



# LARGE LANGUAGE MODELS AS FINANCIAL DATA ANNOTATORS: A STUDY ON EFFECTIVENESS AND EFFICIENCY

**Thành viên:**

Nguyễn Duy Hoàng - 23020368

Nguyễn Trung Hiếu - 23020366

Dương Lý Khánh Hạ - 23020362

# Nội dung

I/ Giới thiệu

II/ Phương pháp nghiên cứu

III/ Tiêu chí đánh giá

IV/ Tổng kết



# I/ Giới thiệu

## 1/ Lý do nghiên cứu

- Gán nhãn dữ liệu tài chính  
tốn kém và cần chuyên gia
- Crowdsourcing có độ chính  
xác thấp
- Cân phương pháp đánh giá  
tự động có hiệu suất cao  
và chi phí thấp



# I/ Giới thiệu

## 2/ Mục tiêu nghiên cứu

- Đánh giá khả năng của các mô hình ngôn ngữ lớn (LLMs) trong việc gán nhãn dữ liệu tài chính
- So sánh LLMs với chuyên gia và nhân viên crowdsourcing
- Phân tích về chi phí độ chính xác, thời gian



# Mục tiêu

1

Hiệu quả về mặt chi phí và thời gian khi sử dụng LLMs

2

Cách tối ưu để sử dụng LLMs trong gán nhãn dữ liệu tài chính ?

3

LLMs có thể thay thế con người trong gán nhãn dữ liệu tài chính không ?



# I/ Giới thiệu

## 3/ Cách tiếp cận

- Tập dữ liệu : Dữ liệu REFinD
- Mô hình sử dụng: GPT-4,PaLM 2,MPT Instruct
- Phương pháp đánh giá chung: chạy mô hình với nhiều thay đổi về data, temprature, prompt, .... và đánh giá bằng các chỉ số cân bằng, ....



## II/ Phương pháp nghiên cứu

### Các phương pháp so sánh

Chạy các mô hình LLM với nhiều prompt khác nhau (zero-shot,few-shot,chain-of-thought)

So sánh độ chính xác của LLM với các chuyên gia và nhân viên crowdsourcing

Đánh giá về chi phí, thời gian thực hiện và tính nhất quán



# Tổng quan

Mô hình ngôn ngữ lớn (Large Language Models - LLMs) là các mô hình học sâu (deep learning) được huấn luyện trên tập dữ liệu văn bản khổng lồ nhằm hiểu, tạo, và thao tác với ngôn ngữ tự nhiên. Chúng được xây dựng dựa trên kiến trúc Transformer, nổi bật với cơ chế self-attention giúp mô hình hiểu được ngữ cảnh của từ trong câu và sinh văn bản có ý nghĩa.



# Nguyên lý hoạt động

1. **Tiền huấn luyện:** Mô hình được huấn luyện trên tập dữ liệu lớn để hiểu ngôn ngữ tự nhiên
2. **Tinh chỉnh:** Mô hình được tinh chỉnh trên tập dữ liệu nhỏ để thực hiện các nhiệm vụ cụ thể
3. **Dự đoán:** Mô hình được sử dụng để dự đoán, phân loại, sinh văn bản, ...



## Gán nhãn dữ liệu trong kinh tế

Gán nhãn dữ liệu (Data Labeling) là quá trình gán các nhãn hoặc thông tin mô tả cho dữ liệu để giúp mô hình học máy (Machine Learning - ML) học cách nhận diện các mẫu (patterns). Dữ liệu đã được gán nhãn được gọi là dữ liệu có giám sát (labeled data) và thường được sử dụng trong các bài toán học có giám sát (Supervised Learning).



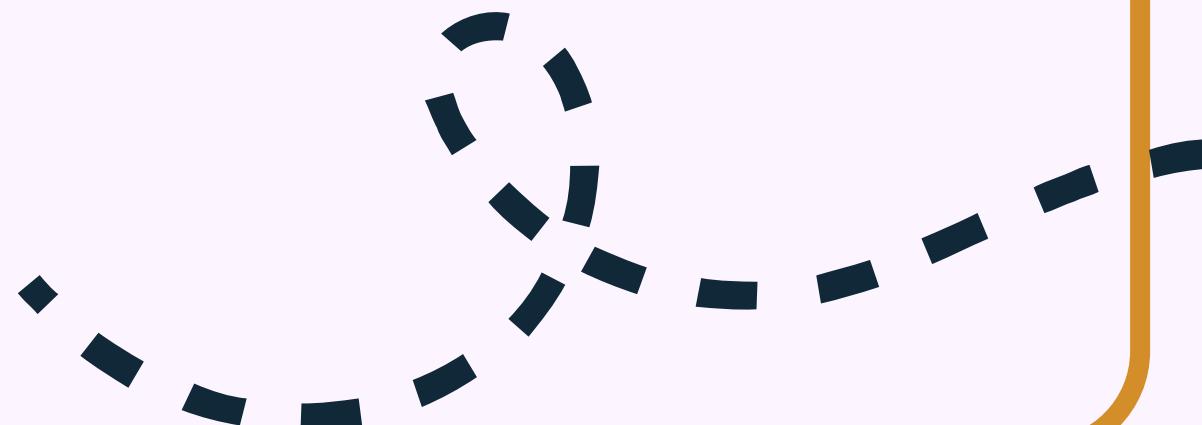
# Các phương pháp truyền thống

## 1/ Gán nhãn thủ công (mantual labeling)

	Gán nhãn thủ công	Gán nhãn bằng crowdsourcing
Đặc điểm	Các chuyên gia hoặc nhân viên nhập liệu trực tiếp xem xét dữ liệu và gán nhãn tương ứng.	Nhiều người tham gia (crowd workers) cùng thực hiện gán nhãn thông qua các nền tảng như Amazon Mechanical Turk, Appen, hoặc Labelbox.
Ưu điểm	Chính xác, đáng tin cậy	Chi phí thấp, nhanh chóng
Nhược điểm	Tốn kém, tốn thời gian, không thể áp dụng cho dữ liệu lớn	Chất lượng không đảm bảo, cần kiểm soát chất lượng

# Vai trò

- 1.Cải thiện độ chính xác trong phân tích kinh tế
- 2.Hỗ trợ mô hình hóa rủi ro và dự báo thị trường
- 3.Cải thiện hiệu suất dự báo kinh tế bằng AI
- 4.Hỗ trợ hoạch định chính sách kinh tế
- 5.Cá nhân hóa dịch vụ tài chính



# Phương pháp gán nhãn sử dụng LLMs

Trong nhiều tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, trích xuất thực thể, hay trích xuất quan hệ, dữ liệu gán nhãn đóng vai trò thiết yếu để huấn luyện các mô hình học máy. Tuy nhiên, quá trình gán nhãn truyền thống thường phụ thuộc vào chuyên gia, tốn kém và khó mở rộng. Sự phát triển của các mô hình ngôn ngữ lớn (LLMs) như GPT, LLaMA, Claude,... đã mở ra khả năng tận dụng kiến thức và khả năng hiểu ngữ cảnh mạnh mẽ của chúng để tự động gán nhãn dữ liệu, giảm phụ thuộc vào gán nhãn thủ công.

Có ba cách tiếp cận phổ biến:

1. Zero-shot labeling: Không cần ví dụ minh họa, chỉ cần prompt rõ ràng.
2. Few-shot labeling: Cung cấp vài ví dụ minh họa để mô hình học cách gán nhãn.
3. Self-training / weak supervision: LLMs gán nhãn cho tập dữ liệu chưa gán, sau đó nhãn đó được dùng để huấn luyện mô hình nhỏ hơn.

# Dateset-REFind

- Dữ liệu REFinD trích xuất từ các bài báo tài chính hàng quý hàng năm của công ty niêm yết (10-K,10-Q của SEC)
- 28676 mẫu, 22 quan hệ tài chính
- Ví dụ về nhiệm vụ trích xuất quan hệ

**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.

# Các loại thực thể và quan hệ trong REFinD

- 6 cặp thực thể chính: ORD-DATE, ORD-MONEY, PER-ORG, PER-TITLE, ...

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

# Mô hình LLMs được thử nghiệm

- GPT-4 (OpenAI) - Khoảng 1.7 nghìn tỷ tham số
- PaLM 2 (Google) - Khoảng 340 tỷ tham số
- MPT Instruct (MosaicML) - Khoảng 7 tỷ tham số

→ GPT-4 và PaLM 2 hoạt động tốt hơn so với crowdworkers, trong khi MPT Instruct có hiệu suất thấp hơn nhưng vẫn có tiềm năng cải thiện.

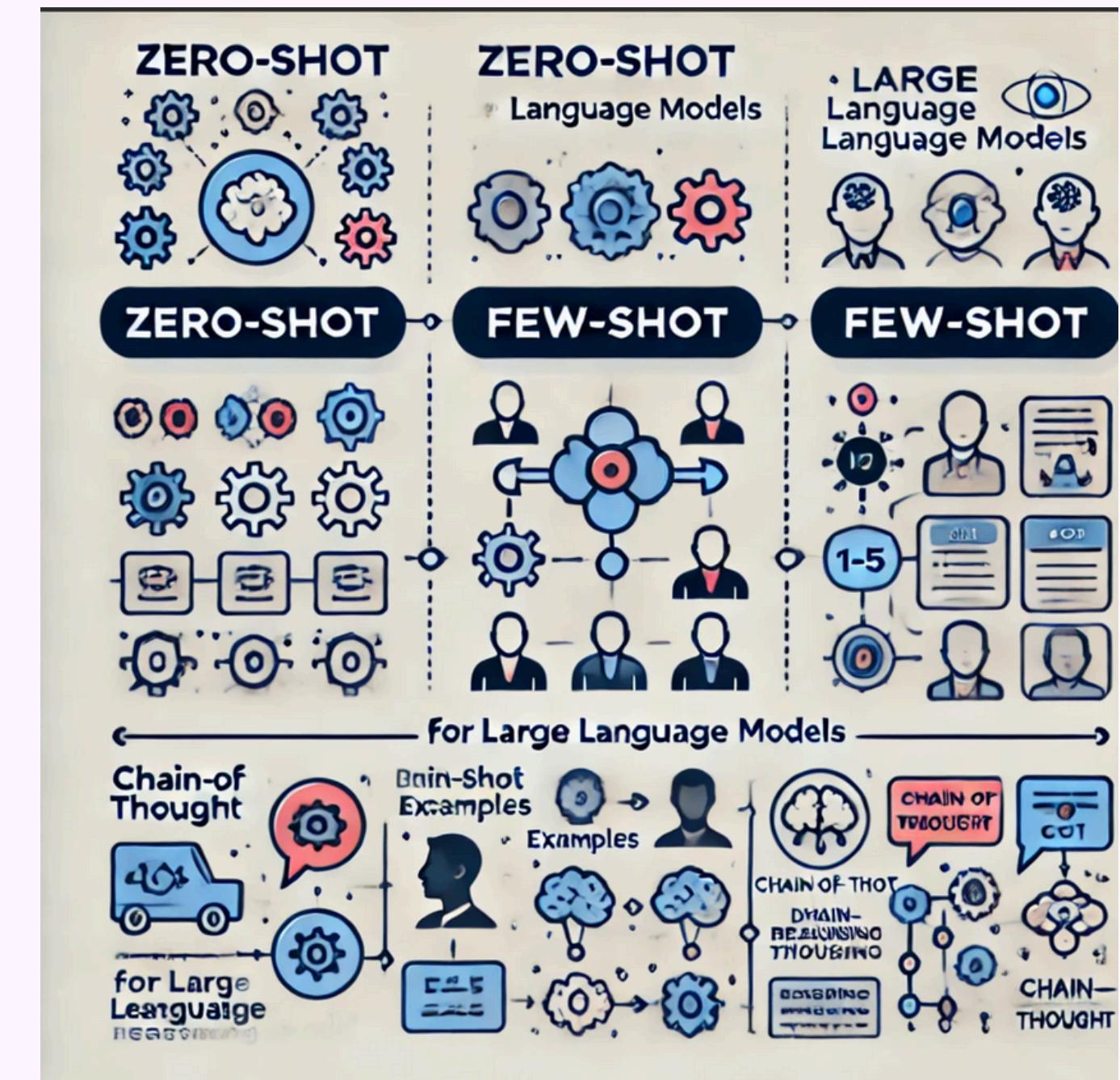


# Phương pháp thử nghiệm - Prompting

- Zero-shot
- Few-shot
- Chain-of-Thought(CoT)

## Ảnh hưởng

- Zero-shot
  - Kết quả không ổn định, đặc biệt khi gặp các trường hợp phức tạp hoặc cần kiến thức chuyên sâu.
  - Phụ thuộc hoàn toàn vào khả năng suy luận và kiến thức sẵn có của mô hình.
- Few-shot và Chain-of-Thought(CoT)
  - Cải thiện đáng kể độ chính xác của các mô hình, đặc biệt là với các nhiệm vụ phức tạp.
  - Cung cấp ví dụ và suy luận từng bước giúp mô hình hiểu rõ hơn và đưa ra dự đoán chính xác.



# III/ Tiêu chí đánh giá

Chính xác và F1-Score	Chi phí và thời gian	Reliability Index (LLM-RelIndex)
Đánh giá hiệu suất của mô hình bằng cách so sánh kết quả dự đoán với nhãn do chuyên gia gán	Phân tích chi phí tính toán và thời gian cần thiết để các mô hình hoàn thành nhiệm vụ	Chỉ số được đề xuất trong nghiên cứu để đo lường độ tin cậy của kết quả dự đoán
Accuracy đo lường tỷ lệ dự đoán đúng trên tổng số dự đoán	GPT-4 và PaLM 2 có chi phí cao hơn do quy mô lớn, trong khi MPT Instruct rẻ hơn nhưng hiệu suất thấp hơn	Giá trị cao cho thấy mô hình tự tin vào dự đoán, còn giá trị thấp cho thấy kết quả có thể cần xem xét lại bởi chuyên gia
F1-Score là trung bình điều hòa của độ chính xác (Precision) và độ phủ (Recall)		

### III/ Tiêu chí đánh giá

#### F1-Score

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

#### Chi phí gán nhãn

$$\text{Chi phí} = \frac{\text{Số token/ký tự đầu vào, đầu ra}}{\text{Giá mỗi token/ký tự}} \times \text{Số mẫu}$$

# III/ Tiêu chí đánh giá

## Rellndex

- Đối với mỗi nhãn dán  $l$ , chúng tôi tính điểm phiếu bầu có trọng số như sau  
$$\text{vote}(i, j) = \text{sim}(a_i, l)$$
- Sau đó tính điểm tin cậy của phiếu bầu bằng cách lấy trung bình điểm phiếu bầu  
$$\text{confid}(l) = \frac{1}{K} * \sum_1^K \text{vote}(i, l)$$
- Cuối cùng chỉ số Rellndex được xác định là độ tin cậy cao nhất trong số các nhãn dán được đề xuất

$$\text{Rellndex} = \max_l \text{confid}(l)$$

# III/ Tiêu chí đánh giá

## Cohen's Kappa

Đánh giá mức độ đồng thuận giữa hai người gán nhãn (annotators) khi họ phân loại một tập hợp dữ liệu, có tính đến xác suất đồng thuận ngẫu nhiên.

- ◆ Công thức:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- $p_o$ : tỷ lệ đồng thuận thực tế (observed agreement)
- $p_e$ : tỷ lệ đồng thuận kỳ vọng do ngẫu nhiên (expected agreement)

Giá trị Kappa ( $\kappa$ )	Mức độ đồng thuận
< 0	Tệ hơn ngẫu nhiên
0,01 - 0,20	Rất thấp
0,21 - 0,40	Thấp
0,41 - 0,60	Vừa phải
0,61 - 0,80	Cao
0,81 - 1,00	Rất cao

# III/ Tiêu chí đánh giá

## Fleiss' Kappa

Đánh giá mức độ đồng thuận giữa hai người gán nhãn (annotators) khi họ phân loại một tập hợp dữ liệu, có tính đến xác suất đồng thuận ngẫu nhiên.

Ý nghĩa:

- $\kappa=1$ : hoàn toàn đồng thuận
- $\kappa=0$ : đồng thuận do ngẫu nhiên
- $\kappa<0$ : ít đồng thuận hơn cả ngẫu nhiên (rất tệ)

### ► Các bước tính toán

1. Tính xác suất đồng thuận cho từng đối tượng  $P_i$ :

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^k n_{ij}(n_{ij} - 1)$$

2. Tính xác suất trung bình đồng thuận kỳ vọng  $\bar{P}_e$ :

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij} \quad (\text{tỉ lệ chọn nhãn } j \text{ trên toàn bộ tập dữ liệu})$$

$$\bar{P}_e = \sum_{j=1}^k p_j^2$$

3. Tính Fleiss' Kappa:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Trong đó:

- $\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i$ : trung bình xác suất đồng thuận thực tế trên tất cả các đối tượng

### Biểu đồ bên 1:

- GPT-4 luôn có Kappa cao nhất ( $> 0.8$  ở nhiều prompt) → rất nhất quán
- PaLM khá ổn định (Kappa  $\sim 0.6-0.8$ )
- MPT có độ đồng thuận thấp ( $< 0.5$  ở hầu hết prompt) → thiếu nhất quán

### Biểu đồ 2:

Với temperature cao hơn (0.7):

- Sự nhất quán giảm nhẹ, đặc biệt ở MPT (Kappa  $< 0.4$  ở nhiều prompt) → GPT-4 vẫn dẫn đầu về độ ổn định
- Các loại prompt như cot\_5shot\_prompt và full\_instruction giữ vững Kappa cao

### Biểu đồ 3 :

- Đo sự nhất quán giữa các điều kiện nhiệt độ khác nhau cho cùng prompt
- GPT-4 vẫn duy trì độ tin cậy cao (Kappa  $> 0.7$  nhiều prompt)
- PaLM và MPT bị ảnh hưởng rõ rệt khi thay đổi nhiệt độ

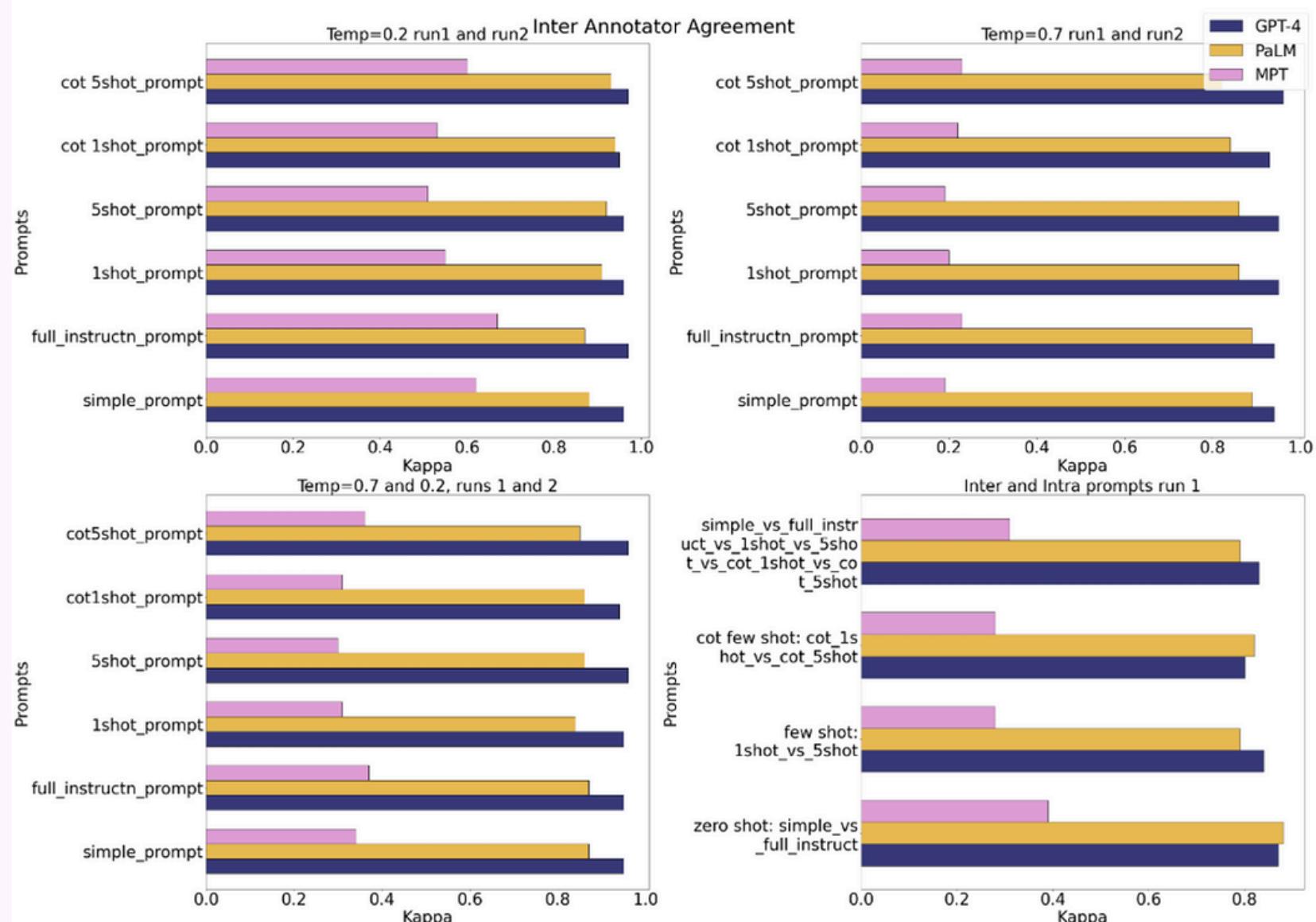


Figure 9: Plots from Inter Annotator Agreement scores

### Biểu đồ 4:

- GPT-4 có độ nhất quán cao giữa các prompt khác nhau (Kappa  $> 0.7$ )
- MPT bị ảnh hưởng nặng (Kappa  $\sim 0.3-0.5$ )
- Dễ thấy CoT prompts tạo ra sự ổn định tốt hơn giữa các kiểu prompt

# Tiêu chí đánh giá

-GPT-4 (Temperature = 0.2):

- Zero-shot Full instruction: F1 = 68.5 / Acc = 64.6 (rất cao)
- Few-shot 5-shot CoT: F1 = 68.6 / Acc = 46.5 → GPT-4 có hiệu suất ổn định và cao nhất trong nhiều chế độ.

-PaLM 2:

- Hiệu suất kém hơn GPT-4, nhưng đáng chú ý ở:
- Few-shot 1-shot: F1 = 67.3 / Acc = 69.0
- Few-shot 5-shot: F1 = 68.1 / Acc = 68.7

-MPT Instruct:

- Hiệu suất rất thấp ở mọi chế độ. Ví dụ:
  - Zero-shot full: F1 = 24.8
  - Few-shot 5-shot CoT: F1 = 33.9 → Không phù hợp để làm annotator tài chính.

-Ensemble Models (Tổ hợp):

- GPT-4 + PaLM 2 hoặc GPT-4 + MPT: đạt hiệu suất cao nhất
  - Few-shot 5-shot CoT: F1 = 68.6 / Acc = 65.2 → Việc tổ hợp các LLM có thể cải thiện hiệu suất và độ tin cậy

Annotator	Type	Temperature Setting	Micro-Averaged F1 Score/ Accuracy(%)					
			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/ <b>65.4</b>
	GPT-4	0.7	<b>67.6/63.6</b>	68.4/64.6	65.0/60.0	67.7/63.9	64.6/ <b>58.4</b>	68.4/ <b>65.4</b>
	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	<b>67.3/60.9</b>	<b>68.7/63.8</b>	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	<b>68.8/63.8</b>
	Ensemble (GPT-4 w Palm 2)	0.2	67.2/63.2	<b>68.6/64.7</b>	65.0/60.1	<b>67.8/64.0</b>	64.3/58.1	68.2/65.2
	Ensemble (GPT-4 w MPT Instruct)	0.2	67.2/63.2	<b>68.6/64.7</b>	65.0/60.1	<b>67.8/64.0</b>	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.

# Độ chính xác của LLM so với con người

## 1. GPT-4 và PaLM 2:

- Vượt trội hơn các crowdworkers với độ chính xác cao hơn đến 29%.
- Cả hai mô hình cho kết quả tốt nhất khi sử dụng các prompt đầy đủ hướng dẫn hoặc phương pháp Few-shot CoT.

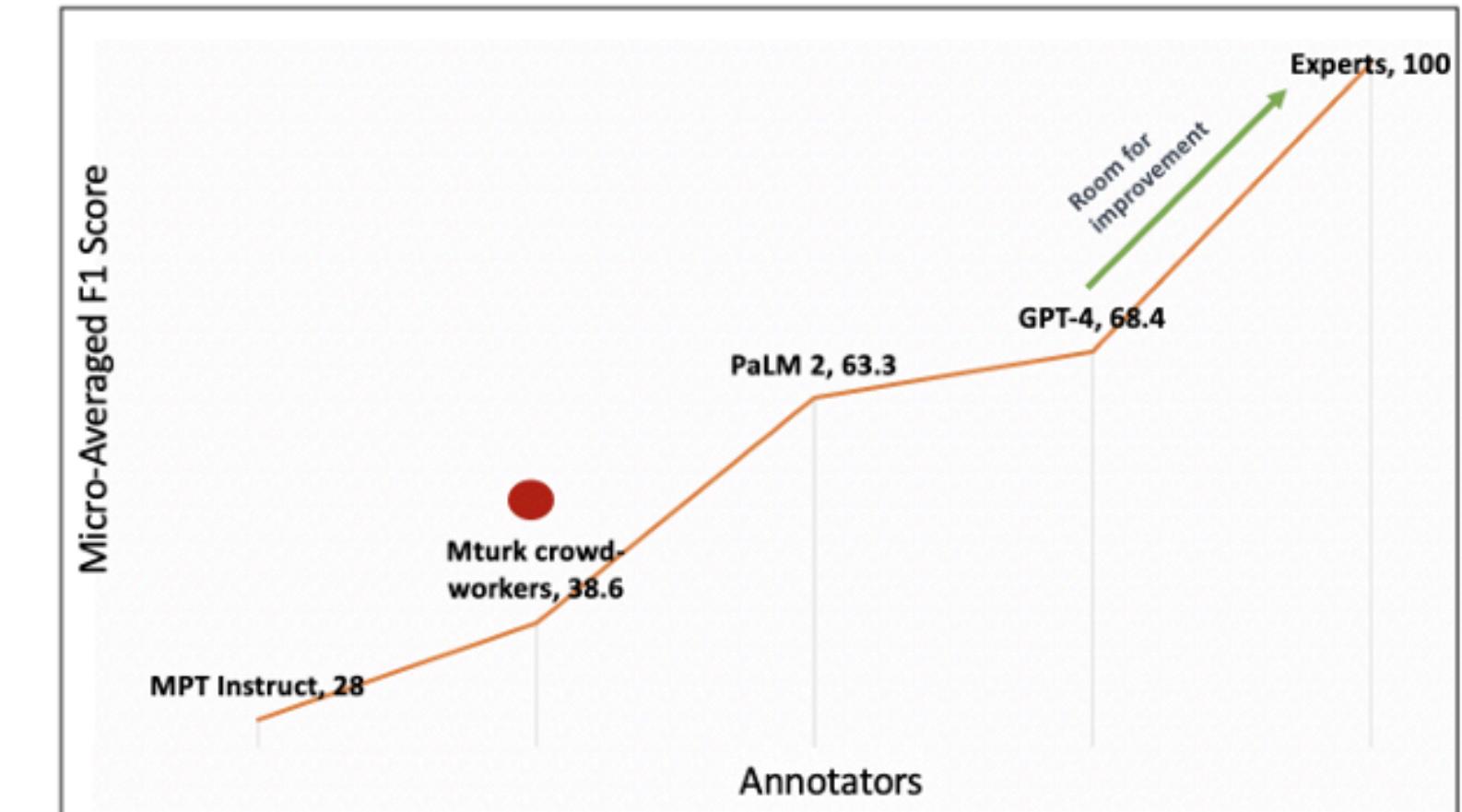


Figure 3: Annotator performance in terms of micro-averaged F1-Score under *full instruction prompt*.

# Độ chính xác của LLM so với con người

## 2. MPT Instruct:

- Mặc dù có hiệu suất thấp hơn so với GPT-4 và PaLM 2, nhưng vẫn cho kết quả tốt hơn so với con người trong nhiều trường hợp.
- Đặc biệt hiệu quả hơn khi sử dụng prompt dạng Few-shot hoặc CoT.

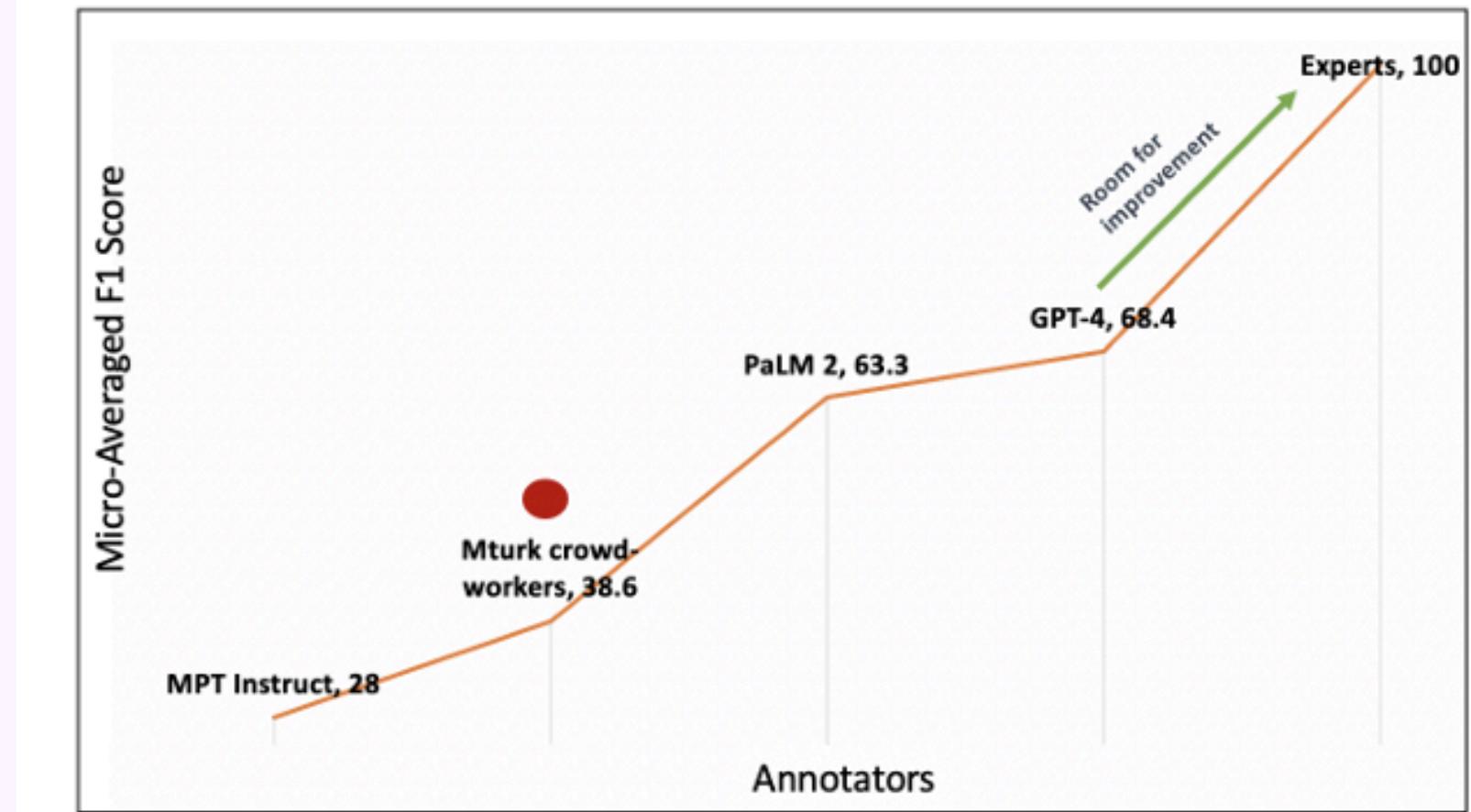


Figure 3: Annotator performance in terms of micro-averaged F1-Score under *full instruction prompt*.

## Đồng thuận giữa các người đánh giá

1. Không mô hình nào hoàn toàn giống nhau về đầu ra
- Tất cả các LLM (kể cả mạnh như GPT-4) không tạo ra đầu ra trùng khớp hoàn toàn.
  - GPT-4 và PaLM 2 có mức độ đồng thuận cao, nghĩa là chúng ổn định khi lặp lại.
  - MPT thì thiếu ổn định, cho ra kết quả khác biệt đáng kể khi thay đổi seed (hạt giống ngẫu nhiên).

2. Ảnh hưởng của cài đặt nhiệt độ (temperature)

- GPT-4 vẫn giữ mức độ đồng thuận cao ngay cả khi thay đổi thông số temperature.
- PaLM 2 và MPT giảm độ đồng thuận rõ rệt  
→ Không ổn định khi thay đổi thiết lập.

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

## Đồng thuận giữa các người đánh giá

### 3. Ảnh hưởng của lời nhắc (prompt)

- Việc sử dụng lời nhắc khác nhau làm giảm sự đồng thuận giữa các đầu ra trên tất cả các LLM.
- Được đo bằng Cohen's Kappa (giữa hai lời nhắc) và Fleiss Kappa (giữa nhiều lời nhắc).

### 4. Kết luận

- GPT-4 và PaLM 2 thể hiện độ tin cậy cao trên các thiết lập khác nhau.
- Sự lựa chọn lời nhắc có tác động mạnh mẽ nhất đến kết quả của mô hình.
- Điều này cho thấy việc thiết kế prompt tốt là cực kỳ quan trọng để đảm bảo hiệu suất ổn định của LLMs trong nhiệm vụ gán nhãn.

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

# LLM-RelIndex Based Accuracy Analysis

- LLM-RelIndex:
  - Là một chỉ số đo lường mức độ tin cậy của dự đoán từ LLMs dựa trên sự nhất quán và mức độ tự tin của mô hình
  - Các dự đoán có LLM-RelIndex cao thường ít lỗi hơn và không cần sự can thiệp từ con người.
- Thay thế Công Việc của Crowdworkers:
  - LLMs có thể thay thế khoảng 65% công việc của crowdworkers trong các nhiệm vụ gán nhãn dữ liệu tài chính.
  - → Điều này giúp tiết kiệm thời gian và chi phí đáng kể
- Vai trò của chuyên gia
  - Khoảng 35% các trường hợp còn lại, đặc biệt là các tình huống phức tạp hoặc mơ hồ, vẫn cần đến sự can thiệp của chuyên gia để đảm bảo độ chính xác.

# LLM-RelIndex Based Accuracy Analysis

- Trong thiết lập Zero-shot, các mô hình lớn như GPT-4 đã thể hiện khả năng vượt qua con người khi được cung cấp hướng dẫn đầy đủ.
- Prompting (đặc biệt là “Full Instruction”) đóng vai trò quan trọng trong việc tăng hiệu suất, đặc biệt đối với các mô hình như PaLM 2.
- MPT không phù hợp với nhiệm vụ phức tạp trong Zero-shot setup, và cần cải tiến đáng kể để đạt hiệu suất chấp nhận được.

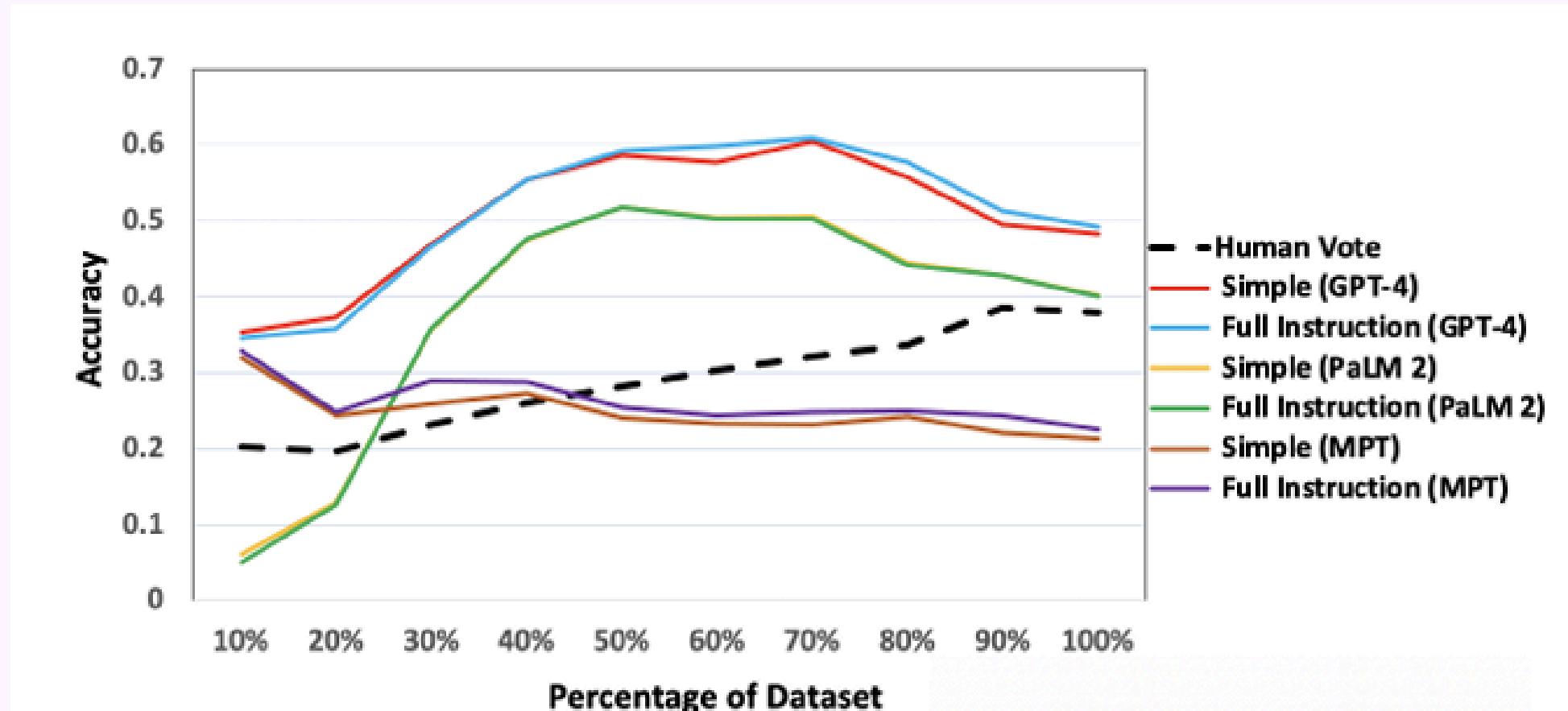


Figure 5: Human vs LLMs at Zero-shot using *LLM-RelIndex*

# LLM-RelIndex Based Accuracy Analysis

- Few-shot prompting rõ ràng mang lại lợi ích lớn cho độ chính xác, đặc biệt với các mô hình mạnh như GPT-4 và PaLM 2.
- 5-shot > 1-shot ở hầu hết các mô hình, cho thấy việc cung cấp thêm ví dụ rất hữu ích.
- MPT không phù hợp cho các tác vụ yêu cầu suy luận phức tạp, ngay cả với Few-shot.
- Dù con người có độ ổn định cao, LLMs (nếu được prompt tốt) vẫn có thể vượt trội hơn

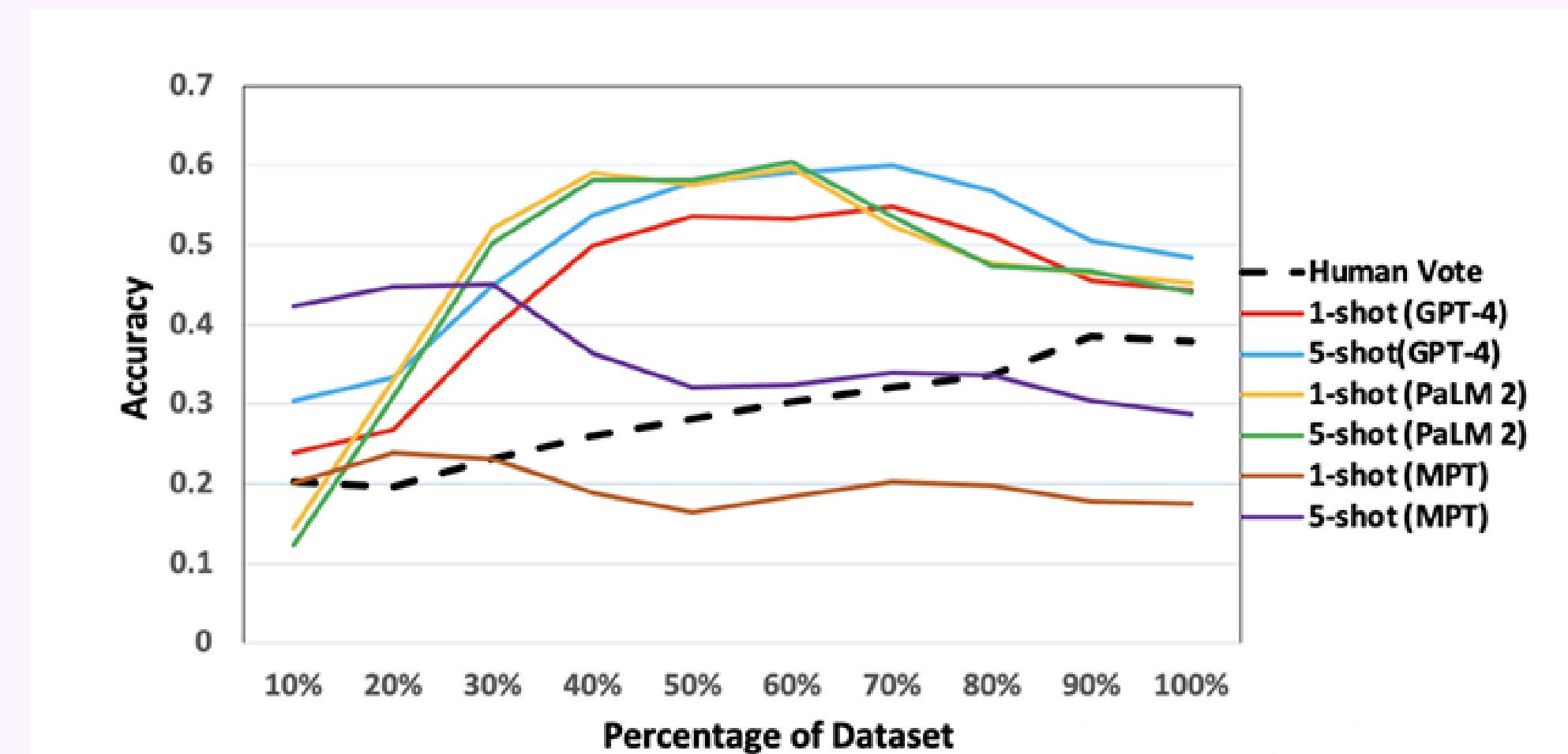


Figure 6: Human vs LLMs at Few-shot using *LLM-RelIndex*

# LLM-RelIndex Based Accuracy Analysis

## Kết luận

- GPT-4 5-shot CoT là phương pháp hiệu quả nhất, vượt xa con người và các mô hình khác
- PaLM 2 có tiềm năng lớn khi sử dụng 5-shot, nhưng vẫn kém GPT-4.
- Con người giữ được mức độ ổn định hơn khi xử lý dữ liệu lớn, nhưng bị vượt qua ở các mức dữ liệu thấp.
- MPT không phù hợp cho các nhiệm vụ phức tạp, đặc biệt khi so sánh với GPT-4 và PaLM 2

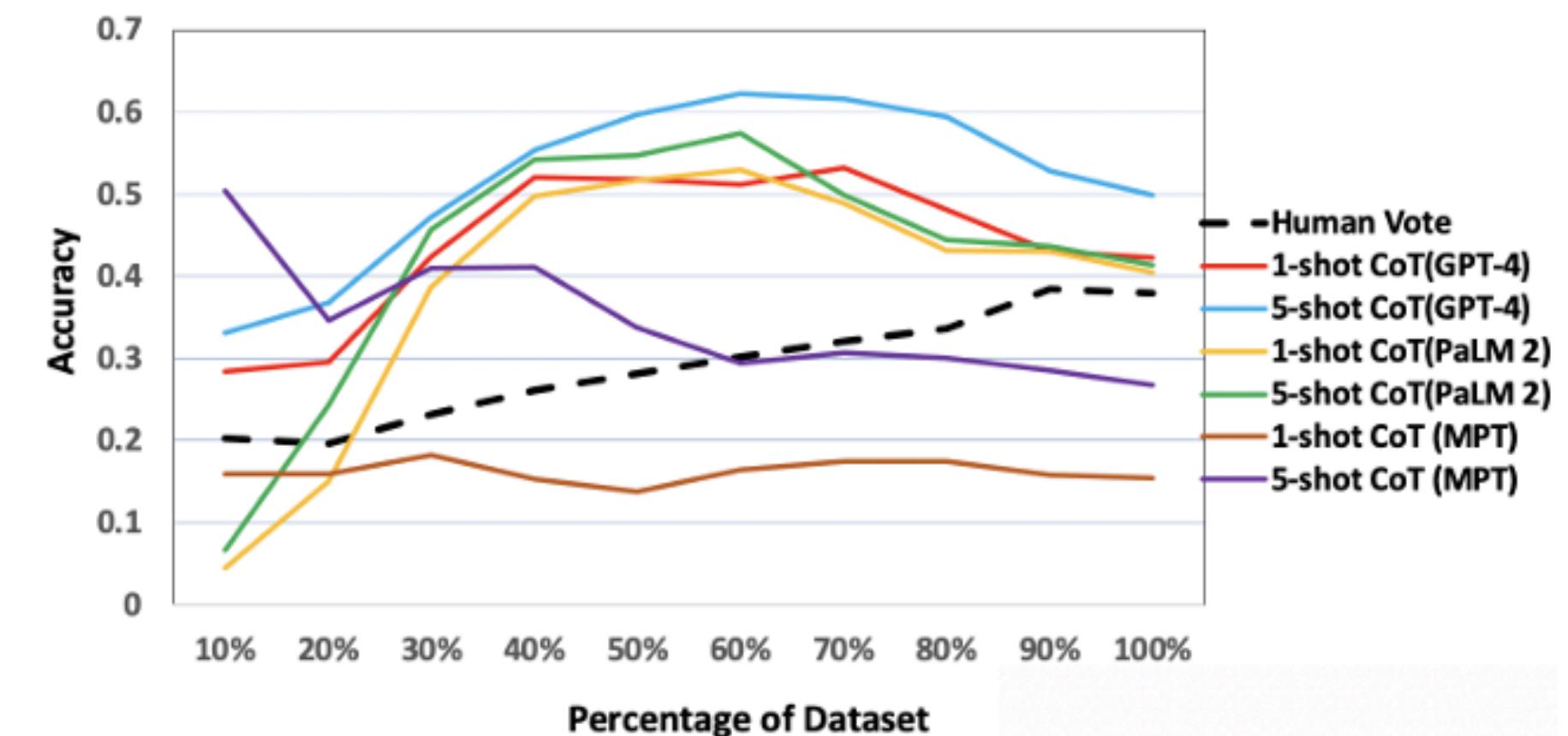


Figure 7: Human vs LLMs at Few-shot CoT using *LLM-RelIndex*

## Nhược điểm

1. Nhầm lẫn giữa các quan hệ gần giống nhau:

- LLMs có thể nhầm lẫn các mối quan hệ tương tự như "Member Of" và "Employee Of".
- Nguyên nhân do sự tương đồng ngữ nghĩa và ngữ cảnh không rõ ràng.

# Nhược điểm

## 2. Hallucination (Ảo tưởng thông tin):

- LLMs có thể tạo ra các quan hệ không có thật hoặc không tồn tại trong văn bản.
- Thường xảy ra khi mô hình quá tin vào kết quả dự đoán mà không có bằng chứng rõ ràng.

# Ứng dụng thực tiễn

## 1.Tự động Gán Nhãn Dữ Liệu Tài Chính

- LLMs có thể thay thế phần lớn công việc của crowdworkers trong việc gán nhãn quan hệ tài chính từ các báo cáo và văn bản tài chính.
- Điều này giúp tiết kiệm thời gian và giảm chi phí đáng kể so với việc thuê nhiều người gán nhãn.

## 2.Tăng Cường Chất Lượng Dữ Liệu Đầu Vào

- LLMs có thể tạo ra các nhãn chất lượng cao hơn bằng cách giảm thiểu các lỗi do con người gây ra, đặc biệt là trong các tác vụ quy mô lớn.

# Ứng dụng thực tiễn

## 3. Kết Hợp LLMs và Chuyên Gia

- Một mô hình kết hợp có thể được áp dụng, trong đó LLMs thực hiện các nhiệm vụ đơn giản hoặc có độ tin cậy cao, còn chuyên gia xử lý các trường hợp phức tạp hoặc có độ tin cậy thấp.
- Sử dụng Reliability Index (LLM-RelIndex) để xác định khi nào cần chuyên gia can thiệp.

# Hướng phát triển

## 1. Mở rộng sang nhiều data set khác

- Thử nghiệm trên các tập dữ liệu tài chính khác như tin tức tài chính, báo cáo thu nhập, hoặc các bài phân tích thị trường
- → Giúp đánh giá độ tổng quát và khả năng thích ứng của LLMs trong các ngữ cảnh thực tế hơn.

## 2. Kết Hợp Đa Mô Hình

- Sử dụng phương pháp ensemble (kết hợp nhiều mô hình) để tăng cường độ chính xác và giảm thiểu lỗi dự đoán.

# Hướng phát triển

## 3.Cải Thiện Độ Chính Xác Bằng Fine-Tuning

- Fine-tuning LLMs trên các tập dữ liệu tài chính chuyên biệt có thể cải thiện độ chính xác, đặc biệt trong các nhiệm vụ trích xuất quan hệ phức tạp.
- Sử dụng các kỹ thuật như Instruction Fine-tuning hoặc Domain Adaptation để tối ưu hóa hiệu suất.Tăng Cường Chất Lượng Dữ Liệu Đầu Vào



# Tổng kết nghiên cứu

## 1. LLMs có thể thay thế Crowdworkers

- LLMs như GPT-4 và PaLM 2 đã chứng minh khả năng vượt trội trong việc gán nhãn dữ liệu tài chính, đạt độ chính xác cao hơn crowdworkers đến 29%
- → LLMs là một lựa chọn hiệu quả để thay thế phần lớn công việc gán nhãn mà không cần sự tham gia trực tiếp từ crowdworkers



# Tổng kết nghiên cứu

## 2.Không thể thay thế chuyên gia

- LMMs hoạt động tốt trong các trường hợp đơn giản và trung bình nhưng với các trường hợp phức tạp hoặc mơ hồ vẫn cần các chuyên gia để đảm bảo độ chính xác
- Khoảng 35% các trường hợp khó vẫn cần chuyên gia



# Tổng kết nghiên cứu

## 3.Kết hợp LLMs và con người là phương pháp tối ưu

- Việc sử dụng LLMs để xử lý các tác vụ đơn giản và trung bình, kết hợp với chuyên gia xử lý các trường hợp phức tạp, là cách tiếp cận tối ưu.
- → Điều này giúp tiết kiệm thời gian, chi phí còn đảm bảo độ chính xác cao nhất có thể



# Thank You!

