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

toyin.d.aguda@jpmchase.com **JPMorgan AI Research**

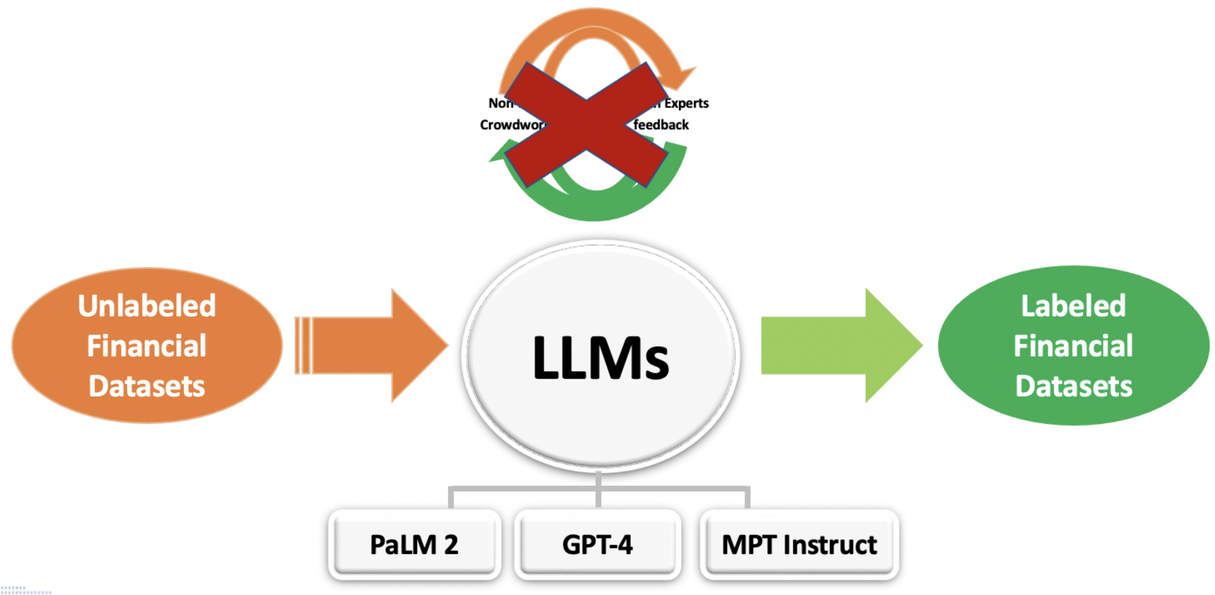
1. Overview

Our study explores the effectiveness of Large Language Models (LLMs) as data annotators in specialized domains, specifically in the financial relation extraction task:

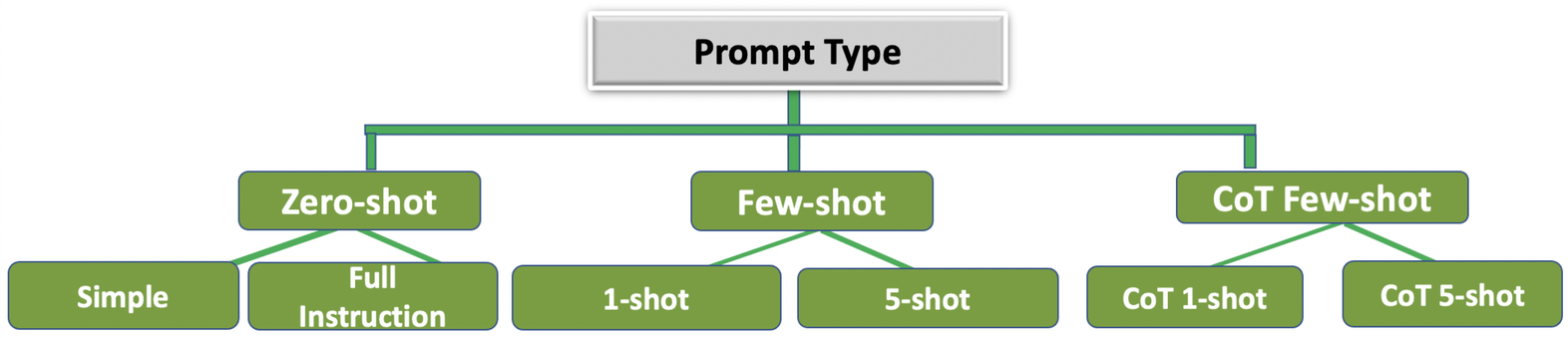
1. Examine if LLMs can replace non-domain expert human annotators.
2. Evaluate 3 different LLMs.
3. Introduce a reliability index (LLM-RelIndex) to measure the trustworthiness of LLM annotations.

2. Problem Statement

Are the **state-of-the-art** LLM(s) adequate alternatives for non-domain expert human annotators?



3. Prompt Type



4. Data & Model Selection

Data: **REFinD**(Kaur et al., 2023).

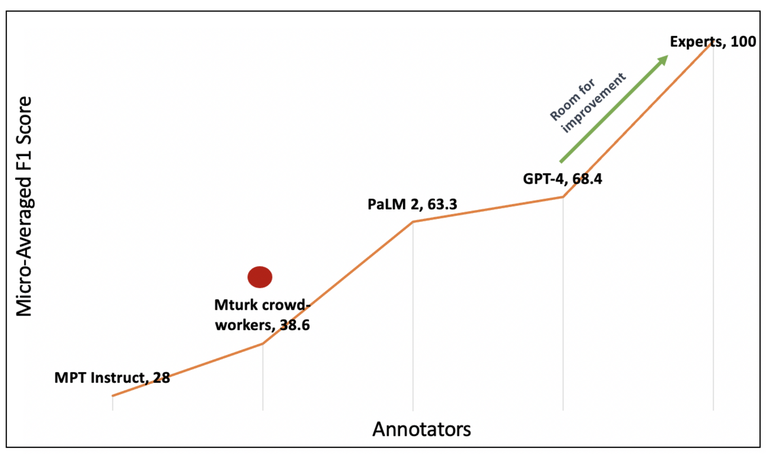
We use the following Large Language Models (LLM):

1. GPT-4
2. PaLM 2
3. MPT Instruct

5. Model Performance

Micro-Averaged F1 Score/ Accuracy(%)								
			Zero-Shot Prompt		Few-Shot Prompt		Few-Shot CoT Prompt	
Annotator	Type	Temperature Setting	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	-	-	-	-

6. Annotator Performance

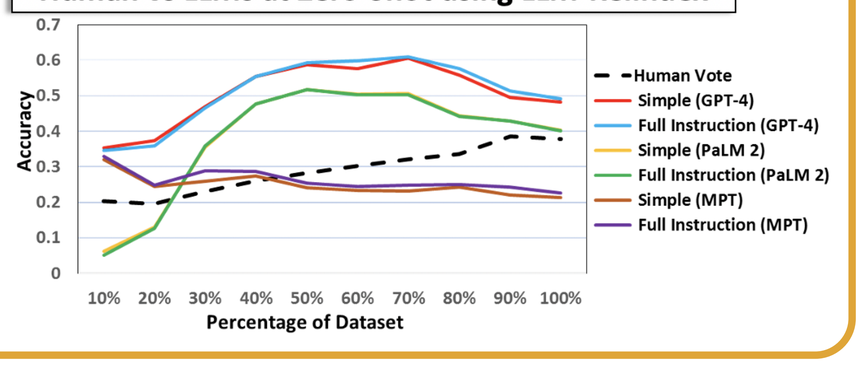


7. Model Consistency & Reliability

Inter- Annotator Agreement

	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

Human vs LLMs at Zero-Shot using LLM-RelIndex



8. Error Analysis

Common Issues encountered by LLMs

1. Semantic Ambiguity

Errors due to proximity and similarity of the answer choices ➔ "Subtle differences" such as "***Member of***" in place of "***Employee of***"

2. Confident Missannotation

Majority of: crowdworker labels = incorrect ; LLM labels = correct

Majority of: crowdworker labels = incorrect ; LLM labels = incorrect

Majority of: crowdworker labels = correct ; LLM labels = incorrect

3. Relation Hallucination

LLMs hallucinated mostly in cases where humans identified the relation as "No/Other relation". In this category, PaLM 2 has the most hallucination at 80.7%

9. Conclusion

Are the state-of-the-art LLMs adequate alternatives for non-domain expert human annotators?

1. **YES!** GPT-4 & PaLM 2 significantly outperform crowdsourced annotations with over **29%**.
2. In term of **Scalability:** LLM greatly reduced time and cost associated with annotation.
3. **Choice of LLMs:** Choose best available models.
4. Expert annotation only where necessary using **LLM-Rellindex**

10. Future Work

1. Explore the uses of the automatically annotated dataset for downstream tasks.
2. Develop smaller efficient models
3. Make use of **sensible hallucinations** for instances labelled as NO/OTHER RELATION.