ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN NHÓM 11

MÔN HỌC: SEMINAR KHOA HỌC

ĐỀ TÀI: PIXIU - A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance

Giảng viên giảng dạy: TS Trần Hồng Việt

Sinh viên thực hiện:

Hoàng Mạnh Hùng - 23020371

Đặng Đức Duy - 23020347

Phạm Trung Hiếu - 23020367

Trịnh Hoàng Đức - 23020359

PIXIU: Mô hình ngôn ngữ lớn, dữ liệu hướng dẫn và Tiêu chuẩn đánh giá tài chính

Qianqian Xie
School of Computer Science
Wuhan University
Wuhan, Hubei, China
xieq@whu.edu.cn

Weiguang Han
School of Computer Science
Wuhan University
Wuhan, Hubei, China
han.wei.guang@whu.edu.cn

Xiao Zhang
Sun Yat-Sen University
Shenzhen, Guangdong, China
zhangx767@mail2.sysu.edu.cn

Yanzhao Lai
School of Economics and Management
Southwest Jiaotong University
Chengdu, Sichuan, China
laiyanzhao@swjtu.edu.cn

Min Peng
School of Computer Science
Wuhan University
Wuhan, Hubei, China
pengm@whu.edu.cn

Alejandro Lopez-Lira
University of Florida
alejandro.lopez-lira@warrington.ufl.edu

Jimin Huang
ChanceFocus AMC.
Shanghai, China
jimin@chancefocus.com

Trừu tượng

Mặc dù các mô hình ngôn ngữ lớn (LLM) đã cho thấy hiệu suất tuyệt vời trong xử lý ngôn ngữ tự nhiên (NLP) trong lĩnh vực tài chính, nhưng không có LLM được thiết kế riêng về mặt tài chính, bộ dữ liệu điều chỉnh hướng dẫn và điểm chuẩn đánh giá, điều này rất quan trọng để liên tục thúc đẩy sự phát triển mã nguồn mở của trí tuệ nhân tạo tài chính (AI). Bài báo này giới thiệu PIXIU, một khuôn khổ toàn diện bao gồm LLM tài chính đầu tiên dựa trên việc tinh chỉnh LLaMA với dữ liệu hướng dẫn, dữ liệu hướng dẫn đầu tiên với 128K mẫu dữ liệu để hỗ trợ tinh chỉnh và điểm chuẩn đánh giá với 8 nhiệm vụ và 15 bộ dữ liệu. Trước tiên, chúng tôi xây dựng dữ liệu hướng dẫn đa nhiệm quy mô lớn xem xét nhiều nhiệm vụ tài chính, loại tài liệu tài chính và phương thức dữ liệu tài chính. Sau đó, chúng tôi đề xuất một LLM tài chính được gọi là FinMA bằng cách tinh chỉnh LLaMA với bộ dữ liệu được xây dựng để có thể làm theo hướng dẫn cho các nhiệm vụ tài chính khác nhau. Để hỗ trợ đánh giá LLM tài chính, chúng tôi đề xuất một tiêu chuẩn tiêu chuẩn bao gồm tài sản của các nhiệm vụ tài chính quan trọng, bao gồm sáu nhiệm vụ NLP tài chính và hai nhiệm vụ dự đoán tài chính. Với tiêu chuẩn này, chúng tôi tiến hành phân tích chi tiết về FinMA và một số LLM hiện có, khám phá điểm mạnh và điểm yếu của họ trong việc xử lý các nhiệm vụ tài chính quan trọng. Mô hình, bộ dữ liệu, điểm chuẩn và kết quả thử nghiệm là mã nguồn mở[1] để tạo điều kiện thuận lợi cho nghiên cứu trong tương lai về AI tài chính.

1 Giới thiệu

[1] https://github.com/chancefocus/PIXIU

Hội nghị lần thứ 37 về Hệ thống xử lý thông tin thần kinh (NeurIPS 2023) theo dõi bộ dữ liệu và điểm chuẩn.

Công nghệ tài chính (FinTech) đã liên tục được phát triển bởi sự phát triển của các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và máy học (ML), mở khóa khả năng đa dạng từ dự đoán biến động giá cổ phiếu đến phân tích tài chính nâng cao (Araci, 2019; Han và cộng sự, 2023; Xie và cộng sự, 2023; Lopez-Lira và Tang, 2023; Li và cộng sự, 2023). Cụ thể, các mô hình ngôn ngữ lớn (LLM) gần đây nhất (Brown et al., 2020)[2] đã thể hiện khả năng đáng chú ý trong việc hiểu ngôn ngữ tự nhiên (NLU) và thực hiện các nhiệm vụ khác nhau bằng cách làm theo các hướng dẫn ngôn ngữ tự nhiên mà không cần dữ liệu đào tạo. Bất chấp những thành công này, bản chất kỹ thuật cao của các văn bản tài chính đòi hỏi các LLM theo lĩnh vực cụ thể để hiểu ngôn ngữ và khái niệm tài chính phức tạp một cách hiệu quả. Những nỗ lực như vậy bao gồm các mô hình ngôn ngữ được đào tạo trước (PLM) tài chính hiện có như finBERT (Araci, 2019), FinBERT (Yang và cộng sự, 2020) và FLANG (Shah và cộng sự, 2022). Tuy nhiên, những mô hình đó được coi là nhỏ vì kích thước tham số của chúng dưới một tỷ, hạn chế khả năng khái quát hóa của chúng. Gần đây, một LLM tài chính độc quyền có tên là BloombergGPT (Wu và cộng sự, 2023) với 50 tỷ tham số đã được đề xuất bằng cách đào tạo trước LLM kiểu Bloom (Scao và cộng sự, 2022) trên dữ liệu tài chính quy mô lớn.

Bất chấp những nỗ lực này, Liu et al. (2023), vẫn còn một số vấn đề, như thể hiện trong Bảng 1. Thứ nhất, BloombergGPT và dữ liệu đào tạo của nó không được công bố công khai. Hiện tại, không có LLM tài chính mã nguồn mở, có thể cản trở sự phát triển trong cộng đồng nghiên cứu. Thứ hai, các PLM tài chính trước đây và BloombergGPT mới nhất không được tinh chỉnh để tuân theo các hướng dẫn ngôn ngữ tự nhiên (còn được gọi là điều chỉnh hướng dẫn), điều này rất quan trọng để cải thiện khả năng không bắn khi xử lý các nhiệm vụ tài chính hạ nguồn (Wei và cộng sự, 2021; Ouyang và cộng sự, 2022). Thứ ba, cũng không có dữ liệu hướng dẫn tài chính để hỗ trợ điều chỉnh hướng dẫn của LLM và các tiêu chuẩn đánh giá để đánh giá và so sánh toàn diện khả năng của LLM cho các nhiệm vụ tài chính. Do đó, chúng tôi có động lực để xem xét các câu hỏi nghiên cứu sau: 1) làm thế nào chúng tôi có thể phát triển các LLM hiệu quả và có sẵn công khai phù hợp với tài chính? 2) Làm thế nào chúng ta có thể xây dựng dữ liệu hướng dẫn tài chính quy mô lớn và chất lượng cao? 3) Làm thế nào chúng ta có thể xây dựng tiêu chuẩn đánh giá tài chính toàn diện để đánh giá LLM tài chính?

Bảng 1: So sánh các mô hình ngôn ngữ được đào tạo trước và các mô hình ngôn ngữ lớn cho tài chính." Hướng dẫn" có nghĩa là liệu mô hình có thể làm theo hướng dẫn hay không. "NLP" và "Fin" có nghĩa là nếu mô hình được đánh giá với các nhiệm vụ NLP tài chính và nhiệm vụ dự đoán tài chính.

Model	Backbone	Size	Open Source		Instruct	Language	Evaluation		Release Date
			Model	Data			NLP	Fin	
finBERT (Araci, 2019)	BERT	110M	✓	✓	✗	English	✓	✗	08/27/19
FinBERT (Yang et al., 2020)	BERT	110M	✓	✗	✗	English	✓	✗	06/15/20
Mengzi-fin (Zhang et al., 2021)	RoBERTa	103M	✓	✗	✗	Chinese	✓	✗	10/13/21
FLANG (Shah et al., 2022)	ELECTRA	110M	✓	✓	✗	English	✓	✗	10/31/22
BBT-FinT5 (Lu et al., 2023)	T5	220M	✓	✓	✗	Chinese	✓	✗	02/18/23
BloombergGPT (Wu et al., 2023)	BLOOM	50B	✗	✗	✗	English	✓	✗	03/30/23
FinMA	LLaMA	7/30B	✓	✓	✓	English	✓	✓	06/01/23

Để giải quyết những câu hỏi nghiên cứu này, chúng tôi đề xuất PIXIU (貔貅)[3], một khuôn khổ toàn diện bao gồm LLM tài chính được tinh chỉnh mã nguồn mở đầu tiên, FinMA, dựa trên tinh chỉnhLLaMA (Touvron et al., 2023) với dữ liệu hướng dẫn đa tác vụ và đa phương thức. Hình 1 trình bày tổng quan về điều chỉnh hướng dẫn đa tác vụ và đa phương thức của FinMA cho các nhiệm vụ tài chính đa dạng. PIXIU cũng chứa dữ liệu hướng dẫn đầu tiên với 128K mẫu dữ liệu để hỗ trợ tinh chỉnh và đánh giá tổng thể với sáu nhiệm vụ NLP tài chính và hai nhiệm vụ dự đoán tài chính. Nó có các đặc điểm phân biệt sau:

[2] https://openai.com/blog/chatgpt

[3] (貔貅)https://en.wikipedia.org/wiki/Pixiu là một sinh vật thần thoại trong văn hóa dân gian Trung Quốc. Nó có đầu của một con rồng và cơ thể của một con sư tử và được cho là một sinh vật tốt lành thu hút tiền và may mắn.

- Tài nguyên mở. Chúng tôi đã công khai phát hành LLM tài chính, dữ liệu điều chỉnh hướng dẫn và bộ dữ liệu có trong tiêu chuẩn đánh giá và triển khai, để khuyến khích nghiên cứu mở và minh bạch trong lĩnh vực nghiên cứu.
- Đa nhiệm. PIXIU bao gồm dữ liệu điều chỉnh lệnh đa nhiệm bao gồm một tập hợp các nhiệm vụ tài chính đa dạng, bao gồm sáu nhiệm vụ NLP tài chính và hai nhiệm vụ dự đoán tài chính. Các Điều chỉnh hướng dẫn đa nhiệm đã được chứng minh là rất quan trọng để cải thiện khả năng tổng quát hóa của mô hình (Sanh và cộng sự, 2022; Longpre và cộng sự, 2023) cho các nhiệm vụ mới.
- Đa phương thức. Dữ liệu điều chỉnh hướng dẫn của chúng tôi bao gồm dữ liệu tài chính đa phương thức như bảng trong báo cáo tài chính và giá cổ phiếu lịch sử làm dữ liệu chuỗi thời gian cho các nhiệm vụ dự đoán chuyển động cổ phiếu ngoài văn bản. Hơn nữa, chúng bao gồm nhiều loại văn bản tài chính khác nhau, bao gồm báo cáo, bài báo, tweet và hồ sơ quy định.
- Đa dạng. So với các nhiệm vụ đánh giá được sử dụng trong BloombergGPT và FLUEbenchmark hiện có (Shah et al., 2022), chủ yếu bao gồm các nhiệm vụ NLP tài chính, benchmark đánh giá của chúng tôi bao gồm các nhiệm vụ dự đoán tài chính như dự đoán biến động cổ phiếu và tính tín dụng. Nó yêu cầu mô hình khai thác triệt để cả văn bản tự nhiên và dữ liệu chuỗi thời gian để trích xuất thông tin cần thiết để dự đoán chính xác. So với các nhiệm vụ NLP tài chính, nhiệm vụ dự đoán tài chính phù hợp hơn với các kịch bản trong thế giới thực và khó khăn hơn.

Để xây dựng dữ liệu hướng dẫn đa nhiệm và đa phương thức, chúng tôi thu thập dữ liệu đào tạo được phát hành mở từ các nhiệm vụ đa dạng, bao gồm phân tích tâm lý tài chính, phân loại tiêu đề tin tức, nhận dạng thực thể được đặt tên (NER), trả lời câu hỏi, tóm tắt văn bản, dự đoán chuyển động cổ phiếu, chấm điểm tín dụng và phân loại diều hâu, đồng thời đề xuất các hướng dẫn cụ thể theo nhiệm vụ đa dạng được viết bởi các chuyên gia lĩnh vực cho từng nhiệm vụ. Chúng tôi tạo ra dữ liệu điều chỉnh hướng dẫn tài chính quy mô lớn (FIT) bằng cách tập hợp các hướng dẫn cụ thể của nhiệm vụ với các mẫu dữ liệu từ mỗi nhiệm vụ. Do đó, chúng tôi đề xuất LLM FinMA dành riêng cho miền bằng cách tiến hành điều chỉnh lệnh đa tác vụ trên LLaMA với tập dữ liệu xây dựng. Để đánh giá mô hình của chúng tôi và các LLM khác một cách toàn diện, chúng tôi xây dựng Điểm chuẩn đánh giá Hiểu biết và PRediction Ngôn ngữ Tài chính (FLARE) bao gồm 6 nhiệm vụ NLP tài chính với 10 bộ dữ liệu và 2 nhiệm vụ dự đoán tài chính với 5 bộ dữ liệu.

Dựa trên FLARE, chúng tôi đánh giá hiệu suất của mô hình của mình, BloombergGPT và LLM nâng cao trong miền chung, chẳng hạn như ChatGPT[4] và GPT-4 (OpenAI, 2023). Kết quả thử nghiệm cho thấy: 1) FinMA vượt trội hơn đáng kể các LLM, bao gồm BloombergGPT, ChatGPT và GPT-4 trên hầu hết các tác vụ trong FLARE, bao gồm phân tích tâm lý tài chính, phân loại tiêu đề tin tức và dự đoán biến động chứng khoán. Điều này cho thấy tầm quan trọng của việc điều chỉnh các LLM dành riêng cho lĩnh vực tài chính. 2) Mặc dù có kết quả đầy hứa hẹn trên hầu hết các tác vụ, FinMA hoạt động kém hơn BloombergGPT, ChatGPT và GPT-4 về câu trả lời, đánh giá khả năng suy luận định lượng của LLM. Phân tích của chúng tôi cho thấy rằng điều này là do hạn chế của LLaMA về lý luận định lượng và toán học. FinMA cũng cho thấy hiệu suất hạn chế trên các tác vụ NER mặc dù nó vượt trội hơn BloomerbergGPT, điều này cũng là do những hạn chế của LLaMA. 3) So với các nhiệm vụ NLP, tất cả LLM, bao gồm FinMA, ChatGPT và GPT-4, vẫn có hiệu suất hạn chế trong dự đoán biến động cổ phiếu, cho thấy khả năng cải thiện hơn nữa. 4) FinMA được tinh chỉnh với cả NLP và nhiệm vụ dự đoán tài chính, trình bày hiệu suất tốt nhất trên một trong các bộ dữ liệu dự đoán cổ phiếu, cho thấy tiềm năng của việc điều chỉnh hướng dẫn cụ thể của LLM đối với các nhiệm vụ dự đoán tài chính.

Những đóng góp của chúng tôi có thể được tóm tắt như sau: 1) Chúng tôi giới thiệu FIT, dữ liệu điều chỉnh lệnh đa tác vụ và đa phương thức đầu tiên trong lĩnh vực tài chính, bao gồm 5 nhiệm vụ và 9 bộ dữ liệu với 128.640 (128K) mẫu dữ liệu. 2) Chúng tôi giới thiệu FLARE, tiêu chuẩn đánh giá đầu tiên với cả hai các nhiệm vụ hiểu và dự đoán ngôn ngữ tự nhiên tài chính. 3) Chúng tôi giới thiệu FinMA, mô hình ngôn ngữ lớn tài chính được phát hành công khai và theo hướng dẫn đầu tiên, đạt được SOTA trên 6 nhiệm vụ NLP tài chính và 2 nhiệm vụ dự đoán tài chính. 4) Chúng tôi so sánh FinMA và các LLM hiện có trên FLARE. Kết quả cho thấy sự vượt trội của FinMA, những hạn chế chính của LLM đối với tài chính và định hướng trong tương lai để thúc đẩy LLM cho tài chính.

2 Công việc liên quan

[4] https://openai.com/blog/chatgpt

Mô hình ngôn ngữ tài chính PLMs cho lĩnh vực tài chính đã được đề xuất bởi các PLM đào tạo trước liên tục với các văn bản tài chính quy mô lớn. Araci (2019) đã đề xuất PLM tài chính đầu tiên được gọi là finBERT đã đào tạo trước BERT (Devlin et al., 2019) với kho tài chính được phát hành mởchẳng hạn như TRC2-financial[5] và Financial Phrase Bank (Malo et al., 2014). finBERT vượt trội hơn các phương pháp mạng nơ-ron như LSTM trong các nhiệm vụ phân loại tâm lý tài chính. Yang et al. (2020) tiếp tục đề xuất FinBERT bằng cách đào tạo trước BERT với giao tiếp tài chính 4,9 tỷ mã thông báo corpus, vượt trội hơn BERT trên ba bộ dữ liệu phân loại tâm lý tài chính. Shah và cộng sự (2022) đề xuất FLANG, một PLM tài chính với BERT và ELECTRA (Clark và cộng sự, 2020) làm xương sống. Bên cạnh tiếng Anh, PLM tài chính bằng các ngôn ngữ khác, chẳng hạn như tiếng Trung, cũng được đề xuất, chẳng hạn như Mengzi-fin (Zhang và cộng sự, 2021) và BBT-FinT5 (Lu và cộng sự, 2023). Mới nhất, Wu et al. (2023) đề xuất BloombergGPT, mô hình ngôn ngữ lớn tài chính đầu tiên với 50 tỷ tham số, được đào tạo trước với các bộ dữ liệu hỗn hợp từ lĩnh vực chung và tài chính. Tuy nhiên, cả mô hình và bộ dữ liệu miền được đào tạo trước đều không được phát hành. Mô hình này cũng không tuân theo hướng dẫn như các LLM khác như ChatGPT và GPT-4.

Tiêu chuẩn đánh giá tài chính Shah et al. (2022) đã đề xuất đánh giá không đồng nhất đầu tiênBenchmark FLUE với 5 nhiệm vụ NLP tài chính, bao gồm phân tích tâm lý tài chính (Malo et al., 2014), phân loại tiêu đề tin tức (Sinha và Khandait, 2021), nhận dạng thực thể được đặt tên (Alvaradoet al., 2015), phát hiện ranh giới cấu trúc[6] và trả lời câu hỏi (Maia et al., 2018). Lu et al. (2023) đã đề xuất tiêu chuẩn đánh giá tài chính đầu tiên của Trung Quốc BBT-CFLEB[7] với nhiệm vụ phân loại, tóm tắt, trích xuất quan hệ, trả lời câu hỏi và xác định tin tức tiêu cực, cũng như nhiệm vụ phân loại cảm xúc của các văn bản truyền thông xã hội tài chính. Tuy nhiên, các điểm chuẩn này chỉ xem xét các nhiệm vụ NLP tài chính và không bao gồm các nhiệm vụ dự đoán tài chính, chẳng hạn như dự đoán biến động cổ phiếu, rất quan trọng để đánh giá hiệu suất của mô hình áp dụng cho các kịch bản trong thế giới thực.

Mô Hình Ngôn Ngữ Lớn Mã Nguồn Mở Các nghiên cứu gần đây đã nỗ lực về AI dân chủ, trong đó công việc đại diện là LLaMA (Touvron và cộng sự, 2023) từ Meta AI, một LLM mã nguồn mở với các thông số từ 7B và 13B đến 65B. LLaMA-13B có hiệu suất tương đương và thậm chí tốt hơn GPT-3 (Brown và cộng sự, 2020) với 175B tham số về các nhiệm vụ suy luận thông thường. Những nỗ lực sau đây đã được đề xuất để cải thiện LLaMA cho các hướng dẫn theo dõi như ChatGPT, điều chỉnh hướng dẫn. Chẳng hạn như Taori và cộng sự (2023) đã đề xuất Alpaca bằng cách tinh chỉnh LLaMA-7B với 52K mẫu theo hướng dẫn được tạo bằng phương pháp tự hướng dẫn (Wang và cộng sự, 2022). Chianget al. (2023) đã đề xuất Vicuna-13B bằng cách tinh chỉnh LLaMA-13B với 70K dữ liệu cuộc trò chuyện từ ShareGPT[8]. Nó có thể tạo ra câu trả lời tốt hơn cho câu hỏi của người dùng so với Alpaca. Tuy nhiên, không có LLM mã nguồn mở và dữ liệu điều chỉnh hướng dẫn tập trung vào lĩnh vực tài chính.

3. FIT: Bộ dữ liệu điều chỉnh hướng dẫn tài chính

Trong phần này, chúng tôi giới thiệu bộ dữ liệu điều chỉnh hướng dẫn tài chính FIT, bao gồm nền tảng của dữ liệu thô, nhiệm vụ trong FIT và quá trình xây dựng dựa trên dữ liệu thô. Khác với các bộ dữ liệu tài chính hiện có, FIT là bộ dữ liệu điều chỉnh hướng dẫn đầu tiên cho LLM tài chính và bao gồm các nhiệm vụ dự đoán tài chính ngoại trừ các nhiệm vụ NLP tài chính, vốn là nền tảng cho các ứng dụng tài chính trong thế giới thực.

3.1 Dữ liệu thô

Bắt nguồn từ các kịch bản tài chính trong thế giới thực, chúng tôi xây dựng bộ dữ liệu điều chỉnh hướng dẫn tài chính FIT dựa trên dữ liệu nguồn mở của các nhiệm vụ dự đoán và NLP tài chính khác nhau. So với phương pháp tự hướng dẫn (Wang et al., 2022) thường được sử dụng bởi các LLM hiện có như Alpaca, chúng tôi chọn xây dựng bộ dữ liệu điều chỉnh lệnh từ các bộ dữ liệu mã nguồn mở vì những lý do sau: 1) bộ dữ liệu mã nguồn mở thường được chú thích bởi các chuyên gia miền, cho thấy chất lượng cao, 2) nó có chi phí rất thấp và không có giới hạn về việc sử dụng thương mại không giống như các bộ dữ liệu được xây dựng từ ChatGPT hoặc GPT-4, 3) Các bộ dữ liệu mã nguồn mở này bao gồm nhiều loại văn bản như tin tức, báo cáo và tweet, cũng

[5] https://trec.nist.gov/data/reuters/reuters.html

[6] https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsbd-3

[7] https://bbt.ssymmetry.com/evaluation.html

[8] https://sharegpt.com

như đa phương thức bao gồm dữ liệu chuỗi thời gian, bảng và văn bản. Chi tiết[9] của dữ liệu thô và dữ liệu hướng dẫn được hiển thị trong Bảng 2

Phân tích tâm lý tài chính. Nhiệm vụ phân tích tâm lý tài chính từ lâu đã là một nhiệm vụ quan trọng trong lĩnh vực tài chính (Araci, 2019; Yang và cộng sự, 2020), nhằm phân tích thông tin tâm lý của các văn bản tài chính đầu vào. Theo tiêu chuẩn FLUE hiện có (Shah và cộng sự, 2022), chúng tôi sử dụng hai bộ dữ liệu: bộ dữ liệu Ngân hàng Cụm từ Tài chính (FPB) (Malo và cộng sự, 2014) và FiQA-SA (Maia và cộng sự, 2018). FPBbao gồm các câu tiếng Anh từ tin tức tài chính và nhãn cảm xúc của chúng là tích cực, tiêu cực hoặc

Bảng 2: Chi tiết dữ liệu thô và dữ liệu hướng dẫn.

Data	Task	Raw	Instruction	Data Types	Modalities	License
FPB	sentiment analysis	4,845	48,450	news	text	CC BY-SA 3.0
FiQA-SA	sentiment analysis	1,173	11,730	news headlines,tweets	text	Public
Headlines	news headline classification	11,412	11,412	news headlines	text	CC BY-SA 3.0
FOMC	hawkish-dovish classification	496	496	FOMC transcripts	text	CC BY-NC 4.0
NER	named entity recognition	609	6,090	financial agreements	text	CC BY-SA 3.0
FiNER-ORD	named entity recognition	1,080	1,080	news articles	text	CC BY-SA 3.0
FinQA	question answering	8,281	8,281	earnings reports	text,table	MIT License
ConvFinQA	question answering	3,458	3,458	earnings reports	text,table	MIT License
ECTSum	text summarization	495	495	earning call transcipts	text	Public
EDTSum	text summarization	2,000	2,000	news articles	text	Public
BigData22	stock movement prediction	7,168	7,168	tweets,historical prices	text,time series	Public
ACL18	stock movement prediction	27,080	27,080	tweets,historical prices	text,time series	MIT License
CIKM18	stock movement prediction	4,971	4,971	tweets,historical prices	text,time series	Public
German	credit scoring	1,000	1,000	credit records	table	CC BY 4.0
Australia	credit scoring	690	690	credit records	table	CC BY 4.0

trung lập được chú thích bởi các chuyên gia miền. FiQA-SA là một bộ dữ liệu được chấp nhận rộng rãi khác, nhằm mục đích dự đoán cảm xúc của tin tức tài chính tiếng Anh và các bài đăng trên microblog trên thang điểm [-1,1], trong đó 1 có nghĩa là tích cực nhất.

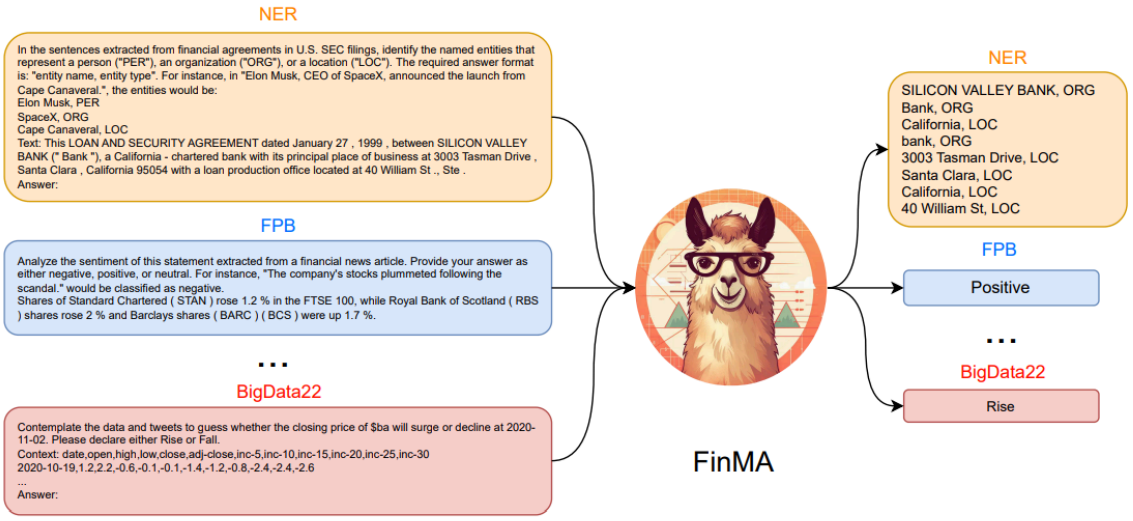
Phân loại tiêu đề tin tức. Nhiệm vụ phân loại tiêu đề tin tức nhằm phân tích các thông tin khác, chẳng hạn như biến động giá trong các văn bản tài chính. Chúng tôi sử dụng bộ dữ liệu tiêu đề tin tức Vàng (Sinha và Khandait, 2021) bao gồm các tiêu đề tin tức từ năm 2000 đến năm 2019 về "vàng" và các thẻ tương ứng của chúng9: "giá hay không", "tăng giá", "giá giảm", "giá ổn định", "giá trong quá khứ", "giá tương lai", "quá khứ chung", "tổng hợp trong tương lai", "so sánh tài sản". Nhiệm vụ là tiến hành phân loại nhị phân cho từng thẻ của mỗi mẫu dữ liệu.

Phân loại Hawkish-dovish. Phân loại Hawkish-Dovish nhằm mục đích phân loại các câu từ các văn bản chính sách tiền tệ thành lập trường 'diều hâu' hoặc 'ôn hòa', không giống như phân tích tâm lý tiêu chuẩn. Chìa khóa của nhiệm vụ này nằm ở việc hiểu ngôn ngữ sắc thái của các văn bản tài chính và có thể xác định các tác động kinh tế được truyền tải thông qua các tín hiệu 'diều hâu' hoặc 'ôn hòa' này. Chúng tôi sử dụng bộ dữ liệu FOMCShah et al. (2023a), bao gồm các câu được trích xuất từ các cuộc họp của Ủy ban Thị trường Mở Liên bang (FOMC), trong đó mỗi câu được chú thích thủ công là ‘Hawkish’ hoặc ‘dovish’

Nhận dạng thực thể được đặt tên. Named Entity Recognition (NER) là phát hiện các thực thể tài chính quan trọng như cá nhân, tổ chức và địa điểm, có thể được sử dụng để xây dựng biểu đồ kiến thức tài chính. Chúng tôi sử dụng hai bộ dữ liệu: NER (Alvarado et al., 2015) và FiNER-ORD Shah et al. (2023b).bao gồm các câu từ các thỏa thuận tài chính công thông qua hồ sơ của Ủy ban An ninh và Giao dịch Hoa Kỳ (SEC), trong khi FiNER-ORD bao gồm các câu từ các bài báo. CÁC THỰC THỂ VỊ TRÍ (LOC), TỔ CHỨC (ORG) và PERSON (PER) của các câu từ cả hai bộ dữ liệu được chú thích thủ công.

[9] Để biết thêm chi tiết về việc phân tách và tiền xử lý dữ liệu, vui lòng tham khảo Phụ lục.

Trả lời câu hỏi. Trả lời câu hỏi là nhiệm vụ tự động trả lời câu hỏi tài chính dựa trên thông tin được cung cấp. Chúng tôi sử dụng hai bộ dữ liệu: FinQA (Chen và cộng sự, 2021) và Con-vFinQA (Chen và cộng sự, 2022). FinQA bao gồm các cặp trả lời câu hỏi được chú thích bởi các chuyên gia và báo cáo thu nhập tương ứng của họ (bao gồm các tài liệu và bảng không có cấu trúc) từ các công ty S&P 500. ConvFinQA là một bản mở rộng của FinQA có các cuộc trò chuyện với câu hỏi nhiều lượt và trả lời các báo cáo thu nhập.
Tóm tắt văn bản. Tóm tắt văn bản nhằm mục đích cô đọng các văn bản tài chính dài không có cấu trúc thành các bản tóm tắt ngắn nắm bắt thông tin quan trọng và duy trì tính nhất quán thực tế với các văn bản dài ban đầu. Chúng tôi sử dụng hai bộ dữ liệu: ECTSum Mukherjee et al. (2022) để tóm tắt chiết xuất và EDTSum Zhou et al. (2021) để tóm tắt trừu tượng. ECTSum bao gồm 2.425 bản ghi cuộc gọi thu nhập dài (ECT) và tóm tắt gạch đầu dòng tương ứng được viết bởi các chuyên gia miền. EDTSum bao gồm các bài báo tài chính và tiêu đề tương ứng dưới dạng tóm tắt.
Dự đoán chuyển động cổ phiếu. Là một trong những nhiệm vụ tài chính cơ bản, dự đoán chuyển động cổ phiếu có giá trị tiềm năng lớn trong các ứng dụng thực tế như chiến lược đầu tư. Theo công việc trước đó (Soun et al., 2022), chúng tôi đóng khung nhiệm vụ như một vấn đề phân loại nhị phân, đó là dự đoán biến động giá cổ phiếu nhị phân dựa trên giá cổ phiếu và tweet lịch sử. Nếu biến động giá cao hơn 0,55%, nó sẽ được gán cho các mẫu dương (1) hoặc các mẫu âm (-1) nếu nó thấp hơn -0.5%. Chúng tôi áp dụng ba bộ dữ liệu được sử dụng phổ biến: BigData22 (Soun và cộng sự, 2022), ACL18 (Xu và Cohen, 2018) và CIKM18 (Wu và cộng sự, 2018).



Hình 1: Tổng quan về điều chỉnh lệnh đa nhiệm và đa phương thức của FinMA cho các nhiệm vụ tài chính đa dạng.

Chấm điểm tín dụng. Chấm điểm tín dụng là một nhiệm vụ quan trọng trong các dịch vụ tài chính, nhằm phân loại người tiêu dùng được mô tả bởi một tập hợp các thuộc tính là rủi ro tín dụng tốt hoặc xấu. Chúng tôi sử dụng hai bộ dữ liệu: Dữ liệu tín dụng Đức (Đức) Hofmann (1994) và Dữ liệu tín dụng Úc (Úc) Quinlan (1987). Tiếng Đức chứa 1000 trường hợp khách hàng, mỗi trường hợp được thể hiện bằng 20 thuộc tính, bao gồm trạng thái của tài khoản séc hiện có, lịch sử tín dụng, v.v. và nhãn tương ứng là rủi ro tín dụng tốt hoặc xấu. Australian chứa 690 trường hợp với 14 thuộc tính và nhãn rủi ro tín dụng tương ứng. Nhiệm vụ là dự đoán xem chúng là rủi ro tín dụng tốt hay xấu dựa trên các thuộc tính này.

3.2 Hướng dẫn xây dựng

Dựa trên các bộ dữ liệu thô, chúng tôi tiếp tục xây dựng các bộ dữ liệu hướng dẫn tài chính của mình, có số liệu thống kê được thể hiện trong Bảng 2. Chúng tôi yêu cầu các chuyên gia lĩnh vực viết 10 hướng dẫn đa dạng cho tất cả các tập dữ liệu ngoại trừ ConvFinQA, nơi chúng tôi chỉ sử dụng một lệnh. Vì ConvFinQA là một bộ dữ liệu trả lời câu hỏi đàm thoại nhiều lượt, có các câu hỏi đa dạng dưới dạng hướng dẫn trong tự nhiên. Đối với BigData22, ACL18, CIKM18, chúng tôi sử dụng cùng một tập lệnh, vì chúng có cùng kiểu dữ liệu đầu vào và công thức tác vụ[10]. Dựa trên những lời nhắc này, chúng tôi chuyển đổi các bộ dữ liệu thô từ các nhiệm vụ này thành các mẫu điều chỉnh hướng dẫn, bằng cách thu thập các hướng dẫn do con người thiết kế và nhập văn bản cùng với phản hồi của từng tập dữ liệu. Đối với bộ dữ liệu FPB, FiQA-SA, Headlines, NER, FiNER-ORD, ECTSum, EDTSum, Đức, Uc, FOMC, BigData22, ACL18 và CIKM18, chúng tôi xây dựng các mẫu điều chỉnh lệnh với mẫu sau:

Hướng dẫn: [nhắc nhiệm vụ] Văn bản: [văn bản đầu vào] Phản hồi: [đầu ra]

[prompt nhiệm vụ] là lời nhắc được thiết kế cho từng dữ liệu, [văn bản đầu vào] là dữ liệu tài chính đầu vào từ mỗi dữ liệu, ví dụ: giá lịch sử và tweet cho bộ dữ liệu dự đoán biến động chứng khoán, [đầu ra] là đầu ra tương ứng cho văn bản đầu vào, ví dụ: nhãn cảm xúc của văn bản đầu vào từ ["Tích cực", "Tiêu cực", "Trung lập"] trong bộ dữ liệu FiQA-SA. Đối với FPB, FiQA-SA và NER, chúng tôi sử dụng tất cả 10 lệnh cho mỗi mẫu, trong khi chúng tôi lấy mẫu ngẫu nhiên một lệnh cho mỗi mẫu trong các tập dữ liệu còn lại. Đối với FinQA và ConvFinQA, chúng tôi sử dụng mẫu sau:

Hướng dẫn: [lời nhắc nhiệm vụ] Ngữ cảnh: [ngữ cảnh đầu vào] Câu hỏi: [câu hỏi đầu vào] Trả lời: [an-swer]

[Ngữ cảnh đầu vào] là thông tin ngữ cảnh đầu vào cho mỗi mẫu dữ liệu. Ví dụ: inputcontext có thể được điền bằng văn bản và bảng từ các tệp điền cho FinQA. ConvFinQA có nhiều cuộc trò chuyện với các câu hỏi và câu trả lời. Chúng tôi chuyển đổi mỗi lượt của cuộc trò chuyện cho mỗi mẫu dữ liệu thành một hướng dẫn thông qua mẫu, mẫu này sẽ thêm các câu hỏi trước đó và trả lời trong [ngữ cảnh đầu vào].

4. FinMA: Mô hình ngôn ngữ lớn tài chính

Chúng tôi tiếp tục xây dựng FinMA bằng cách tinh chỉnh LLaMA (Touvron và cộng sự, 2023) với FIT. Chúng tôi đào tạo bốn mô hình: FinMA-7B và FinMA-30B bằng cách tinh chỉnh điểm kiểm tra LLaMA 7B và 30B với dữ liệu điều chỉnh lệnh bao gồm các tác vụ NLP, FinMA-7B-trade bằng cách tinh chỉnh điểm kiểm tra LLaMA 7B với dữ liệu điều chỉnh lệnh bao gồm các nhiệm vụ dự báo và FinMA-7B-full bằng cách tinh chỉnh LLaMA 7B với dữ liệu điều chỉnh toàn lệnh. Chúng tôi tinh chỉnh LLaMA-7B và LLaMA-7B-trade với 15 kỷ nguyên và LLaAM-7B-full với 3 kỷ nguyên dựa trên trình tối ưu hóa AdamW (Loshchilov và Hutter, 2017). Kích thước lô được đặt thành 32, tốc độ học ban đầu là 8e-6 và phân rã trọng lượng là 1e-5. Chúng tội cũng thiết lập các bước khởi động lên 5% tổng số bước tập luyện. Độ dài tối đa của văn bản đầu vào là 2048. FinMA-7B được tinh chỉnh 8 GPU A100 40GB. Đối với mô hình FinMA-30B, chúng tôi tinh chỉnh LLaMA-30B với 20 kỷ nguyên, cũng dựa trên trình tối ưu hóa AdamW. Kích thước lộ được đặt thành 24, tỷ lệ học ban đầu là 8e-6, giảm trọng lượng là 1e-5 và các bước khởi động đến 5% của tất cả các bước đào tạo. Độ dài tối đa của văn bản đầu vào là 2048. Khác với FinMA-7B, nó chỉ có thể được phân phối tinh chỉnh trên 128 GPU A100 40GB.

5 FLARE: Bộ Tiêu Chuẩn Đánh Giá Hiểu Biết và Dự Đoán Ngôn Ngữ Tài Chính

Dựa trên FIT, chúng tôi thiết kế điểm đánh giá dự đoán và hiểu ngôn ngữ tự nhiên tài chính (FLARE). Chúng tôi chọn ngẫu nhiên các bộ xác thực từ FIT để chọn điểm kiểm tra mô hình tốt nhất và các bộ kiểm tra để đánh giá. So với tiêu chuẩn FLUE hiện có (Sanh et al., 2022), FLARE bao gồm các nhiệm vụ dự đoán tài chính ngoài các nhiệm vụ NLP[11]. Chúng tôi tin rằng điều

[10] Các ví dụ hướng dẫn được trình bày trong Phụ lục

[11] Sau BloombergGPT, chúng tôi không bao gồm nhiệm vụ phát hiện ranh giới cấu trúc có trong FLUE vì nó khó được chuyển đổi thành nhiệm vụ theo hướng dẫn.

quan trọng là phải bao gồm các nhiệm vụ dự đoán tài chính như dự đoán biến động cổ phiếu, để đánh giá toàn diện hiệu suất của LLM trên các ứng dụng thực tế của lĩnh vực tài chính. Chúng tôi hiển thị số liệu thống kê dữ liệu của xác nhận và bộ kiểm tra cho từng tập dữ liệu trong Bảng 3. Theo các phương pháp trước đó Li et al. (2023); Shah và cộng sự (2022),

Bảng 3: Chi tiết về bộ dữ liệu đánh giá của chúng tôi. Để so sánh hiệu suất với BloombergGPT mà dữ liệu thử nghiệm không được công bố công khai, chúng tôi giữ nguyên số lượng và phân phối dữ liệu của bộ dữ liệu thử nghiệm của chúng tôi với BloombergGPT. Để đánh giá thêm khả năng xuất hiện và khái quát hóa của LLM, chúng tôi chỉ áp dụng dữ liệu thử nghiệm của FOMC, FINER-ORD, ECTSum, EDTSum, German và Australian để đánh giá trên FLARE. Đối với ConvFinQA, chúng tôi lấy mỗi lượt của cuộc trò chuyện làm hướng dẫn, số lượng của nó sẽ khác với số lượng cuộc trò chuyện.

Data	Task	Valid	Test	Evaluation
FPB (Malo et al., 2014)	sentiment analysis	775	970	F1, Accuracy
FiQA-SA (Maia et al., 2018)	sentiment analysis	188	235	F1
Headlines (Sinha and Khandait, 2021)	news headline classification	1,141	2,283	Avg F1
NER (Alvarado et al., 2015)	named entity recognition	103	980	Entity F1
FiNER-ORD (Shah et al., 2023b)	named entity recognition	-	1080	Entity F1
FinQA (Chen et al., 2021)	question answering	883	1,147	EM Accuracy
ConvFinQA (Chen et al., 2022)	question answering	2,210	1,490	EM Accuracy
BigData22 (Soun et al., 2022)	stock movement prediction	798	1,470	Accuracy, MCC
ACL18 (Xu and Cohen, 2018)	stock movement prediction	2,560	3,720	Accuracy, MCC
CIKM18 (Wu et al., 2018)	stock movement prediction	431	1,140	Accuracy, MCC
ECTSum Mukherjee et al. (2022)	text summarization	-	495	ROUGE, BERTScore, BARTScore
EDTSum Zhou et al. (2021)	text summarization	-	2000	ROUGE, BERTScore, BARTScore
German Hofmann (1994)	credit scoring	-	1000	F1, MCC
Australian Quinlan (1987)	credit scoring	-	690	F1, MCC
FOMC Shah et al. (2023a)	hawkish-dovish classification	-	496	F1, Accuracy

chúng tôi đánh giá hiệu suất của nhiệm vụ phân loại cảm xúc trên bộ dữ liệu FPB và FiQA-SA, với độ chính xác (ACC) và Điểm F1 có trọng số (F1). Hiệu suất của nhiệm vụ phân loại tiêu đề tin tức được đánh giá bằng điểm trung bình có trọng số của điểm F1 trên tất cả chín hạng mục (F1 trung bình). Đối với hiệu suất của nhiệm vụ NER, chúng tôi đánh giá với điểm F1 cấp thực thể (Thực thể F1). Hiệu suất của nhiệm vụ trả lời câu hỏi được đánh giá với độ chính xác khớp chính xác (EM Acc). Đối với các nhiệm vụ tóm tắt như ECTSum và EDTSum, chúng tôi đánh giá mức độ liên quan và tính thực tế của các bản tóm tắt được tạo ra với sự thật cơ bản bằng cách sử dụng các số liệu như điểm ROUGE Lin (2004), BERTScore Zhang và cộng sự (2019) và BARTScore Yuan và cộng sự (2021). Đối với nhiệm vụ dự đoán tài chính, theo các phương pháp trước đó (Xu và Cohen, 2018; Xie và cộng sự, 2023), chúng tôi đánh giá hiệu suất với độ chính xác (ACC) và hệ số tương quan Matthews (MCC) để dự đoán biến động giá cổ phiếu và điểm F1 với MCC để chấm điểm tín dụng. Mặc dù Macro-F1 công bằng hơn đối với bộ dữ liệu không cân bằng, nhưng trong bài báo này, chúng tôi áp dụng các số liệu tương tự theo các phương pháp trước đó để so sánh công bằng.

6 Thí nghiệm trên FLARE

FIT và FLARE được đề xuất cho phép đào tạo, lựa chọn mô hình và đánh giá hiệu suất của LLM về sự hiểu biết và dự đoán tài chính. Trong phần này, chúng tôi điều tra mức độ mạnh mẽ của FinMA được tinh chỉnh FIT và các LLM khác trên FLARE. Chúng tôi so sánh FinMA với các LLM sau: 1) BloombergGPT (Wu và cộng sự, 2023). Mô hình ngôn ngữ lớn duy nhất với 50B tham số được đào tạo trước với các văn bản tài chính. 2) GPT-4 (OpenAI, 2023). Một hướng dẫn mạnh mẽ theo mô hình ngôn ngữ lớn với các tham số khoảng 1T do OpenAI đề xuất. 3) ChatGPT. Một hướng dẫn theo mô hình ngôn ngữ lớn với 175B tham số từ OpenAI. 4) Vicuna-7B (Zhang và cộng sự, 2022). Một hướng dẫn theo mô hình ngôn ngữ lớn bằng cách tinh chỉnh LLaMA-7B

Theo các phương pháp trước đó (Wu và cộng sự, 2023; Li và cộng sự, 2023), chúng tôi báo cáo hiệu suất 20 lần bắn của BloombergGPT và hiệu suất 5 lần bắn của các phương pháp cơ sở khác trên bộ dữ liệu FIN. Chúng tôi báo cáo hiệu suất 5 lần bắn của BloombergGPT trên FPB và FiQA-SA. Chúng tôi báo cáo hiệu suất 5 cảnh quay của tất cả các đường cơ sở trên tập dữ liệu Tiêu đề. Đối với các kết quả còn lại, chúng tôi báo cáo hiệu suất không bắn. Kết quả của một số

đường cơ sở dựa trên đánh giá của con người, vì LLM không tinh chỉnh sẽ không tạo ra câu trả lời được xác định trước trong hướng dẫn nhất định. Tất cả các kết quả của FinMA được tiến hành trên zero-shot và có thể được đánh giá tự động.

Bảng 4: Hiệu suất không bắn và ít bắn của các LLM khác nhau trên điểm chuẩn FLARE. Kết quả của BloombergGPT, ChatGPT và GPT4 trên FPB, FiQASA, Headlines, NER, FinQA và ConvFinQA được tham khảo từ bài báo (Li và cộng sự, 2023). Kết quả của BloombergGPT được tham khảo từ bài báo gốc Wu et al. (2023). Bộ dữ liệu thử nghiệm được xây dựng để có cùng phân phối dữ liệu với BloombergGPT và hiệu suất của FinMA được so sánh trực tiếp với BloombergGPT theo phương pháp trước đó (Li và cộng sự, 2023). Tất cả các kết quả thông qua đánh giá của chúng tôi là trung bình của ba lần chạy. "-" đại diện cho kết quả hiện không thể mang lại do kích thước mô hình hoặc tính khả dụng và "*" đại diện cho kết quả từ bài báo trước đó.

Dataset	Metrics	Chat GPT	GPT 4	Bloom berg GPT	Vicuna 7B	FinMA 7B	FinMA 7B-trade	FinMA 7B-full	FinMA 30B
FPB	F1	0.78*	0.78*	0.51*	0.29	**0.94**	0.03	**0.94**	0.88
	Acc	0.78*	0.76*	-	0.26	**0.94**	0.12	**0.94**	0.87
FiQA-SA	F1	0.60	0.80	0.75*	0.32	0.85	0.16	0.82	**0.87**
Headlines	AvgF1	0.77*	0.86*	0.82*	0.60	**0.97**	0.28	**0.97**	**0.97**
NER	EntityF1	0.77*	**0.83***	0.61*	0.12	0.59	0.00	0.64	0.62
FINER-ORD	EntityF1	0.28	**0.77**	-	0.00	0.00	0.00	0.00	0.00
FinQA	EmAcc	0.58*	**0.63***	-	0.00	0.06	0.00	0.04	0.11
ConvFinQA	EmAcc	0.60*	**0.76***	0.43*	0.00	0.25	0.00	0.20	0.40
BigData22	Acc	0.53	**0.54**	-	0.44	0.45	0.45	0.51	0.47
	MCC	-0.025	0.03	-	-0.05	0.02	0.00	0.02	**0.04**
ACL18	Acc	0.50	**0.52**	-	0.50	0.49	0.49	0.51	0.49
	MCC	0.005	0.02	-	0.02	-0.01	**0.03**	0.03	0.00
CIKM18	Acc	0.55	**0.57**	-	0.44	0.43	0.43	0.50	0.43
	MCC	0.01	0.02	-	-0.03	-0.02	-0.003	**0.08**	-0.05
EDTSUM	Rouge-1	0.17	0.2	-	**0.22**	0.09	0.05	0.13	0.17
	Rouge-2	0.08	0.09	-	**0.10**	0.04	0.02	0.06	0.08
	Rouge-N	0.13	0.15	-	**0.17**	0.08	0.05	0.10	0.14
	BertScore	0.66	**0.67**	-	0.61	0.56	0.51	0.38	0.54
	BartScore	-3.64	**-3.62**	-	-4.13	-6.12	-6.91	-5.71	-5.24
ECTSUM	Rouge-1	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
	Rouge-2	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
	Rouge-N	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
	BertScore	0.00	0.00	-	0.00	0.00	0.00	0.00	0.00
	BartScore	-5.18	-5.18	-	-5.18	-5.18	-5.18	-5.18	-5.18
German	F1	0.20	**0.55**	-	0.52	0.17	0.52	0.17	0.53
	MCC	-0.10	-0.02	-	**0.00**	0.00	-0.07	0.00	-0.07
Australian	F1	0.41	**0.74**	-	0.26	0.41	0.26	0.41	0.46
	MCC	0.00	**0.47**	-	0.00	0.00	0.00	0.00	-0.01
FOMC	F1	0.64	**0.71**	-	0.19	0.49	0.10	0.49	0.43
	Acc	0.6	**0.69**	-	0.28	0.47	0.25	0.46	0.53

6.1 Kết quả

Hiệu suất tổng thể. Đối với các nhiệm vụ NLP tài chính, như thể hiện trong Bảng 4[12], mô hình FinMA được tinh chỉnh của chúng tôi vượt trội hơn đáng kể các LLM khác trên bộ dữ liệu FPB, FiQA-SA và Headlines, cho thấy tầm quan trọng của việc điều chỉnh hướng dẫn cụ thể trong việc cải thiện hiệu suất của LLM trong miền cụ thể. Ví dụ: FinMA-30B vượt trội hơn GPT-4 10% điểm F1 và BloombergGPT 37% điểm F1 trên tập dữ liệu FPB. Trên bộ dữ liệu NER, FinMA-7B cũng vượt trội hơn BloombergGPT và các LLM khác, đồng thời đạt được kết quả cạnh tranh so với ChatGPT và GPT-4. Đối với FinQA và ConvFinQA yêu cầu suy luận số phức tạp, có một khoảng cách lớn giữa hiệu suất của GPT và FinMA. Như đã báo cáo trong các nghiên cứu hiện có (Touvron và cộng sự, 2023; Lewkowycz và cộng sự, 2022), LLaMA không

[12] Để biết hiệu suất của các LLM khác, vui lòng xem Phụ lục.

bao gồm bộ dữ liệu toán học để đào tạo trước, dẫn đến hiệu suất kém trên các bộ dữ liệu điểm chuẩn toán học như GSM8K (Cobbe và cộng sự, 2021). Kết quả của chúng tôi cũng phù hợp với các nghiên cứu trước đây cho thấy LLaMA với các tham số lớn hơn thể hiện hiệu suất tốt hơn trên các bộ dữ liệu điểm chuẩn toán học. Hiệu suất của FinMA-30B tốt hơn đáng kể so với FinMA-7B trên FinQA và ConvFinQA. Phát hiện này chỉ ra tầm quan trọng của lý luận số đối với việc trả lời câu hỏi tài chính, có thể là hướng đi tiềm năng để thúc đẩy LLM trong lĩnh vực tài chính. Bất chấp những điểm mạnh được thể hiện bởi các phương pháp của chúng tôi đối với các nhiệm vụ đã biết, chúng hoạt động kém hơn so với GPT-4 đối với các tác vụ vô hình như FINER-ORD, EDTSUM, ECTSUM và FOMC. Khoảng cách hiệu suất này cho thấy rằng cần có một tập hợp các nhiệm vụ cụ thể theo lĩnh vực đa dạng hơn để tinh chỉnh hiệu quả. Cụ thể, trong FINER-ORD và ECTSUM, chúng tôi sử dụng một thiết kế nhắc nhở phức tạp yêu cầu mô hình tạo trình tự nhãn trực tiếp. Kết quả chỉ ra rằng các mô hình tinh chỉnh của chúng tôi liên tục không tạo ra đầu ra ở các định dạng mong muốn. Trong khi các mô hình như ChatGPT và GPT-4 chứng minh một số khả năng gắn nhãn mã thông báo trong FINER-ORD, chúng cũng gặp khó khăn trong việc tạo nhãn câu cho ECTSUM, đặc biệt là khi đối mặt với thông tin ngữ cảnh dài hơn. Trái ngược với các tiêu chuẩn tài chính hiện có, FLARE cung cấp một bộ toàn diện hơn về cả nhiệm vụ tạo và phân loại, do đó cung cấp đánh giá đầy đủ hơn về khả năng của các mô hình ngôn ngữ lớn trong lĩnh vực NLP tài chính.

Đối với các nhiệm vụ dự đoán tài chính, tất cả các LLM bao gồm FinMA, ChatGPT và GPT-4 đều gặp khó khăn trong dự đoán biến động cổ phiếu như các phương pháp trước đây[13]. Mặc dù FinMA-7B-trade đã được tinh chỉnh đặc biệt cho nhiệm vụ dự đoán biến động cổ phiếu, nhưng mức tăng hiệu suất quan sát được trong lĩnh vực này là tốt nhất. Sau khi tinh chỉnh cả NLP và nhiệm vụ dự đoán tài chính, FinMA-7B-full có thể đạt được hiệu suất tốt hơn đáng kể trên tập dữ liệu ACL18 so với ChatGPT và GPT-4. Tuy nhiên, nó vẫn gần như không có MCC trên hai bộ dữ liệu còn lại như ChatGPT và GPT-4. Tương tự, các phương pháp được kiểm tra tất cả, ngoại trừ GPT-4, thể hiện các giá trị Matthews CorrelationCoefficient (MCC) bằng không hoặc âm trong nhiệm vụ chấm điểm tín dụng. Điều này làm nổi bật những hạn chế của họ trong việc dự báo chính xác rủi ro vỡ nợ cá nhân. Mặc dù GPT-4 cho thấy sự cải thiện rõ rệt về hiệu suất trên bộ dữ liệu của Úc, nhưng nó không đăng ký MCC dương trên bộ dữ liệu của Đức. Những hạn chế như vậy có thể là do những thách thức liên quan đến đầu vào dữ liệu dạng bảng và phân phối nhãn mất cân bằng cao vốn có của các nhiệm vụ chấm điểm tín dụng. Điều này cho thấy sự phức tạp và thách thức củacác nhiệm vụ dự đoán tài chính trong FLARE. So với các tiêu chuẩn tài chính hiện có tập trung vào các nhiệm vụ NLP, FLARE mang đến những cơ hội thú vị để cải thiện LLM dựa trên nền tảng của các nghiên cứu và ứng dụng học thuật tài chính. Nó cũng chứng minh tầm quan trọng của việc học nhiều nhiệm vụ trong lĩnh vực tài chính đối với LLM, có thể cung cấp kiến thức và kỹ năng lĩnh vực cần thiết để xử lý các ứng dụng phức tạp trong lĩnh vực này.

Phân tích thêm. Chúng tôi phân tích thêm ảnh hưởng của kích thước mô hình và dữ liệu điều chỉnh lệnh đối với hiệu suất của LLM trên các tác vụ khác nhau. FinMA-30B không có hiệu suất tốt hơn đáng kể so với FinMA-7B trên hầu hết các tác vụ NLP và nhiệm vụ dự đoán chuyển động cổ phiếu. Rõ ràng, chất lượng của các hướng dẫn hơn là kích thước mô hình là rất quan trọng đối với việc thực hiện các tác vụ này. Đối với các tác vụ trả lời câu hỏi phức tạp như ConvFinQA, như thể hiện trong Bảng 4, mô hình LLaMA lớn hơn thường có hiệu suất tốt hơn. Đặc biệt, Vicuna-7B dựa trên LLaMA-7B có hiệu suất kém nhất, điều này cũng phù hợp với những phát hiện trước đó (Cobbe và cộng sự, 2021) rằng LLaMA với các tham số lớn hơn thể hiện hiệu suất tốt hơn trên các bộ dữ liệu điểm chuẩn toán học. Ngược lại, đối với các tác vụ thế hệ như tóm tắt trừu tượng (EDTSUM), Vicuna-7B có hiệu suất tốt nhất trong khi các mô hình tinh chỉnh cho thấy hiệu suất giảm trên hầu hết các chỉ số. Điều này có thể chỉ ra rằng việc tinh chỉnh chỉ với các nhiệm vụ phân loại có thể dẫn đến hiệu suất phân loại tốt hơnnhưng cũng ảnh hưởng đến khả năng tạo. Đối với nhiệm vụ dự đoán tài chính và nhiệm vụ NLP không được đưa vào bộ dữ liệu hướng dẫn tinh chỉnh của FinMA-7B và FinMA-30B, tức là FOMC, FINER-ORD, Australian và ACL18, các mô hình của chúng tôi có những cải tiến hạn chế. Mặc dù nó thể hiện một số mức độ khả năng nổi lên, nhưng khoảng cách hiệu suất so với GPT-4 cho thấy nhu cầu tối ưu hóa hơn nữa. Tuy nhiên, FinMA-7B đầy đủ được tinh chỉnh với cả bộ dữ liệu NLP và dự đoán, đã cho thấy hiệu suất tốt hơn đáng kể trên các bộ dữ liệu dự đoán tài chính và hiệu suất tương đương trên các tác vụ NLP với FinMA-7B và GPT-4. Điều này cho thấy tiềm năng của LLM sẽ được điều chỉnh hơn nữa và áp dụng trực tiếp vào các nhiệm vụ dự đoán tài chính thông qua đào tạo trước và tinh chỉnh trên các bộ dữ liệu miền đa dạng.

[13] Để biết hiệu suất của các phương pháp truyền thống, vui lòng xem Phụ lục.

7 Hạn chế

Bất chấp những đóng góp tích cực của nghiên cứu này, chúng tôi nhận ra những hạn chế sau: **1) Mô hình và Hạn chế đào tạo**: Chúng tôi chỉ trình bày các mô hình FinMA lên đến 30B. Do các ràng buộc về tính toán, FinMA-30B chưa được tinh chỉnh trên tập dữ liệu đầy đủ. **2) Hiệu suất nhiệm vụ phức tạp**: FinMA, do hạn chế của mô hình đường trục LLaMA, phải vật lộn với các nhiệm vụ đòi hỏi suy luận lượng tử, chẳng hạn như trả lời câu hỏi tài chính và nhiệm vụ dự đoán tài chính khó khăn. **3) Hạn chế tài nguyên và khả năng khái quát hóa**: Sự phát triển của FinMA, FIT và FLARE bị ảnh hưởng bởi các nguồn lực sẵn có và các hướng dẫn thủ công, có khả năng ảnh hưởng đến sự đa dạng và khả năng khái quát hóa mô hình. Kích thước đầu vào tối đa của FinMA cũng bị giới hạn bởi các văn bản đầu vào tối đa có thể được xử lý bởi mô hình đường trục LLaMA. **4) Tác động tiêu cực tiềm ẩn**: Mặc dù nghiên cứu của chúng tôi chủ yếu tập trung vào các khía cạnh tích cực và tiến bộ của các mô hình hiểu ngôn ngữ tài chính, nhưng điều quan trọng là phải thừa nhận các tác động tiêu cực tiềm ẩn liên quan đến việc sử dụng chúng, chẳng hạn như sự lan truyền thông tin sai lệch về tài chính hoặc ảnh hưởng phi đạo đức của thị trường. Chúng tôi khuyến bạn nên sử dụng phương pháp của chúng tôi chỉ cho nghiên cứu học thuật.[14]

8 Kết luận

Trong công trình này, chúng tôi đã trình bày PIXIU, bao gồm mô hình ngôn ngữ lớn tài chính nguồn mở đầu tiên FinMA, bộ dữ liệu điều chỉnh lệnh FIT và điểm chuẩn đánh giá FLARE. Thông qua đánh giá rộng rãi, chúng tôi đã chứng minh hiệu quả của FinMA trong các nhiệm vụ tài chính khác nhau, cho thấy tiềm năng của việc điều chỉnh hướng dẫn theo lĩnh vực cụ thể của các mô hình ngôn ngữ lớn trong lĩnh vực tài chính. Tuy nhiên, những thách thức như cải thiện hiệu suất trong các nhiệm vụ phức tạp và giải quyết các hạn chế về nguồn lực vẫn còn. Đóng góp nguồn mở của chúng tôi nhằm tạo điều kiện cho nghiên cứu và đổi mới hơn nữa trong việc hiểu, dự đoán và LLM ngôn ngữ tài chính, hướng tới các LLM hữu ích và an toàn hơn trong lĩnh vực tài chính.

Lời cảm ơn

Dự án PIXIU chủ yếu được hỗ trợ bởi ChanceFocus AMC và được hỗ trợ một phần bởi Chương trình Nghiên cứu và Phát triển Quốc gia Trung Quốc (Số 2021ZD0113304) và Chương trình Tổng hợp của Quỹ Khoa học Tự nhiên Trung Quốc (NSFC) (Tài trợ số 62072346), Tài trợ bổ sung được cung cấp bởi Dự án Nghiên cứu và Phát triển trọng điểm của tỉnh Hồ Bắc (Tài trợ số 2021BBA099, Số 2021BBA029) và bởi Phòng thí nghiệm Liên hợp và Phòng thí nghiệm về Công nghệ Tín dụng.

Tham khảo

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology* Association Workshop 2015. 84–90.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv*:1908.10063 (2019).

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

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3697–3711.

[14] Để biết tuyên bố đạo đức và pháp lý chi tiết liên quan đến công việc này, vui lòng xem Phụ lục.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv*:2210.03849 (2022).

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https: //lmsys.org/blog/2023-03-30-vicuna/

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pretraining text encoders as discriminators rather than generators. *arXiv preprint arXiv*:2003.10555 (2020).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. 4171–4186.

Weiguang Han, Boyi Zhang, Qianqian Xie, Min Peng, Yanzhao Lai, and Jimin Huang. 2023. Select and Trade: Towards Unified Pair Trading with Hierarchical Reinforcement Learning. *arXiv preprint arXiv*:2301.10724 (2023).

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv*:2206.14858 (2022).

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? An Examination on Several Typical Tasks. *arXiv preprint arXiv*:2305.05862 (2023).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. FinGPT: Democratizing Internet-scale Data for Financial Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv*:2301.13688 (2023).

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv*:2304.07619 (2023).

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv*:1711.05101 (2017).

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark. *arXiv preprint arXiv*:2302.09432 (2023)

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*. 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al.

2022. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 10893–10906.

Jingwei Ni, Zhijing Jin, Qian Wang, Mrinmaya Sachan, and Markus Leippold. 2023. When Does Aggregating Multiple Skills with Multi-Task Learning Work? A Case Study in Financial NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 7465–7488. https://doi.org/10.18653/v1/2023.acl-long.412

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

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* (2022), 27730–27744.

Ross Quinlan. 1987. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR 2022-Tenth International Conference on Learning Representations.*

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman ´ Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv*:2211.05100 (2022).

Agam Shah and Sudheer Chava. 2023. Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks. *ArXiv* abs/2305.16633 (2023). https://api.semanticscholar. org/CorpusID:258941519

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv*:2302.11157 (2023).

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. WHEN FLUE MEETS FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. *arXiv preprint arXiv:2211.00083* (2022).

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In Advances in Information and Communication: Proceedings of the 2021 *Future of Information and Communication Conference (FICC), Volume 2. Springer,* 589–601.

Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, and U Kang. 2022. Accurate Stock Movement Prediction with Self-supervised Learning from Sparse Noisy Tweets. In 202*2 IEEE International Conference on Big Data (Big Data)*. IEEE, 1691–1700.

A Hướng Dẫn

Bảng 5 trình bày tất cả các lời nhắc cho từng tập dữ liệu trong tập dữ liệu hướng dẫn điểm chuẩn FLARE và FIT.

Bảng 5: Lời nhắc ví dụ cho mỗi tập dữ liệu. FiQA-SA có hai loại văn bản, bao gồm tiêu đề tin tức và tweet. Chúng tôi sẽ điền loại văn bản chi tiết vào {category} cho mỗi mẫu dữ liệu. Đối với dữ liệu dự đoán chuyển động như BigData22, chúng tôi sẽ điền vào {tid} và {point} với tên và thời gian chi tiết từ mỗi mẫu dữ liệu.

Data	Prompt
FPB	"Analyze the sentiment of this statement extracted from a financial news article. Provide your answer as either negative, positive or neutral. For instance, 'The company's stocks plummeted following the scandal.' would be classified as negative."
FiQA-SA	"What is the sentiment of the following financial {category}: Positive, Negative, or Neutral?"
Headlines	"Consider whether the headline mentions the price of gold. Is there a Price or Not in the gold commodity market indicated in the news headline? Please answer Yes or No."
NER	"In the sentences extracted from financial agreements in U.S. SEC filings, identify the named entities that represent a person ('PER'), an organization ('ORG'), or a location ('LOC'). The required answer format is: 'entity name, entity type'. For instance, in 'Elon Musk, CEO of SpaceX, announced the launch from Cape Canaveral.', the entities would be: 'Elon Musk, PER; SpaceX, ORG; Cape Canaveral, LOC'"
FiNER-ORD	"In the list of tokens, identify 'Person', 'Location', and 'Organisation' and label each accordingly. If the entity spans multiple tokens, use the prefix B-PER, B-LOC, or B-ORG for the first token, and I-PER, I-LOC, or I-ORG for the subsequent tokens of that entity. The beginning of each separate entity should always be labeled with a B-PER, B-LOC, or B-ORG prefix. If the token does not fit into any of the three named categories, or is not a named entity, label it as 'O'. Each line should contain one token and its corresponding label, separated by a colon. Do not combine tokens on your own. The format for each line should be: 'token:label'. Text: And all because you failed to prepare ! Answer:"
FinQA	"Given the financial data and expert analysis, please answer this question:"
ConvFinQA	"In the context of this series of interconnected finance-related queries and the additional information provided by the pretext, table data, and post text from a company's financial filings, please provide a response to the final question. This may require extracting information from the context and performing mathematical calculations. Please take into account the information provided in the preceding questions and their answers when formulating your response:"
BigData22	" Contemplate the data and tweets to guess whether the closing price of {tid} will surge or decline at {point}. Please declare with either Rise or Fall."
ECTSum	"Given the following article, please produce a list of 0 and 1, each separated by ' ' to indicate which sentences should be included in the final summary. The article's sentences have been split by ' '. Please mark each sentence with 1 if it should be included in the summary and 0 if it should not."
EDTSum	"You are given a text that consists of multiple sentences. Your task is to perform abstractive summarization on this text. Use your understanding of the content to express the main ideas and crucial details in a shorter, coherent, and natural sounding text."
German	"Assess the creditworthiness of a customer using the following table attributes for financial status. Respond with either 'good' or 'bad'. And the table attributes including 13 categorical attributes and 7 numerical attributes are as follows:"
FOMC	"Examine the excerpt from a central bank's release below. Classify it as HAWKISH if it advocates for a tightening of monetary policy, DOVISH if it suggests an easing of monetary policy, or NEUTRAL if the stance is unbiased. Your response should return only HAWKISH, DOVISH, or NEUTRAL."

B Các phương pháp truyền thống để dự báo chuyển động cổ phiếu

Trong bối cảnh dự đoán chuyển động cổ phiếu, các mô hình truyền thống, như được tóm tắt trong Bảng 6, đã được sử dụng từ lâu nhưng phải đối mặt với những thách thức đáng kể trong việc đạt được mức độ chính xác cao nhất quán. Điều này nhấn mạnh sự khó khăn cố hữu của nhiệm vụ trước mắt. Ngược lại, Mô hình ngôn ngữ lớn (LLM) giới thiệu một mức độ thích ứng bằng cách học từ nhiều nhiệm vụ, mặc dù chúng cũng có những hạn chế, chẳng hạn như hiểu số và suy luận. Do đó, độ khó của nhiệm vụ là một thách thức chung đối với cả mô hình truyền thống và LLM.

Bảng 6: Hiệu suất dự đoán chuyển động của các mô hình không phải LLM so với FinMA, được đo bằng độ chính xác (ACC) và hệ số tương quan Matthews (MCC). Tốt nhất trong số các mô hình không phải LLM là in đỏ và tốt nhất là in đậm.

Method	BIGDATA22		ACL18		CIKM18	
	ACC	MCC	ACC	MCC	ACC	MCC
Logistic regression (LR)	0.53	0.02	0.52	0.04	0.53	-0.04
Random forest (RF)	0.47	-0.11	0.52	0.03	0.54	0.01
LSTM	0.51	0.01	0.53	0.06	0.53	0.02
Attention LSTM (ALSTM)	0.49	-0.03	0.52	0.04	0.53	-0.01
Adv-ALSTM	0.50	0.01	0.53	0.07	0.54	0.02
DTML	0.52	0.07	0.58	0.18	0.54	-0.00
XGBoost	0.52	-0.04	0.49	-0.02	**0.58**	0.07
XGBRefressor	0.46	-0.13	0.50	-0.01	0.53	-0.03
ALSTM-W	0.48	-0.01	0.53	0.08	0.54	0.03
ALSTM-D	0.49	0.01	0.53	0.07	0.50	-0.04
StockNet	0.53	-0.00	0.54	-0.03	0.52	-0.02
SLOT	**0.55**	**0.10**	**0.59**	**0.21**	0.56	**0.09**
FinMA-7B	0.48	0.04	0.50	0.00	0.56	-0.02
FinMA-30B	0.47	0.04	0.49	0.00	0.43	-0.05
FinMA-7B-full	0.49	0.01	0.56	0.10	0.53	-0.03

C Hiệu suất của các mô hình ngôn ngữ lớn dựa trên BERT

Bảng 7 trình bày hiệu suất được báo cáo trước đây của các mô hình ngôn ngữ được đào tạo trước (PLM) bao gồm FinBERT và FLANG-BERT trên ba nhiệm vụ được chọn theo tiêu chuẩn FLARE. Mặc dù cả hai mô hình đều thể hiện kết quả ấn tượng trong các chỉ số hiệu suất cụ thể, nhưng điều quan trọng cần lưu ý là chúng đã được đào tạo trước với các văn bản tài chính quy mô lớn và sử dụng tiêu đề dành riêng cho các nhiệm vụ khác nhau, đó là lý do chính cho hiệu suất tốt hơn của chúng. Tuy nhiên, so với LLM, các PLM này không thể thích ứng với các tác vụ không nhìn thấy nếu không có tinh chỉnh có giám sát, do đó có nút thắt cổ chai trong việc thích ứng với môi trường đa nhiệm và học tập không bắn. Đối với LLM, kết quả của chúng tôi cho thấy hiệu suất của chúng vẫn kém hơn các PLM được điều chỉnh tinh tế này trong một số tác vụ, nhưng mang lại tính linh hoạt, khả năng thích ứng và khả năng không bắn cao hơn. Họ có lợi thế là học trực tiếp từ lời nhắc và linh hoạt hơn trong việc xử lý một loạt các nhiệm vụ, ngay cả khi không cần dữ liệu đào tạo được gắn nhãn. Hiệu suất kém của LLM so với PLM cũng làm nổi bật sự cần thiết phải đào tạo trước lĩnh vực cụ thể của LLM trong tương lai, để cải thiện hơn nữa hiệu suất của họ trong lĩnh vực cụ thể. Những quan sát này phù hợp với các nghiên cứu gần đây, chẳng hạn như Shah và Chava (2023), cũng quan sát thấy rằng các LLM zero-shot như ChatGPT mang lại hiệu suất đáng nể trong các nhiệm vụ tài chính mà không cần dữ liệu được gắn nhãn. Công trình gần đây của Ni et al. (2023), cũng nêu bật những thách thức của việc PLM tinh chỉnh thích ứng với nhiều nhiệm vụ vô hình và không nhìn thấy.

Bảng 7: Kết quả của các mô hình ngôn ngữ lớn (LLM) mã hóa-giải mã dựa trên BERT trên nhiều tác vụ được báo cáo trong các bài báo trước.

Method	FPB	Headline	NER
	Accuracy	AvgF1	F1
FinBERT	0.872	0.968	0.8
FLANG-BERT	0.912	0.972	0.83

D Hiệu suất của các mô hình ngôn ngữ lớn trước đây

Bảng 8 cung cấp so sánh chi tiết về các chỉ số hiệu suất không bắn và ít lần bắn cho các Mô hình ngôn ngữ lớn (LLM) khác nhau — GPT NeoX 20B, OPT 66B và BLOOM — trên nhiều bộ dữ liệu. Các số liệu này, thu được từ một bài báo trước đó, đóng vai trò là cơ sở có giá trị để hiểu khả năng của các mô hình này trong cả kịch bản không bắn và ít bắn theo điểm chuẩn FLARE.

Bảng 8: Hiệu suất không bắn và ít bắn của các LLM khác nhau trên điểm chuẩn FLARE. Kết quả được tham khảo từ bài báo trước.

Dataset	**Metrics**	**GPT NeoX 20B**	**OPT 66B**	**BLOOM**
FPB	F1	0.45*	0.49*	0.50*
	Acc	0.38	-	-
FiQA-SA	F1	0.51*	0.52*	0.53*
Headlines	AvgF1	0.73*	0.79*	0.77*
NER	EntityF1	0.61*	0.57*	0.56*
FinQA	EmAcc	0.00	-	-
ConvFinQA	EmAcc	0.28*	0.30*	0.36*
BigData22	Acc	0.41	-	-
	MCC	0.08	-	-
ACL18	Acc	0.35	-	-
	MCC	0.00	-	-
CIKM18	Acc	0.25	-	-
	MCC	-0.12	-	-
EDTSUM	Rouge-1	0.02	-	-
	Rouge-2	0.01	-	-
	Rouge-N	0.02	-	-
	BertScore	0.48	-	-
	BartScore	-5.72	-	-
ECTSUM	Rouge-1	0.00	-	-
	Rouge-2	0.00	-	-
	Rouge-N	0.00	-	-
	BertScore	0.00	-	-
	BartScore	-5.18	-	-
German	F1	0.17	-	-
	MCC	0.02	-	-
Australian	F1	0.00	-	-
	MCC	0.00	-	-
FOMC	F1	0.37	-	-
	Acc	0.27	-	-