parameter. While PaLM 2 consistently exhibits higher performance at 0.7, the observed performance differences are not statistically significant at the 0.05 significance level using a two-tailed t-test A.2.6. MPT Instruct performance is heavily affected by temperature settings, but no consistent pattern of superiority emerges for either setting. The highest scores are achieved at 0.2 with 5-shot example prompts.

Additionally, we evaluate the performance of an ensemble of models using a simple majority voting approach, which mimics having multiple annotators. While this approach results in the highest overall accuracy score, it does not consistently improve performance across all prompt types compared to a single model approach.

5.2. Inter-Annotator Agreement

High performance alone is insufficient for LLMs to serve as annotators, their output must also be consistent to be considered reliable. Therefore, we assess the consistency of the output by measuring agreement scores for models in different experiment settings shown in Table 3. First, we evaluate whether the models produce consistent outputs with the exact same parameters. For each experiment setting, we measure the IAA between the two runs of each model and then present an average score (row 1).

We observe that none of the models exactly replicate the outputs. GPT-4 and PaLM 2 exhibit high levels of agreement, while MPT runs with two different random seeds display significant differences. We then evaluate the agreement between outputs produced under two different temperature settings (row 2). GPT-4 agreement remains high even when varying the temperature parameter, while scores of PaLM 2 and MPT decrease. Furthermore, we compare the agreement between outputs produced using different prompts, both pairwise (using Cohen's Kappa, rows 3-5) and between the group of prompts (using Fleiss Kappa, row 6). We find that the choice of prompt has a more substantial impact on the outputs of the model, reducing the agreement for all LLMs. Overall, GPT-4 and PaLM 2 demonstrate reasonably high agreement across various experiment settings, indicating their overall reliability for the annotation task.

5.3. Error Analysis

In our error analysis, we aim to identify and categorize common issues encountered by LLMs during the annotation process. By examining instances with incorrect answers, hallucinated relations, and confident misannotations, we aim to gain insights into the challenges faced by LLMs and explore potential improvements for their performance in complex tasks, such as relation

LLM	Zero-Shot Prompt		Few-Shot Prompt		Few-Shot CoT Prompt	
	simple	full instruction	1-shot	5-shot	1-shot CoT	5-shot CoT
GPT-4	69.3	70.2	71	68.8	72.5	66.2
PaLM 2	74.5	73.8	74.9	76.1	79.8	80.7
MPT Instruct	46.4	52.5	48.4	57.5	49.7	64.9

Table 2: Proportion of LLM Hallucinations for instances labeled as NO/OTHER RELATION by experts

	GPT-4	PaLM 2	MPT
Random seed run1 vs run2	**0.95**	0.88	0.395
Temperature 0.2 vs 0.7	**0.95**	0.85	0.30
Zero-shot: simple vs full	0.87	**0.88**	0.39
Few-shot: 1- vs 5-shot	**0.84**	0.79	0.28
Few-shot CoT: 1- vs 5-shot	0.8	**0.82**	0.28
All prompts (Fleiss)	**0.83**	0.79	0.31

Table 3: Pairwise IAA in terms of Cohen Kappa (top 5 rows) and IAA between outputs for all prompts in terms of Fleiss Kappa (last row). First two rows present mean averaged values of pairwise Cohen Kappa for each prompt type.

extraction.

5.3.1. Semantic Ambiguity

We analyze instances where LLMs return incorrect answers and observe that these errors often stem from the proximity and similarity of the answer options, causing confusion in identifying the most accurate response. Common trends include MEMBER OF instead of EMPLOYEE OF and FORMED IN rather than OPERATIONS IN. This highlights the need to improve LLM's comprehension of subtle differences. For instance, in the example "W. Howard Keenan , Jr. has served as a director of Midstream Management since February 2014", both GPT-4 and PaLM 2 incorrectly choose MEMBER OF over the correct relation EMPLOYEE OF. Although MPT Instruct's result is also inaccurate, its answer varies significantly by prompt type, exhibiting a level of randomness not observed in the other two LLMs. Its also worth noting that MPT Instruct returns blanks for some instances. 0.5% of the responses from MPT Instruct for each prompt variation were blanks.

5.3.2. Relation Hallucinations

In our relation extraction task, we provide the LLMs with limited label options, including an option for NO/OTHER RELATION available for every entity pair. Consequently, we expect minimal instances of hallucinations, i.e., LLMs inventing new relations between specified entities not present in the label set or generating off-topic responses. We analyze the LLM outputs for instances labeled as NO/OTHER RELATION by the experts and report the proportion of hallucinations among them (Table 2). We observe that hallucinations primarily emerge from PaLM 2 for *5-shot CoT*, where 80.7% of instances labeled as NO/OTHER RELATION by the experts were misidentified by PaLM 2 as hallucinations. Overall, LLMs exhibit a higher tendency to generate new relations when the expert label is NO/OTHER RELATION. GPT-4 and PaLM 2 tend to hallucinate more than MPT Instruct. We post-process the hallucinated relations to extract relation styles similar to those in the label options. The most common relations extracted from these are AGREEMENT WITH, SHARES OF, MEMBER OF and SUBSIDIARY OF.

5.3.3. Confident Misannotations

We analyze instances where LLMs and crowdworkers return incorrect answers with high confidence (answers selected by majority of annotators). The relationship between high confidence and incorrect answer choice varies, and we observe three scenarios: (i) the majority of crowdworker labels are incorrect while the majority of LLM labels are correct, (ii) the majority of both crowdworker and LLM annotations are incorrect, and (iii) the majority of crowdworker labels are correct while the majority of LLM labels are incorrect. Qualitative examples of these can be found in Figure 4. This analysis demonstrates that the varying dynamics between LLMs and crowdworkers emphasize the importance of refining LLMs to better understand nuanced distinctions and improve their reliability in annotation tasks. Furthermore, the analysis highlights the potential benefits of combining the expertise of both LLMs and human annotators to achieve more accurate and reliable annotations in complex tasks, such as relation extraction.

5.4. LLM-RelIndex Based Accuracy Analysis

In this analysis, we employ the LLM-RelIndex majority voting scheme to assess the accuracy derived from human votes and LLM results across all six prompt variations on the dataset. The data is arranged in descending order of LLM-RelIndex that is we moved from instances that were simple to annotate to the more complex ones and we present the accuracy for incremental percentages of the dataset. We showcase the plots for three distinct cases: (i) zero-shot (Figure 5), (ii) few-shot (Figure 6), and (iii) few-shot CoT (Figure 7).

Our observations indicate that for all three cases,

Scenario 1 (Crowdworkers incorrect, LLMs correct):
Instance: Personal Lines underwriting profit for the three months ended September 30, 2017 was $ 40.8 million, compared to $ 23.3 million for the three months ended September 30, 2016 , an improvement of $ 17.5 million.

Expert Label: PROFIT OF
Crowdworker Label: PROFIT OF, NO/OTHER RELATION , LOSS OF
LLMs Label: PROFIT OF

Scenario 2 (Crowdworkers and LLMs incorrect):
Instance: Our Hawaii Gas entered into licensing agreements with Utility Service Partners , Inc. and America's Water Heater Rentals , LLC , both indirect subsidiaries of Macquarie Group Limited , to enable these entities to offer products and services to Hawaii Gas's customer base.

Expert Label: SUBSIDIARY OF
Crowdworker Label: NO/OTHER RELATION, SUBSIDIARY OF, SHARES OF
LLMs Label: AGREEMENT WITH

Scenario 3 (Crowdworkers correct, LLMs incorrect):
Instance: On December 10, 2014 , Orbital Tracking Corp. purchased certain contracts from Global Telesat Corp , a Virginia corporation (GTC) for $ 250,000 pursuant to an asset purchase agreement by and among Orbital Tracking Corp i, its wholly owned subsidiary Orbital Satcom, GTC and World Surveillance Group , Inc. (World) , GTC's parent.

Expert Label: SUBSIDIARY OF
Crowdworker Label: SUBSIDIARY OF
LLMs Label: AGREEMENT WITH

Figure 4: Error Analysis: Qualitative examples illustrating different scenarios of how MTurk Crowdworkers and LLMs demonstrated high confidence on incorrect answer choices.

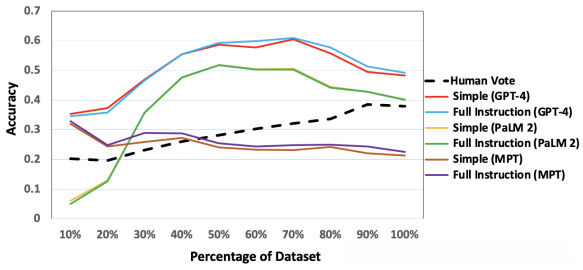

Figure 5: Human vs LLMs at Zero-shot using *LLM-RelIndex*

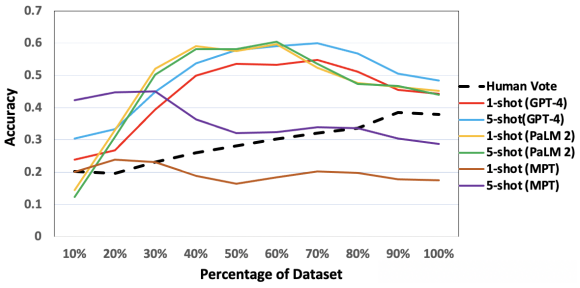

Figure 6: Human vs LLMs at Few-shot using *LLM-RelIndex*

GPT-4 and PaLM 2 outperform both Human Votes and MPT Instruct when considering ∼65% of the dataset. However we also observed a drop in accuracy in the top 20% of the dataset where there were high level agreements among LLMs. This can be attributed to the instances which were simple to annotate but easier to error on. Hence we observe that in those instances most of the LLMs made the same mistakes as human annotators which were inconsistent with expert choices. For example "The number of shares that are sold by Cowen after delivering a sales notice will fluctuate based on the market price of Dermira, Inc common stock during the sales period and limits Dermira, Inc. set with Cowen." Most of the LLMs chose AGREEMENT WITH over SHARES OF where the latter is the correct relation.

Additionally, PaLM 2's performance exhibits an upward trend, as we transition from zero-shot to few-shot, and ultimately to few-shot CoT scenarios. We also find that all LLMs demonstrate improved results for *5-shot* and *5-shot CoT*, suggesting that having more examples and explanations enhances the reliability of LLM-generated annotations.

As we progress towards complete dataset coverage, again we see a decline in performance is noted. This outcome is anticipated since instances with lower LLM-RelIndex scores become more prevalent as we approach more complex instances. Here the LLMs likely lack confidence in relations between specific entity pairs.

Overall, LLM-RelIndex allows us to confidently assert that LLMs can serve as more reliable annotators for ∼65% of this dataset. For cases beyond this threshold, expert intervention is necessary to determine the appropriate annotation. This strategy effectively reduces the cost and time associated with human annotation of the entire dataset, streamlining the process considerably.

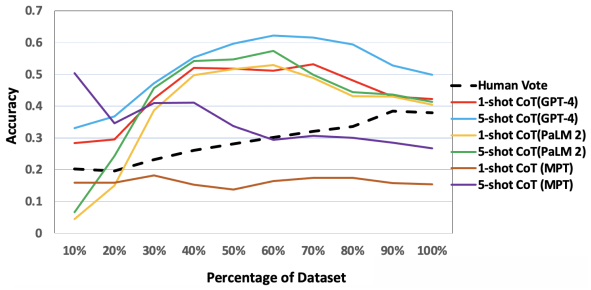

Figure 7: Human vs LLMs at Few-shot CoT using *LLM-RelIndex*

5.5. Time and Cost Analysis

We calculate the annotation cost for each of the LLMs (detailed in Evaluation section) and compare it to our estimated cost of MTurk annotations. The average input prompt size ranges from 191 tokens (814 characters) for *simple prompts* to 441 tokens (1954 characters) for *5-shot CoT prompts*. On average, GPT-4 generates an output of 17 tokens (65 characters) for all prompt types. The outputs of PaLM 2 vary more in size, from 70 to 36 tokens (298 and 147 characters), with shorter outputs for longer prompts like few-shot and few-shot CoT. Each model can process an instance within 1-5 seconds, with longer prompts requiring more processing time. MPT Inference on average takes 0.96 seconds for *simple prompts* and 1.81 for the longest *5-shot CoT prompts*. The annotation price increases with the prompt size, and for our dataset of 3598 instances, it ranges from $24-51 for GPT-4, $5-9 for PaLM 2 and $29-55 for MPT Instruct.

For crowdsourced human annotators, the time and associated cost would be higher. Assuming a human annotator takes 45 seconds per instance and is paid the US minimum wage of $7.25 per hour[8], the dataset's annotation cost using a single annotator amounts to $389. However, the crowdsourced annotation process typically involves multiple annotators per instance. These outcomes demonstrate that automated annotations are more efficient in terms of time and cost compared to human labelling

6. Discussion

In this section we discuss our findings and share recommendations for future annotation tasks. Our experiments demonstrated the potential of LLMs as data annotators for tasks within the financial domain. Specifically, GPT-4 and PaLM 2 have exhibited exceptional performance, surpassing the accuracy of the non-expert crowdworkers, while delivering time and cost savings. PaLM 2 has achieved comparable results to GPT-4, despite its smaller size, at a fraction of the cost (∼5 times less). These models have also displayed robustness by producing consistent outputs across various parameter and prompt configurations. However, it is crucial to recognize that LLMs' performance does not yet match that of domain experts and expert involvement remains necessary for obtaining high-quality annotations with minimal or no noise.

The next generation of annotation approaches in domain-specific contexts should consider adopting a hybrid strategy, harnessing both automated and expert-generated annotations to optimize results. In these settings, approximating model uncertainty, e.g., via the LLM-RelIndex, can help prioritize instances that require expert attention. In all annotation tasks, the ability to formulate detailed instructions is a vital factor, regardless of whether annotators are human or LLMs. Carefully crafting prompts, guided by an understanding of the task and the specific LLM being used helps optimize the outputs generated by the LLMs.

We, therefore, recommend that researchers conduct small preliminary experiment on a data subset to assess model capabilities and identify optimal parameter and prompt configurations. The specifics of the task should inform researchers about the *tolerance for annotation noise*, allowing them to train new models using automatically annotated data accordingly. Moreover, future annotation tasks can benefit from more open task formulations, leveraging the generative abilities of LLMs. For instance, in our task, LLMs have the potential to help identify more relations than the original pre-defined set. As such, future experiment can be done to check if these LLM-annotated data boost downstream performances. Lastly, it is essential to remain mindful that model biases may differ from those of crowdworkers and to account for these differences where necessary.

7. Limitations

One of the main limitations of this work is that the evaluation is performed only on a single dataset, covering a single task. The dataset contains the texts from one particular source, SEC filings, and it would be interesting to compare the results when the texts come from other financial sources, such as news or earning calls. This limitation partially comes from the costs of using the LLMs, and partially from the absence of financial datasets with annotations produced by individual crowdworkers released publicly.

In this work we present the breakdown of the results and their analysis by relation categories in the Appendix due to the page limit. We found that model performance varies strongly between the

[8] https://www.dol.gov/agencies/whd/minimum-wage, Accessed on 31/07/2023.

entity pair groups similar to Kaur et al. (2023) with ORGANIZATION-ORGANIZATION being the most challenging category. In future work, we aim to expand our analysis further with respect to categories of errors frequently associated with this task and financial domain such as numerical inference, semantic and directional ambiguity.

We observe that our LLM-RelIndex metric is subject to error, particularly with instances that are easy to annotate. Efforts are underway to enhance this metric. Furthermore, we are exploring the adoption of an automated and systematic approach for calculating similarity scores rather than depending on experts' judgment. Additionally, we intend to incorporate multi-label samples into our approach, given that some similar labels may closely align for some cases.

Finally, while providing the discussion, we do not experimentally demonstrate how the automatically annotated dataset can be used, either to improve relation extraction model performance, or to develop smaller efficient models. We recognize the importance of this and leave this to future work.

8. Conclusion

In this study, we have showcased the remarkable potential of using LLMs as a robust alternative to non-expert crowdworkers for domain-specific task by comparing three LLMs of varying sizes. Due to large volume of unstructured documents within financial domain, leveraging LLMs for annotations significantly reduces the time spent by humans on manual annotation, while providing valuable insights for making well-informed downstream decisions and driving efficient business outcomes. Our evaluation shows that larger models like GPT-4 and PaLM 2 excel in these tasks, while incorporating more examples into prompts for smaller models like MPT Instruct can yield improved results. We also introduced the reliability index, a metric that identifies reliable labels and detects outputs requiring expert attention, enhancing quality control and decision-making. Our error analysis provides valuable insights for future improvements.

The integration of LLMs streamlines the annotation process, delivering consistent, high-quality outputs that result in substantial time savings and cost-effectiveness. However, their performance does not yet match that of experts who possess a nuanced understanding of the subject matter. While LLMs offer scalability and reduced time and costs compared to employing experts, there exists a trade-off between the convenience and efficiency of LLMs and the precision provided by expert annotators. Consequently, the decision to employ LLMs as annotators should be carefully guided by the desired level of accuracy and the complexity of the task at hand, striking the right balance between automation and human expertise.

9. Acknowledgments

We would like to thank Armineh Nourbakhsh, Natraj Raman, Xiaomo Liu, Manuela Veloso, and our anonymous reviewers for their thoughtful comments and feedback which greatly contributed to the quality of this work.

10. Bibliographical References

Meysam Alizadeh, Fabrizio Gilardi, Emma Hoes, K Jonathan Klüser, Mael Kubli, and Nahema Marchal. 2022. Content moderation as a political issue: The twitter discourse around trump's ban. *Journal of Quantitative Description: Digital Media*, 2.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2021. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 workshop on open challenges in online social networks*, pages 27–32.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2):103200.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Cohen. 1960a. A coefficient of agreement for nominal scales. educational and psychological measurement, 20(1), 37–46. Accessed: 2023-07-24.

Jacob Cohen. 1960b. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fabrizio Gilardi, Meysam Alizadeh, and Ma´l Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. Annollm: Making large language models to be better crowdsourced annotators.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. SIGIR '23, Taipei, Taiwan. Association for Computing Machinery.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Nelson Liu, Tony Lee, Robin Jia, and Percy Liang. 2021. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches.

NLP MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26837–26867. PMLR.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597.

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023. Beyond classification: Financial reasoning in state-of-the-art language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 34–44, Macao. -.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Livia Van Vliet, Petter Törnberg, and Justus Uitermark. 2020. The twitter parliamentarian database: Analyzing twitter politics across 26 countries. *PLoS one*, 15(9):e0237073.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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. Supplementary Materials

A.1. Ethical Considerations

This paper explores the use of LLMs for data annotation. As such, the prevailing concerns around the use of LLMs apply to this work. This includes the potential to generate text containing bias, stereotypes, misinformation and, as noted in the discussion, hallucinations. Outside of issues concerning LLM usage, we do not anticipate other ethical concerns with this work.

A.2. Appendices

A.2.1. Dataset relation distribution

Entity-Pair	No. of Instances
ORG-GPE	710
ORG-ORG	913
ORG-DATE	554
ORG-MONEY	281
PER-ORG	485
PER-TITLE	655
Total	3598

Table 4: Dataset Relation Distribution

A.2.2. Metrics for MTurk Annotators

Micro F1 Score/ Accuracy (%)	
Entity Pair	**MTurk Annotators**
ORG-GPE	37.3/35.8
ORG-ORG	13.5/21.6
ORG-DATE	31.4/45.0
ORG-MONEY	26.4/29.1
PER-ORG	33.9/32.3
PER-TITLE	89.0/80.4
Total	38.6/40.7

Table 5: MTurk Annotator Micro-average F1 Score/Accuracy by Entity Pair

A.2.3. LLM Setup and Configuration

The setup and configuration of each LLMs have some overlap such as specifying the location of each entity in the text, however, there are notable differences as well. These differences enabled each LLM to perform at its best. Figure 1 explains the different piece of LLM setup and configuration. Unlike GPT-4, where we had "system role", in PaLM 2 we had "Additional Instruction". This is the unique prompt design for the different prompt type.

Setup & Configuration: ORG-DATE
INSTRUCTION:Select the statement that best describes the relation in the example sentence below. Ignore any grammatical errors. If there are multiple options, please choose the one that is clearest and most obvious from the sentence.

PROMPT: The predecessor Mississippi Power Company was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on December 23, 1924 and in Alabama on December 7, 1962.

PROMPT+: Please choose the MOST appropriate relation from the following options:

1. Entity1 is/was formed on Entity2.
2. Entity1 is/was acquired on Entity2.
3. No/other relation between Entity1 and Entity2.

SYSTEM ROLE: You are an AI assistant and relation extraction checker. You read the prompt, note where the entities in question are and determine the relation between them. Once done, please select from option which best suite the relation.

GPT-4 setup follows: Using the context from "setup piece and configuration"
Zero-shot:
This starts with **PROMPT**, followed by **PROMPT+** and finally **SYSTEM ROLE** and **RESPONSE**.
Few-shot:
This starts with **INSTRUCTION**, followed by **PROMPT WITH EXAMPLE(S)**, **PROMPT+** and finally **SYSTEM ROLE** and **RESPONSE**.
Few-shot CoT:
This starts with **INSTRUCTION** followed by **PROMPT WITH EXAMPLE(S)**, **REASONING**, **PROMPT+** and finally **SYSTEM ROLE** and **RESPONSE**.

PaLM 2 setup follows: Using the context from "setup piece and configuration"
Zero-shot:
This starts with **SYSTEM ROLE** called **ADDITIONAL INSTRUCTION** in PaLM 2, followed by **PROMPT** and finally **PROMPT+** and **RESPONSE**.
Few-shot:
This starts with **SYSTEM ROLE** called **ADDITIONAL INSTRUCTION** followed by **INSTRUCTION**, **PROMPT WITH EXAMPLE(S)**, and finally **PROMPT+** and **RESPONSE**.
Few-shot CoT:
This starts with **INSTRUCTION** followed by **PROMPT WITH EXAMPLE(S)**, **REASONING**, **PROMPT+** and finally **SYSTEM ROLE** and **RESPONSE**.

MPT Instruct setup follows: Using the context from "setup piece and configuration"
Zero-shot:
This starts with **SYSTEM ROLE** also called **INSTRUCTION** in MPT Instruct, followed by **PROMPT** and finally **PROMPT+** and **RESPONSE**.
Few-shot:
This starts with **SYSTEM ROLE** called **INSTRUCTION**, followed by **PROMPT WITH EXAMPLE(S)**, and finally **PROMPT+** and **RESPONSE**.
Few-shot CoT:
This starts with **INSTRUCTION**, followed by **PROMPT WITH EXAMPLE(S)**, **REASONING**, **PROMPT+** and finally **SYSTEM ROLE**.

Figure 8: LLM Setup and Configuration.

A.2.4. Prompt Description

Title	Prompt style based on LLM setup
Simple Prompt	In the context of this sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962 . Note the location of the Mississippi Power Company and December 23, 1924 as highlighted to help determine the relation given the listed options below. Please choose the MOST appropriate relation from the following options: 1. Mississippi Power Company is/was acquired on December 23, 1924. 2. Mississippi Power Company is/was formed on December 23, 1924. 3. no/other relation between Mississippi Power Company and December 23, 1924.
Full Instruction Prompt	Select date of formation relationship described in one sentence. Given a single sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962. With 2 highlighted phrases:Mississippi Power Company and December 23, 1924, select a multiple choice answer from options below, which best describes the relation between Mississippi Power Company and December 23, 1924. Please choose the MOST appropriate relation from the following options: 1. Mississippi Power Company is/was formed on December 23, 1924. 2. Mississippi Power Company is/was acquired on December 23, 1924. 3. no/other relation between Mississippi Power Company and December 23, 1924.
1-Shot Prompt	Select the statement that best describes the relation in the example sentence below. Ignore any grammatical errors. If there are multiple options, please choose the one that is clearest and most obvious from the sentence. \n\nExample Sentence 1:**LecTec** was organized in 1977 as a Minnesota corporation and went public in __December 1986__. \n Answer to Example 1: LecTec was formed/incorporated on/in December 1986. \n Following the example above, read through this sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962 . Given the location of the Mississippi Power Company and December 23, 1924 as highlighted, choose an answer from listed options below. \n Please choose the MOST appropriate relation from the following options: \n 1. Mississippi Power Company is/was acquired on December 23, 1924\n 2. Mississippi Power Company is/was formed on December 23, 1924\n 3. no/other relation between Mississippi Power Company and December 23, 1924.
5-Shot Prompt	Select the statement that best describes the relation in the example sentence below. Ignore any grammatical errors. If there are multiple options, please choose the one that is clearest and most obvious from the sentence. \n\n Example Sentence 1:**LecTec** was organized in 1977 as a Minnesota corporation and went public in __December 1986__. \n Answer to Example 1: LecTec was formed/incorporated on/in December 1986. \n Example Sentence 2: The assets of **Unified Payments , LLC** were acquired by us in __April 2013__.\n Answer to Example 2: Unified Payments, LLC was acquired in April 2013. \n Example Sentence 3: Since __July 6, 2016__ , Pinnacle West has issued four parental guarantees for 4CA relating to payment obligations arising from 4CA s acquisition of El Paso s 7 % interest in **Four Corners** , and pursuant to the Four Corners participation agreement payment obligations arising from 4CA s ownership interest in Four Corners. \n Answer to Example 3: No relation between Four Corners and July 6 , 2016. \n Example Sentence 4: In__ 2014__ , $ 148 million cash proceeds , net of cash sold , from Sempra Renewables sale of 50 - percent equity interests in **Copper Mountain Solar 3** ($ 66 million) and Broken Bow 2 Wind ($ 58 million) , and Sempra Mexico s sale of a 50 - percent equity interest in Energ a Sierra Ju rez ($ 24 million) ; and .\n Answer to Example 4: No relation between Copper Mountain Solar 3 and 2014.\n Example Sentence 5: **Zendex** was incorporated in the state of Utah in __March 2011__ to create an online platform for the sale of art .\n Answer to Example 5:Zendex was formed in March 2011. \n\n Following the example above, read through this sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962 . Given the location of the Mississippi Power Company and December 23, 1924 as highlighted, choose an answer from listed options below. \n Please choose the MOST appropriate relation from the following options: \n 1. Mississippi Power Company is/was formed on December 23, 1924\n 2. Mississippi Power Company is/was acquired on/in December 23, 1924\n 3. no/other relation between Mississippi Power Company and December 23, 1924.

Title	Prompt style based on LLM setup
1-Shot CoT Prompt	Select the statement that best describes the relation in the example sentence below. Ignore any grammatical errors. If there are multiple options, please choose the one that is clearest and most obvious from the sentence. \n\n Example Sentence 1:**LecTec** was organized in __1977__ as a Minnesota corporation and went public in December 1986. \n Answer to Example 1: LecTec was formed/incorporated on/in 1977. \n The reasoning for the above answer is that the highlighted portion of the question, LecTec, corresponds with the entity being discussed, and the year 1977 refers to when LecTec was organized or incorporated, both of which are accurately reflected in the answer.\n Following the example above, read through this sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962 . Given the location of the Mississippi Power Company and December 23, 1924 as highlighted, choose an answer from listed options below. \n Please choose the MOST appropriate relation from the following options: \n 1. Mississippi Power Company is/was acquired on December 23, 1924\n 2. Mississippi Power Company is/was formed on December 23, 1924\n 3. no/other relation between Mississippi Power Company and December 23, 1924.
5-Shot CoT Prompt	Select the statement that best describes the relation in the example sentence below. Ignore any grammatical errors. If there are multiple options, please choose the one that is clearest and most obvious from the sentence. \n\n Example Sentence 1:**LecTec** was organized in __1977__ as a Minnesota corporation and went public in December 1986. \n Answer to Example 1: LecTec was formed/incorporated on/in 1977. \n The reasoning for the above answer is that the highlighted portion of the question, LecTec, corresponds with the entity being discussed, and the year 1977 refers to when LecTec was organized or incorporated, both of which are accurately reflected in the answer. \n Example Sentence 2: The assets of **Unified Payments , LLC** were acquired by us in __April 2013__.\n Answer to Example 2: Unified Payments, LLC was acquired in April 2013. \n The reasoning for the answer above is that the highlighted portions of the question indicate the key elements of the event being asked about: Unified Payments, LLC being the entity that was acquired and April 2013 being the time when the acquisition took place, both of which are directly stated in the answer. \n Example Sentence 3: Since __July 6, 2016__ , Pinnacle West has issued four parental guarantees for 4CA relating to payment obligations arising from 4CA s acquisition of El Paso s 7 % interest in **Four Corners** , and pursuant to the Four Corners participation agreement payment obligations arising from 4CA s ownership interest in Four Corners. \n Answer to Example 3: No relation between Four Corners and July 6, 2016. \n We are only interested in identifying if the organization mentioned was formed on the specified date or acquired by another organization on the specified date. Since Four Corners was neither formed on July 6, 2016 nor acquired by another company on July 6, 2016, there is no relation between Four Corners and July 6, 2016.\n Example Sentence 4: In__ 2014__ , $ 148 million cash proceeds , net of cash sold , from Sempra Renewables sale of 50 - percent equity interests in **Copper Mountain Solar 3** ($ 66 million) and Broken Bow 2 Wind ($ 58 million) , and Sempra Mexico s sale of a 50 - percent equity interest in Energ a Sierra Ju rez ($ 24 million) ; and .\n Answer to Example 4: No relation between Copper Mountain Solar 3 and 2014.\n We are only interested in identifying if the organization mentioned was formed on the specified date or acquired by another organization on the specified date. Since Copper Mountain Solar 3 was neither formed in 2014 nor acquired by another company in 2014, there is no relation between Copper Mountain Solar 3 and 2014. \n Example Sentence 5: **Zendex** was incorporated in the state of Utah in __March 2011__ to create an online platform for the sale of art .\n Answer to Example 5:Zendex was formed in March 2011. \n\n The incorporation of Zendex in March 2011 suggests that this is the official date when the company was legally established and recognized as a corporate entity in the state of Utah. Hence Zendex was formed on March 2011. \n Following the example above, read through this sentence: The predecessor **Mississippi Power Company** was incorporated under the laws of the State of Maine on November 24, 1924 and was admitted to do business in Mississippi on __December 23, 1924__ and in Alabama on December 7, 1962 . Given the location of the Mississippi Power Company and December 23, 1924 as highlighted, choose an answer from listed options below. \n Please choose the MOST appropriate relation from the following options: \n 1. Mississippi Power Company is/was formed on December 23, 1924\n 2. Mississippi Power Company is/was acquired on/in December 23, 1924\n 3. no/other relation between Mississippi Power Company and December 23, 1924.

Table 6: Prompts for Entity-Pair: ORG-DATE

A.2.5. Metrics for LLM Annotators

RUN 1: Micro F1 Score / Accuracy(%)									
LLM	**Annotator**	**Annotator Description**	**Entity-Pair**						**Total**
			ORG-GPE	**ORG-ORG**	**ORG-DATE**	**ORG-MONEY**	**PER-ORG**	**PER-TITLE**	
GPT-4	annotator1	simple prompt, temp=0.2	80.1/74.8	15.4/38.4	48.9/67.0	48.4/43.4	70.8/67.8	92.7/86.9	67.2/63.2
	annotator2	simple prompt, temp=0.7	80.3/74.6	15.1/37.1	51.6/70.0	**48.9/44.5**	72.1/69.3	93.2/87.6	67.8/63.7
	annotator3	full instruction, temp=0.2	80.9/75.8	15.7/36.7	56.3/74.5	47.8/43.1	**72.8/69.9**	93.7/88.7	**68.6/64.7**
	annotator4	full instruction, temp=0.7	**80.9/75.8**	15.5/36.9	55.4/73.8	47.6/42.7	71.8/69.3	93.5/88.2	68.2/64.4
	annotator5	1-shot, temp=0.2	79.6/73.9	15.4/30.1	48.0/65.7	47.4/42.0	62.5/60.0	94.4/89.9	65.0/60.1
	annotator6	1-shot, temp=0.7	79.8/73.9	14.5/28.9	50.3/68.1	48.0/42.3	62.5/60.0	94.3/89.8	65.1/60.1
	annotator7	5-shot, temp=0.2	79.3/73.9	15.7/35.5	59.3/78.0	48.0/41.3	71.1/67.8	93.3/87.9	67.8/64.0
	annotator8	5-shot, temp=0.7	78.9/73.5	15.0/35.3	58.9/77.6	47.1/40.2	72.0/69.1	93.2/87.8	67.6/63.8
	annotator9	COT 1-shot, temp=0.2	79.6/74.1	16.1/32.5	36.8/47.3	48.9/43.4	63.7/61.0	94.4/89.8	64.3/58.1
	annotator10	COT 1-shot, temp=0.7	80.2/74.6	16.2/32.7	37.8/48.9	47.5/41.3	63.7/60.8	**94.7/90.2**	64.6/58.4
	annotator11	COT 5-shot, temp=0.2	79.4/73.7	16.2/37.8	**65.4/83.2**	46.3/42.7	70.6/67.6	92.8/87.0	68.2/65.2
	annotator12	COT 5-shot, temp=0.7	79.6/73.8	**17.0/38.4**	65.2/83.0	46.0/43.1	70.9/67.8	92.9/87.0	68.4/65.5
PaLM 2	annotator1	simple prompt, temp=0.2	81.0/76.9	**13.5**/14.7	50.1/67.0	43.5/29.2	68.3/62.9	87.2/78.5	62.6/54.3
	annotator2	simple prompt, temp=0.7	80.0/76.1	13.5/13.3	49.3/65.9	43.7/29.9	69.0/63.5	93.8/90.4	64.4/55.9
	annotator3	full instruction, temp=0.2	79.3/75.5	13.2/14.0	49.3/66.2	44.2/31.3	67.7/62.1	87.0/77.9	61.9/53.6
	annotator4	full instruction, temp=0.7	80.5/76.5	13.2/12.5	49.9/66.6	43.7/29.9	68.0/62.7	94.1/90.7	64.3/55.8
	annotator5	1-shot, temp=0.2	86.4/81.4	13.0/**33.0**	48.7/64.6	42.7/29.2	67.5/63.7	90.9/84.0	66.7/60.5
	annotator6	1-shot, temp=0.7	**87.1/82.1**	13.0/22.1	57.8/77.6	42.2/31.0	64.2/60.4	**95.9/93.0**	67.5/61.3
	annotator7	5-shot, temp=0.2	81.3/74.9	12.4/20.2	55.4/73.3	41.4/31.3	70.6/67.2	95.2/91.9	66.1/59.4
	annotator8	5-shot, temp=0.7	86.5/82.3	12.6/27.2	**63.8/81.8**	**44.6/35.9**	68.4/64.9	95.0/91.5	**69.0/63.9**
	annotator9	COT 1-shot, temp=0.2	84.7/81.0	12.4/12.7	46.7/63.7	43.2/28.5	67.2/63.3	93.2/87.6	64.5/55.7
	annotator10	COT 1-shot, temp=0.7	84.4/80.0	12.2/16.9	49.0/68.6	40.8/29.5	65.3/61.0	93.5/88.9	64.9/57.3
	annotator11	COT 5-shot, temp=0.2	81.6/77.7	13.4/11.4	50.4/66.6	40.7/32.4	**71.4/67.6**	95.5/92.4	65.4/56.9
	annotator12	COT 5-shot, temp=0.7	83.5/79.3	12.6/16.0	54.2/73.5	41.6/34.2	70.0/66.6	93.2/89.6	65.8/59.0
MPT Instruct	annotator1	simple prompt, temp=0.2	16.7/16.2	6.2/14.9	18.4/31.0	25.0/27.4	40.2/37.5	17.1/15.4	19.9/21.8
	annotator2	simple prompt, temp=0.7	25.3/23.1	5.8/20.8	16.5/35.7	13.1/29.2	31.3/28.5	23.5/20.3	20.9/25.2
	annotator3	full instruction, temp=0.2	39.2/34.8	5.4/15.2	**24.2**/24.7	**34.0**/30.2	31.8/30.1	41.8/35.1	30.3/27.3
	annotator4	full instruction, temp=0.7	31.9/28.6	5.3/**24.9**	19.6/29.2	22.8/**31.0**	31.0/29.1	32.4/27.6	25.4/27.8
	annotator5	1-shot, temp=0.2	25.7/24.1	5.9/7.2	21.4/29.8	29.4/22.8	16.2/15.3	17.5/16.5	18.3/18.0
	annotator6	1-shot, temp=0.7	27.6/25.2	5.5/17.0	13.0/28.0	23.5/23.5	22.1/21.0	36.1/31.5	22.9/24.0
	annotator7	5-shot, temp=0.2	49.5/45.9	4.4/9.4	23.1/**43.5**	20.5/15.7	54.3/50.1	**69.5**/56.9	41.6/**36.5**
	annotator8	5-shot, temp=0.7	37.2/33.7	5.0/24.4	13.2/35.9	16.5/18.5	37.6/33.8	46.1/36.0	29.8/30.9
	annotator9	COT 1-shot, temp=0.2	22.7/21.4	6.5/6.2	17.5/23.5	29.8/21.7	16.1/15.1	30.1/27.6	20.0/18.2
	annotator10	COT 1-shot, temp=0.7	25.5/23.0	**8.5**/18.8	14.0/28.0	19.3/19.9	18.3/16.9	37.7/33.0	22.3/23.5
	annotator11	COT 5-shot, temp=0.2	**58.7/55.1**	6.5/7.4	23.6/22.4	21.6/14.6	**64.1/59.0**	67.8/**58.8**	**45.2**/36.0
	annotator12	COT 5-shot, temp=0.7	41.9/37.7	5.6/18.8	22.3/27.3	11.8/13.9	49.1/44.1	47.7/39.7	33.7/30.7

Table 7: First Run LLM Annotators: Micro-Averaged F1 Score/Accuracy

RUN 2: Micro F1 Score/ Accuracy(%)									
LLM	**Annotator**	**Annotator Description**	**Entity-Pair**						**Total**
			ORG-GPE	**ORG-ORG**	**ORG-DATE**	**ORG-MONEY**	**PER-ORG**	**PER-TITLE**	
GPT-4	annotator1	simple prompt, temp=0.2	80.5/75.2	15.3/38.0	50.6/68.4	48.6/**44.8**	71.1/68.0	92.9/87.2	67.6/63.6
	annotator2	simple prompt, temp=0.7	80.3/74.9	15.0/38.3	49.9/67.9	**48.7**/44.1	71.0/68.2	92.7/86.9	67.4/63.4
	annotator3	full instruction, temp=0.2	**81.3/75.9**	15.4/36.6	55.9/74.4	47.4/42.0	72.3/69.5	93.5/88.4	68.4/64.5
	annotator4	full instruction, temp=0.7	80.9/75.8	15.4/37.5	57.7/75.6	47.6/42.7	**72.4/69.7**	93.1/87.8	68.5/64.8
	annotator5	1-shot, temp=0.2	79.9/74.5	14.3/29.4	48.8/66.4	47.4/42.0	62.8/60.2	94.2/89.6	65.0/60.1
	annotator6	1-shot, temp=0.7	79.1/73.4	14.3/29.1	48.2/65.7	48.0/42.3	62.5/60.0	**94.7/90.4**	64.8/59.8
	annotator7	5-shot, temp=0.2	78.5/73.5	15.9/34.9	58.2/76.9	47.6/40.2	71.1/67.8	93.4/88.1	67.4/63.5
	annotator8	5-shot, temp=0.7	79.7/74.4	15.0/35.2	58.9/77.6	48.0/41.3	71.7/68.2	93.2/87.8	67.8/64.0
	annotator9	COT 1-shot, temp=0.2	80.4/74.8	16.4/32.9	37.3/48.0	48.3/43.1	64.7/61.9	94.5/89.9	64.7/58.6
	annotator10	COT 1-shot, temp=0.7	80.2/74.8	16.3/32.2	36.4/46.9	47.2/41.6	65.5/62.9	94.5/89.9	64.6/58.3
	annotator11	COT 5-shot, temp=0.2	79.9/74.5	**16.6**/37.7	**65.2/83.0**	47.0/43.1	71.3/68.2	92.9/87.0	**68.5/65.5**
	annotator12	COT 5-shot, temp=0.7	80.2/74.5	16.5/**37.9**	64.9/82.9	46.8/42.7	70.7/67.4	92.9/87.0	68.4/65.3
PaLM 2	annotator1	simple prompt, temp=0.2	80.1/75.9	14.0/14.1	49.7/67.0	43.5/29.2	66.9/60.8	87.0/77.7	62.0/53.5
	annotator2	simple prompt, temp=0.7	80.6/76.6	13.8/13.8	48.4/65.3	**43.9**/30.6	67.8/61.9	94.1/91.1	64.5/56.0
	annotator3	full instruction, temp=0.2	80.2/76.3	13.6/13.8	49.6/66.2	**43.9**/30.6	69.1/63.7	87.1/78.0	62.4/53.9
	annotator4	full instruction, temp=0.7	79.9/75.8	**14.1**/14.7	49.5/66.1	43.6/29.5	68.5/63.1	94.1/91.0	64.5/56.2
	annotator5	1-shot, temp=0.2	86.1/81.0	13.1/31.7	47.3/63.0	42.5/28.5	67.2/63.3	90.3/83.1	66.1/59.6
	annotator6	1-shot, temp=0.7	**87.4/82.0**	13.1/21.8	55.3/74.7	42.2/29.9	63.5/60.0	**95.7/92.2**	67.1/60.4
	annotator7	5-shot, temp=0.2	82.1/75.8	11.8/18.3	56.1/74.0	41.6/32.7	69.6/66.4	94.4/90.5	65.8/59.0
	annotator8	5-shot, temp=0.7	86.0/81.8	13.3/**27.3**	**64.3/82.1**	43.2/**35.9**	68.5/65.4	93.6/89.3	**68.4/63.6**
	annotator9	COT 1-shot, temp=0.2	84.7/81.0	12.5/12.8	47.4/65.2	43.3/28.8	68.7/64.7	92.7/87.0	64.8/56.1
	annotator10	COT 1-shot, temp=0.7	84.6/80.1	11.7/16.3	49.1/68.6	41.1/29.2	65.5/61.9	94.0/89.8	64.9/57.5
	annotator11	COT 5-shot, temp=0.2	82.4/78.6	13.3/11.8	50.7/67.0	43.2/35.2	**71.5/67.8**	95.1/92.1	65.8/57.5
	annotator12	COT 5-shot, temp=0.7	82.4/78.2	**14.1**/17.6	54.5/74.2	41.5/32.4	69.0/65.6	94.6/91.6	66.1/59.4
MPT Instruct	annotator1	simple prompt, temp=0.2	17.7/17.0	6.5/15.8	17.7/31.2	25.2/26.0	40.4/37.5	16.7/14.7	20.1/21.9
	annotator2	simple prompt, temp=0.7	21.0/20.3	6.4/20.0	15.1/34.7	15.1/22.4	33.2/30.5	25.5/21.7	20.6/24.2
	annotator3	full instruction, temp=0.2	42.2/38.0	7.2/13.9	24.0/24.2	31.2/**28.5**	32.9/30.7	44.7/37.4	31.9/27.9
	annotator4	full instruction, temp=0.7	31.4/28.3	5.2/**23.3**	19.0/31.8	18.2/27.0	26.9/24.9	33.0/26.9	24.2/26.8
	annotator5	1-shot, temp=0.2	28.2/25.9	4.7/5.7	21.5/29.6	**33.4**/23.1	16.3/15.3	17.3/16.5	18.9/18.0
	annotator6	1-shot, temp=0.7	28.7/26.1	5.3/17.5	16.3/29.6	25.6/27.8	19.1/19.4	33.2/29.3	22.4/24.3
	annotator7	5-shot, temp=0.2	54.9/51.1	4.5/7.3	22.6/38.3	18.8/14.6	55.4/52.2	**71.1/59.8**	43.3/36.9
	annotator8	5-shot, temp=0.7	41.1/36.3	4.4/21.7	13.0/**35.9**	16.2/17.1	34.3/31.1	50.2/41.7	31.1/31.3
	annotator9	COT 1-shot, temp=0.2	24.1/22.4	**7.8**/8.5	17.1/24.0	31.2/22.1	15.0/14.2	28.4/26.3	20.1/18.7
	annotator10	COT 1-shot, temp=0.7	28.3/26.6	7.0/16.1	11.5/25.3	19.8/21.0	17.0/16.9	37.1/31.8	22.1/22.9
	annotator11	COT 5-shot, temp=0.2	**58.8/55.5**	6.0/8.0	**25.4**/23.5	20.1/13.9	**61.2/56.3**	68.6/59.7	**45.2/36.1**
	annotator12	COT 5-shot, temp=0.7	45.5/40.8	4.3/15.3	15.9/26.9	21.0/22.1	46.2/42.1	48.2/40.6	34.0/30.9

Table 8: LLM Annotators: Micro-average F1-Score / Accuracy for second run

A.2.6. Statistical Tests

Are Difference Statistically significant at alpha = 0.05?			
Null Hypothesis	**LLM**	**Micro-Averaged F1 Scores P-values**	**Accuracy P-values**
Ho:There is no significant difference in metric when we change temperature setting i.e. Ho: metrics at temp0.2 =metrics at temp0.7	GPT-4	0.950	0.975
	PaLM 2	0.053	0.062
	MPT Instruct	0.481	0.734
Ho:At temperature setting = 0.2, there is no significant difference in metric when we compare run1 and run2 i.e. Ho: metrics at temp0.2_first_run =metrics at temp0.2_second_run	GPT-4	0.935	0.959
	PaLM 2	0.964	0.932
	MPT Instruct	0.921	0.954
Ho:At temperature setting = 0.7, there is no significant difference in metric when we compare run1 and run2 i.e. Ho: metrics at temp0.7_first_run =metrics at temp0.7_second_run	GPT-4	0.973	0.976
	PaLM 2	0.948	0.992
	MPT Instruct	0.974	0.889
Ho:There is no significant difference when we compare average metrics across first run and second for the different temperature setting i.e Ho: avg_metric at temp0.2 =avg_metric at temp0.7	GPT-4	0.967	0.983
	PaLM 2	0.195	0.213
	MPT Instruct	0.493	0.909
Ho:There is no significant difference between LLM metrics when we compare one to the other. i.e Ho: LLM1 avg_metric at temp 0.2 = LLM2 avg_metric at temp 0.2	GPT-4-vs-PaLM 2	0.055	**0.004***
	GPT-4-vs-MPT Instruct	**0.000***	**0.000***
	PaLM 2-vs-MPT Instruct	**0.000***	**0.000***

Table 9: Statistical Significance of Metric Difference: At alpha = 0.05, we test if the difference captured are statistical significant. For these hypotheses, we either reject or fail to reject the null hypothesis.

A.2.7. Inter Annotator Agreement

What we measure	Prompt Type	Inter annotator agreement (IAA)		
		GPT-4	PaLM 2	MPT Instruct
Same Prompt at same temperature setting (run twice).Using Cohen Kappa	simple_prompt, temp = 0.2	0.96	0.88	0.62
	simple_prompt, temp = 0.7	0.94	0.89	0.19
	full_instructn_prompt, temp = 0.2	0.97	0.87	0.67
	full_instructn_prompt, temp = 0.7	0.94	0.89	0.23
	1shot_prompt, temp = 0.2	0.96	0.91	0.55
	1shot_prompt, temp = 0.7	0.95	0.86	0.2
	5shot_prompt, temp = 0.2	0.96	0.92	0.51
	5shot_prompt, temp = 0.7	0.95	0.86	0.19
	cot 1shot_prompt, temp = 0.2	0.95	0.94	0.53
	cot 1shot_prompt, temp = 0.7	0.93	0.84	0.22
	cot 5shot_prompt, temp = 0.2	0.97	0.93	0.6
	cot 5shot_prompt, temp = 0.7	0.96	0.82	0.23
Same Prompt at different temperature setting (0.2 and 0.7) for only run 1. Using Cohen Kappa	simple_prompt	0.94	0.87	0.3
	full_instructn_prompt	0.95	0.86	0.34
	1shot_prompt	0.95	0.82	0.27
	5shot_prompt	0.96	0.84	0.26
	cot1shot_prompt	0.95	0.85	0.28
	cot5shot_prompt	0.96	0.83	0.34
	simple_prompt	0.95	0.87	0.34
	full_instructn_prompt	0.95	0.87	0.37
	1shot_prompt	0.95	0.84	0.31
	5shot_prompt	0.96	0.86	0.3
	cot1shot_prompt	0.94	0.86	0.31
	cot5shot_prompt	0.96	0.85	0.36
Compare prompts within prompt types. Using Cohen Kappa	zero shot: simple_vs_full_instruction	0.87	0.88	0.39
	few shot: 1_shot_vs_5_shot	0.84	0.79	0.28
	cot few shot: cot_1_shot_vs_cot_5_shot	0.8	0.82	0.28
Compare among prompt types. Using Fleiss Kappa	simple_vs_full_instruction_1_shot_ vs_5shot_vs_cot_1_shot_vs_cot_5_shot	0.83	0.79	0.31

Table 10: LLM Inter Annotator Agreement: This table shows how consistent outputs from each LLMs are within and accross prompt types and within and accross different temperature settings.

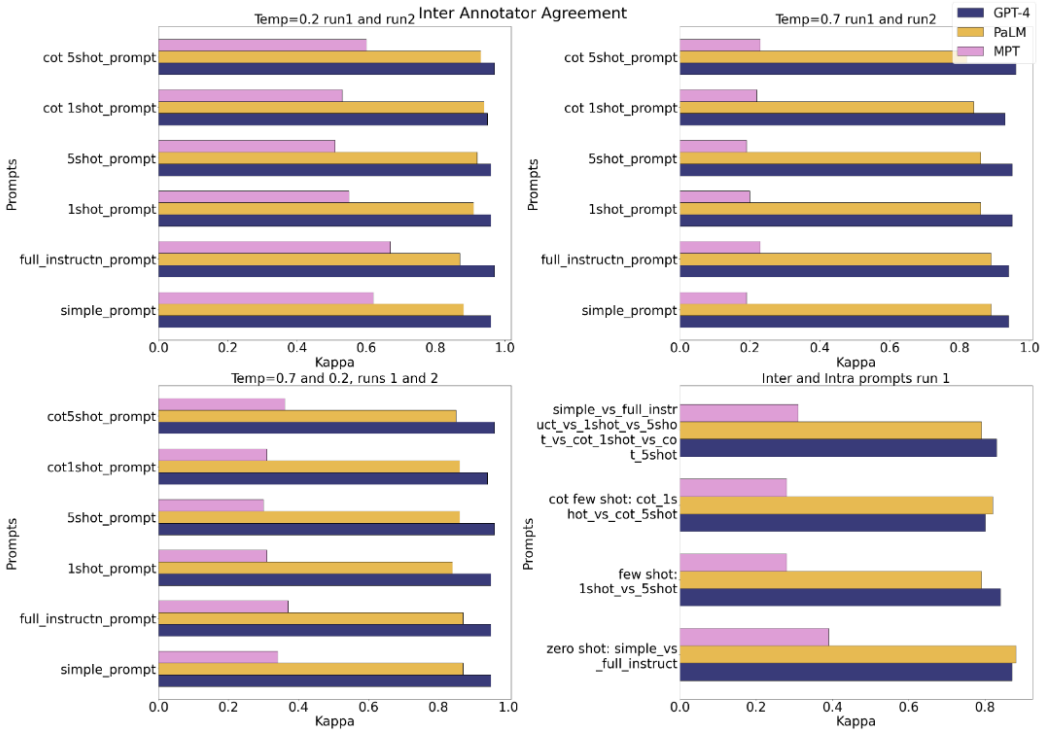

Figure 9: Plots from Inter Annotator Agreement scores

A.2.8. Error Analysis

Entity Pair	Scenario 1: Crowd workers = Wrong Answer, LLMs = Correct Answer	Scenario 2: Crowd workers = Wrong Answer, LLMs = Wrong Answer	Scenario 3: Crowd workers = Correct Answer, LLMs = Wrong Answer
ORG-GPE	Properties Atlas Financial Holdings , Inc. corporate headquarters is located at 150 Northwest Point Boulevard , Elk Grove Village , Illinois 60007 , USA. Expert Label:Headquartered in Crowd worker Label: Operations in LLMs Label:Headquartered in	Our eWellness Corporate Office is located in Culver City , California . eWELLNESS Expert Label: Operations in Crowd worker Label: Formed in LLMs Label: Headquartered in	This Settlement Agreement (" Agreement ") is made effective this 20th day of May , 2015 by and between ActiveCare , Inc , a Delaware corporation (the " Company ") , and Advance Technology Investors , LLC (" ATI ") . Expert Label:Operations in Crowd worker Label: Operations in LLMs Label: No/Other Relation, Formed in
ORG-ORG	Michael D. Huddy , President / CEO and Director , joined INTERNATIONAL BARRIER TECHNOLOGY INC in February 1993 as President of the newly - formed US Subsidiary, Barrier Technology Corporation . Expert Label:Subsidiary of Crowd worker Label: No/Other Relation, Shares of LLMs Label: Subsidiary of	Our Hawaii Gas entered into licensing agreements with Utility Service Partners , Inc. and America's Water Heater Rentals , LLC , both indirect subsidiaries of Macquarie Group Limited, to enable these entities to offer products and services to Hawaii Gas's customer base Expert Label:Subsidiary of Crowd worker Label: No/Other Relation, Subsidiary of, Shares of LLMs Label: Agreement with	On December 10 , 2014 , Orbital Tracking Corp. purchased certain contracts from Global Telesat Corp ,a Virginia corporation (GTC) for $ 250,000 pursuant to an asset purchase agreement by and among Orbital Tracking Corp i, its wholly owned subsidiary Orbital Satcom, GTC and World Surveillance Group , Inc. (World) , GTC's parent Expert Label: Subsidiary of Crowd worker Label: Subsidiary of LLMs Label: Agreement with
ORG-DATE	Wishbone Pet Products Inc. was incorporated in the State of Nevada on July 30 , 2009 . Expert Label: Formed on Crowd worker Label: No/Other Relation LLMs Label: Formed on	None	None
ORG-MONEY	Personal Lines underwriting profit for the three months ended September 30 , 2017 was $40.8 million , compared to $23.3 million for the three months ended September 30 ,2016 , an improvement of $17.5 million . Expert Label: Profit of Crowd worker Label: No/Other Relation, Loss of LLMs Label: Profit of	None	None
PERS-ORG	Mr. Untermeyer also serves as senior program manager with Southwest Research institute , San Antonio Expert Label: Employee of Crowd worker Label: Founder of, Member of LLMs Label: Employee of	Currently , Mr. Morrison serves on the board of directors of the Texas AM university , kingsville foundation and the Rockport center for the arts. Expert Label: Employee of Crowd worker Label: Founder of, Member of LLMs Label: Member of	From September 2012 through June 2015 , Mr. Kimmel has also served on the board of directors of Electronic Magnetic Power Solutions , which implements disruptive patented technology licensed from Virginia Tech University for the express purpose of alternative energy use in the consumer space . Expert Label: Employee of Crowd worker Label: Employee of LLMs Label: Member of, No/Other Relation
PERS-TITLE	Information regarding Harel Gadot , Microbot Medical Inc. Chairman , President and Chief Executive Officer , is set forth above under Board of Directors . Expert Label:Title Crowd worker Label: No/Other Relation LLMs Label:Title	None	Yvonne should contact her manager , segment or region leader , or FTI Consulting s Chief Ethics and Compliance Officer to discuss the gift . Expert Label: Title Crowd worker Label: Title LLMs Label: No/Other Relation

Table 11: Qualitative Examples from our Error Analysis depicting the 3 prominent scenarios of how MTurk Crowd workers and LLMs demonstrated high confidence on answer choice

A.2.9. Confusion Matrix for GPT-4

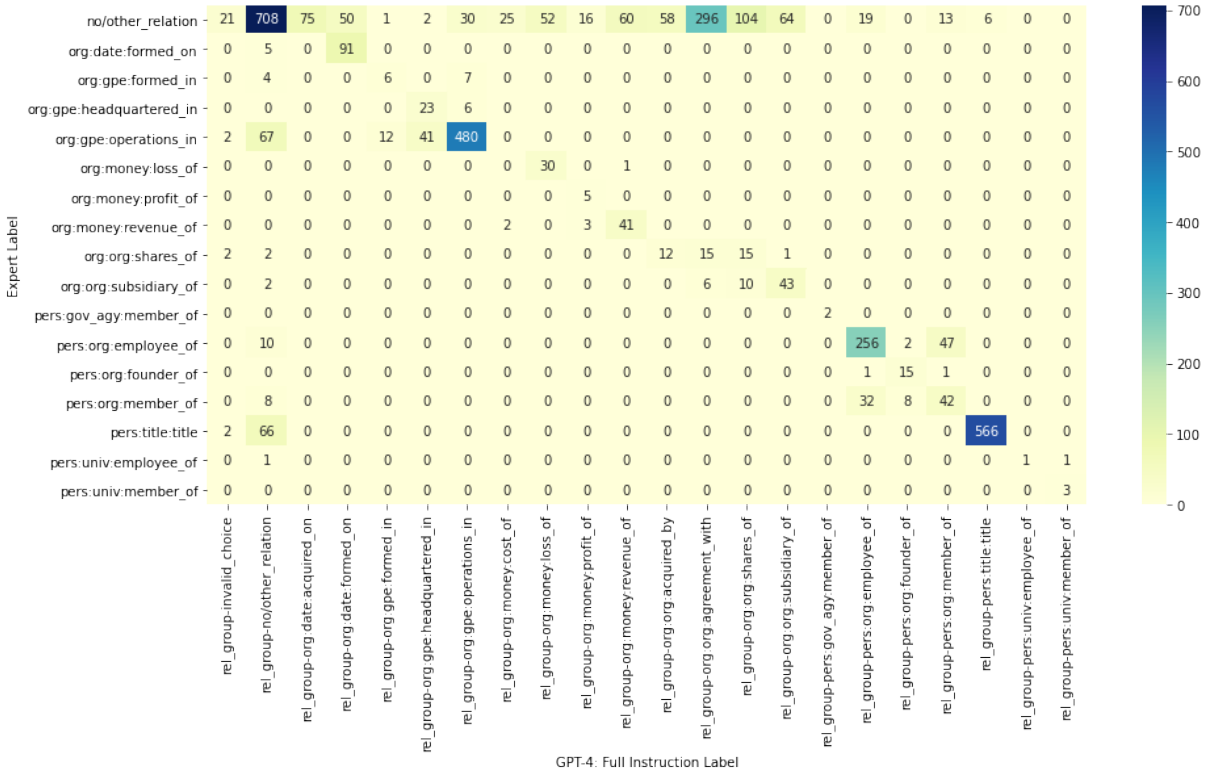

Figure 10: Confusion Matrix for GPT-4 Zero Shot Prompt

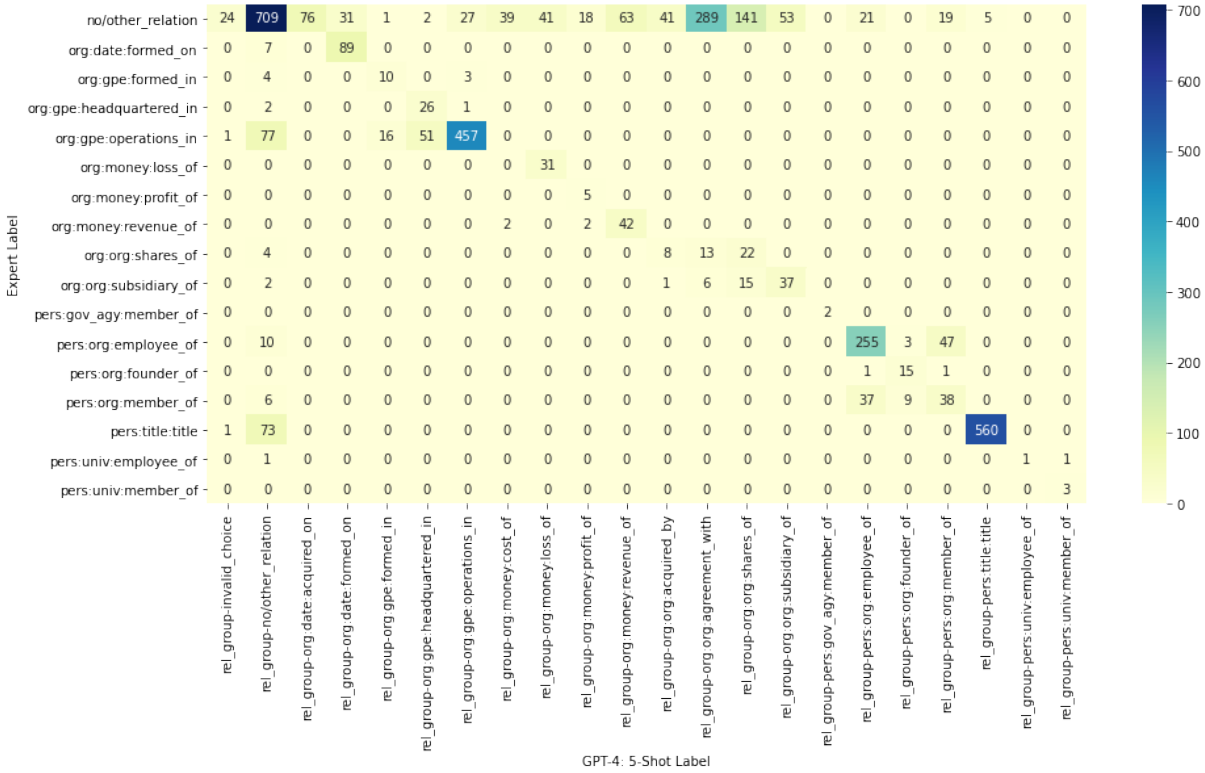

Figure 11: Confusion Matrix for GPT-4 Few Shot Prompt

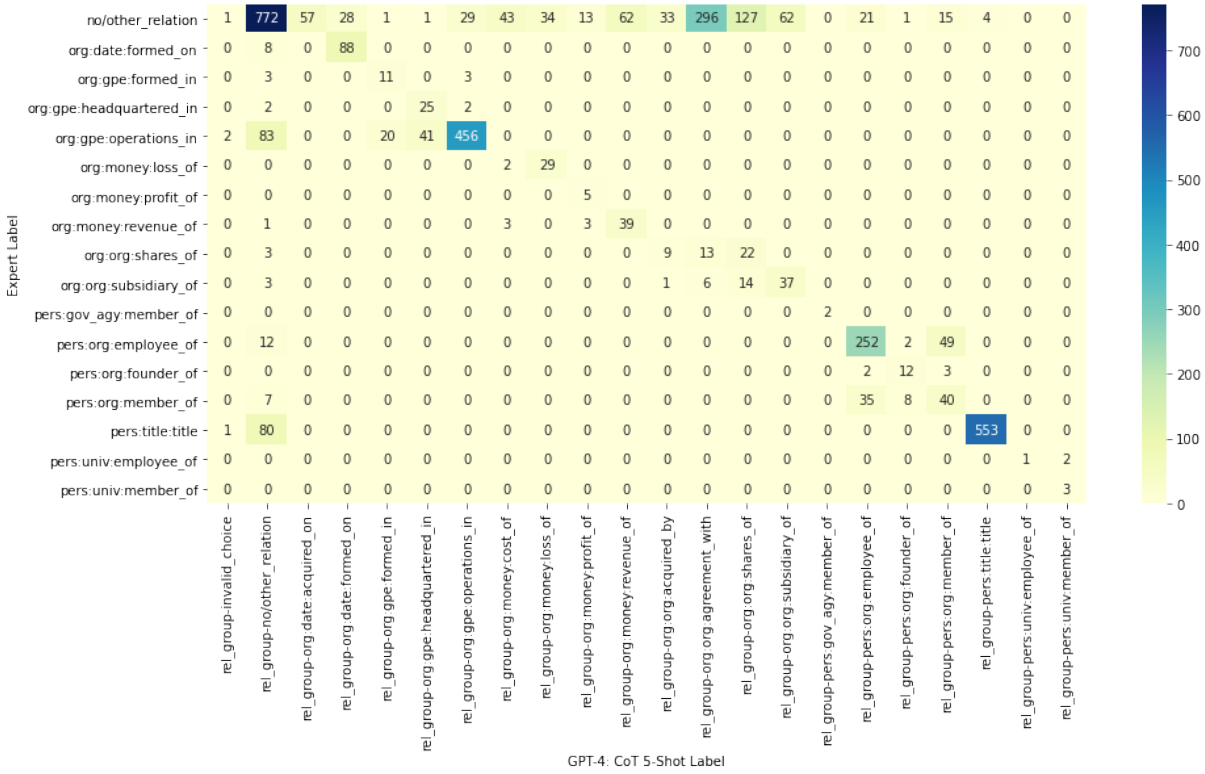

Figure 12: Confusion Matrix for GPT-4 Few Shot CoT Prompt