LSEG

Using GPT-4 with prompt engineering for financial industry tasks

LSEG ANALYTICS, MAY 2023

Contents

Section 1: Overview of GPT

Innovations and advancements in cutting-edge technologies in the field of Large Language Models (LLMs) are growing exponentially. With vast quantities of available data and increases in computing power, they have wide-ranging application potential in the financial industry.

LSEG is well placed to provide commentary in this space given our experience with LLMs in combination with the wide variety of financial data used to improve these models and our exploration of opportunities with Microsoft.

In this paper, we aim to demonstrate the role of prompt engineering in GPT models in improving performance for sentiment and theme classification using financial text data. The results were obtained using GPT-4, which was released in March 2023, and at the time of writing is new for all users. Our aim with the exploratory studies is to provide clear and concise communication for an improved understanding of GPT models in the financial industry.

Using GPT-4 for sentiment and theme classification, we found that GPT-4 outperforms the GPT-3 and GPT-3.5 models. Also, GPT-4 slightly outperformed other LLMs that are used as benchmarks, such as BART and FinBERT. Using prompt engineering for sentiment classification, GPT-4's performance was seen to further improve, indicating that prompt engineering is a valuable area for performance optimisation.

Introduction

The release of GPT-4 series marks an exciting time for those industries looking to apply Large Language Models (LLMs) (1; 2). The improved capability of LLMs has made them increasingly relevant for financial industry tasks.

LSEG leverages LLM capabilities in a variety of products: summarisation, entity recognition, topic detection and sentiment analysis are used in products including SentiMine (1,2), MarketPsych (3) and Transcripts Summarisation, and are being implemented into Starmine (4). SentiMine provides aspect-based sentiment analysis of earnings/ conference call transcripts for a wide range of financially significant themes. Such understanding is crucial to the productivity of bankers and portfolio managers. Call transcripts often touch on a range of topics within the same sentence or paragraph. For instance, the manager of a company may have to report a drop in customer retention and want to move on quickly to positive news about hitting an ESG milestone – SentiMine will clearly set those apart and help analysts to find the relevant updates needed to make a better-informed decision.

LSEG also offers a wide range of financial data – unstructured and textual data being particularly relevant in this context. This data is a prime candidate to improve the performance of GPT models given the increased role of ***prompt engineering.***

Prompt engineering is the careful construction of the input into a GPT model to improve performance of a specific task (5). The input into GPT can be quite detailed, including not just questions or chat history, but also significant amounts of relevant financial data; it is this aspect that is of particular interest to the financial industry given the existence of a variety of data-driven models and the significant industry knowledge thereof. With the release of GPT-4 series in March 2023 the input prompt has increased significantly, offering improvements in an area already ripe for exploration.

This paper explores using prompt engineering with GPT-4 for sentiment and theme classification. Section 1 provides a concise overview of the GPT family of models, with an emphasis on the practicalities relevant for the average user of GPT in the financial industry, such as costs. Section 2 provides results comparing GPT models with popular existing models such as BART (6).

1 To distinguish between series and individual models for GPT, we use capitals for the series, e.g., GPT-3, GPT-4 and we denote models by their lowercase names which represent endpoints available from OpenAI, matching their nomenclature. Examples of models are gpt-4 and gpt-4-32k. Note gpt-4 is the 8K token model, and gpt-4-32k is the 32K token model.

2 Aspect-based sentiment analysis categorises data by aspect and then identifies the sentiment attributed to each aspect.

LSEG. [Online] https://www.lseg.com/content/dam/lseg/en_us/documents/white-papers/ discovering-sentiment-in-finances-unstructured-data.pdf. Accessed: April 2023.

LSEG. [Online] https://www.lseg.com/en/labs/sentimine. Accessed: April 2023.

LSEG. [Online] https://www.refinitiv.com/en/financial-data/financial-news-coverage/marketpsych

LSEG. [Online] https://www.refinitiv.com/en/financial-data/company-data/quantitative-models/credit-risk-models/starmine-combined-credit-risk-model

Evolution of the GPT series

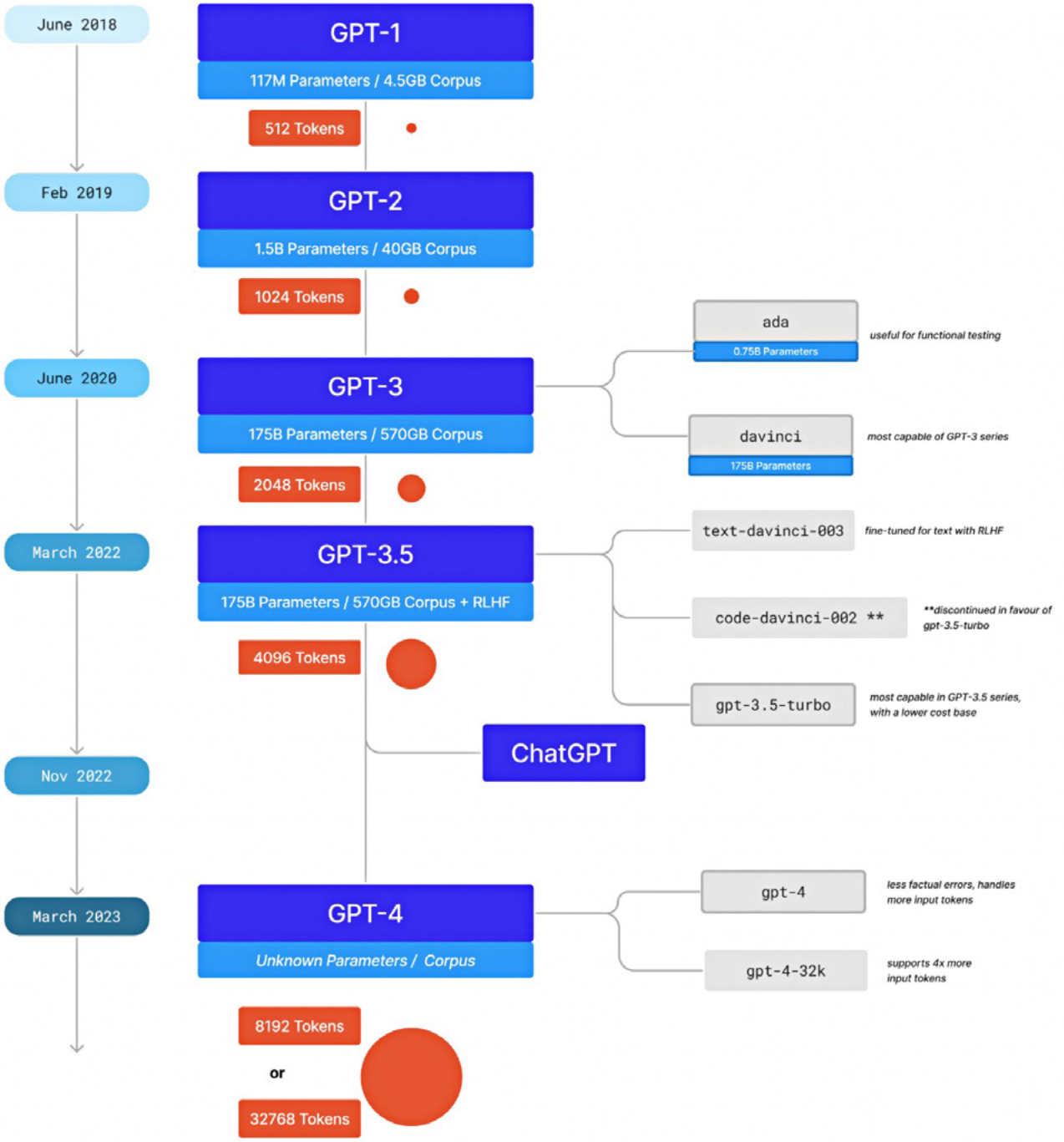

Fig. 1: GPT models available from OpenAI demonstrating the pace of updates up to the latest release of GPT-4 series in March 2023. The number of tokens is relevant to the quantity of data that can supplied to the model with each query. RLHF (Reinforcement Learning from Human Feedback) is a key differentiator between GPT-3 and GPT-3.5 series.

Understanding the current state of GPT is important when considering the practicalities of product development, such as performance, competitive advantage, limitations and costs. Generative Pre-trained Transformer (GPT) models are considered "general", in the sense that they can perform a wide range of tasks well but can still under-perform against existing techniques given a specific task; this is important to finance use cases given the highly competitive nature of the industry for specific tasks. The increasing pace of releases of GPT models poses the question as to if/when such generalised models could outperform existing LLM techniques.

Figure 1 outlines several key aspects of the GPT family of models such as release date, number of parameters, size of training corpus and size of input tokens.

- GPT-4 series was released in March 2023, less than six months after the previous release. It is plausible that model updates might become even more frequent; a major consideration when considering product development cycles.
- ChatGPT is a product powered initially by GPT-3 series and now analogous to gpt-3.5-turbo. Among the general public the term "ChatGPT" is sometimes used incorrectly as a catchall phrase for all GPT models. We wish to emphasise that while acceptable for more general conversations, this isn't suitable when comparing performance in financial use cases, where distinctions between GPT models is important.
- The input from a user into GPT models, known as a prompt, has increased in size with each subsequent release. This is of value when considering the potential to improve performance through the information included in the prompt; within the financial industry, this may prove of significant value when considering including financial data.
- The GPT-3, GPT-3.5 and GPT-4 series make several models available for use in each release (7) – more than are shown in Figure 1 (see references for details).
  - These are fine-tuned by OpenAI for specific tasks, such as for chat or coding. When comparing results, the specific model version is important.
  - We include code-davinci-002 as an example both of a fine-tuned task for code but also as an example of an API that is now discontinued in favour of a different API – in this case, gpt-3.5-turbo.

Tokenisation

Tokenisation of language is the process of splitting text into word, sub-word or character tokens (8). The GPT models process text using tokens. Tokens are the units that dictate the maximum amount of information passed into a GPT model and the units on which costs are calculated.

The increase in the number of tokens in GPT-4 series to 32,768 is important when considering the maximum quantity of information one can include in the prompts.

We recommend that interested readers consider the OpenAI documentation and tokenisation application (9). An example screenshot from this web application is shown in Figure 2, demonstrating how words and tokens are similar but not identical. As a rough guide, 100 tokens ≈ 75 words.

For context, the median length of a news article in the financial industry is approximately 250 tokens. Given an estimation of 5,000 news articles per day, this would equate to approximately 1.25 million tokens within news per day. This calculation demonstrates that currently it is possible to pass only small subsets of information to GPT-4 models, given the 32K token limit; therefore, for optimal performance careful consideration of processing input is required.

Fig. 2: An example of tokenisation of the headline text of a Reuters news article containing 98 tokens (9). Text taken from Reuters News website (10). Included in the final sentence are some names, highlighting how these are treated by GPT tokenisers.

Costs

A practical consideration of using GPT models is cost. Below is a table showing the costs of usage, as taken from OpenAI (11). To provide context, we consider 2,500 prompts each containing 150 tokens, for a total of 375K tokens – this is a similar number of tokens required for a typical exploratory experiment in the results section of this paper.

Model	Usage per 1K tokens	Cost of 375K tokens
ada	$0.0004	$0.15
davinci	$0.0200	$7.50
gpt-3.5-turbo	$0.0020	$0.75
gpt-4 8K context	$0.0300	$11.25

Prompt engineering, fine-tuning and pre-training

The performance of LLMs can be improved by adapting the base model using a different set of data from that on which it was trained. For example, GPT-3 series (base models) were trained using CommonCrawl, WebText, English Wikipedia and two book corpora (Books1 and Books2) (12). However, for financial applications we require that model be optimised for financial data. Domain adaptation is a very active research area for LLM with a wide range of techniques available (13). We shall comment on three popular methods; prompt engineering, fine-tuning and pre-training. Prompt engineering is the simplest and cheapest option; pre-training the most complex and costly. These methods are not mutually exclusive.

Prompt engineering

Providing more useful information to a GPT model via the input text – the prompt – can result in improved performance (14). This is by far the most accessible and cost-effective domain adaptation solution compared with fine-tuning and pre-training. The increase in tokens for GPT-4 series greatly increases the amount of information that can be passed to the model via the prompt.

Different pieces of information can be provided in the same prompt and can be quite varied. For example, it is reasonable to consider that the prompt may follow any of the following formats:

1. Perform a task, without context.
2. Perform a task, with example outputs.
3. Perform a task, with supplementary data.
4. Perform a task, with previous chat history from a user.
5. Perform a task, with example text, outputs and constraints.
6. Perform a task, specifying for polite and professional language.

Note that the costs of prompt engineering are directly proportional to the costs of usage as provided in the previous section; there is no additional overhead to prompt engineering.

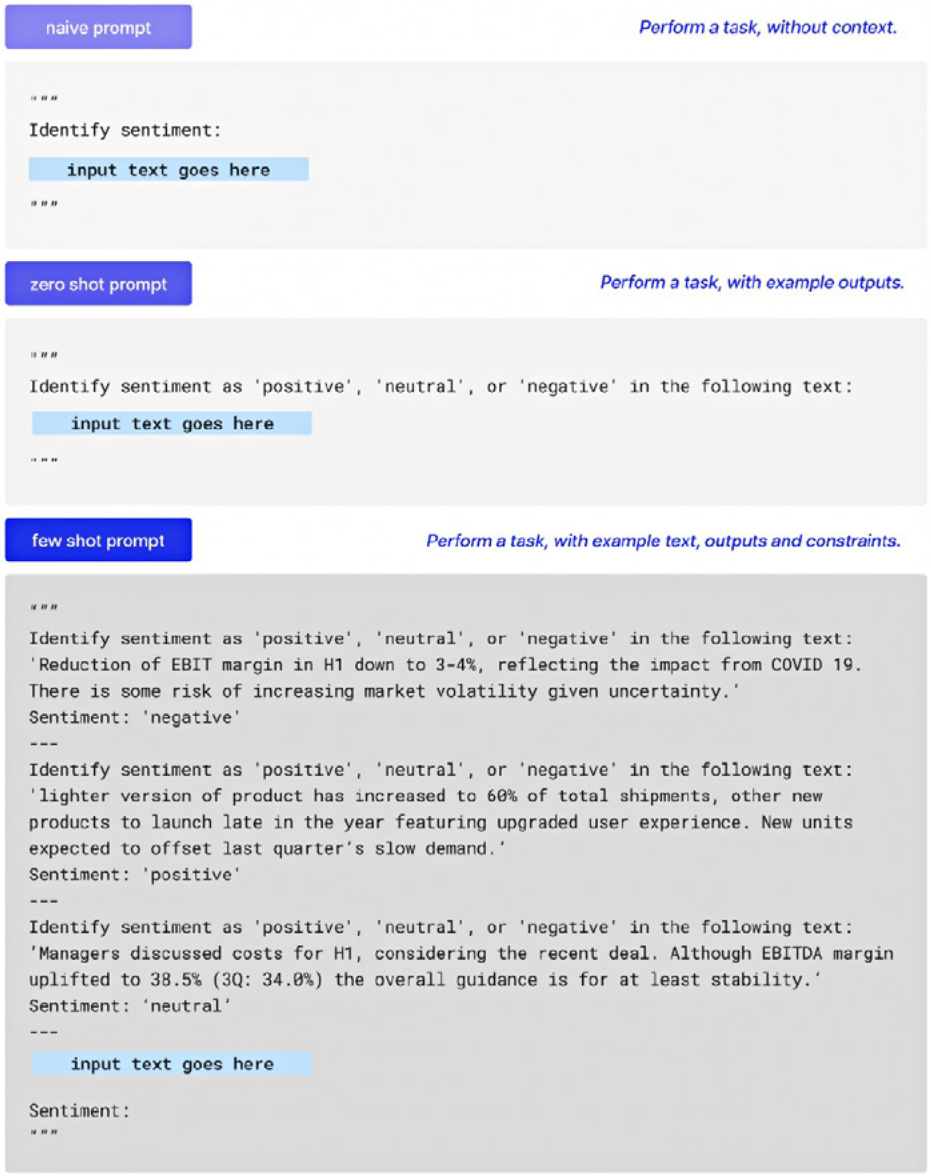

Fig. 3: Examples of prompting structure used to classify sentiment. The naïve[3] prompt was not sufficient to return reliable results. Zero-shot passes in no examples, few-shot passes in three examples. Note that these prompts include only items 1, 2 and 5 from the list in the above section; we envisage that a production system would have all items included in a longer prompt.

Fine-tuning

In fine-tuning, model weights are adjusted to fit domain specific information. This is typically in the format of a smaller corpus of examples that reflect the intended task (7). These could be hand-crafted examples as performance improvements are possible with small sample sizes – e.g., a few dozen examples. At the time of writing, fine-tuning is not available for gpt-3.5 turbo or gpt-4; it remains to be seen whether fine-tuning will be an accessible feature of future GPT models.

Costs for fine-tuning

The costs of fine-tuning GPT consist of two aspects; the cost to train the model and the cost to use the fine-tuned model (11). The following table shows a comparison of costs for fine-tune training and fine-tune usage per 1K tokens. The costs for the fine-tune training are lower than the costs for usage of a fine-tuned model. The right-hand column is the regular base model usage, showing that usage for a fine-tuned model is more expensive.

Model	Fine-tune training	Fine-tune usage	Usage of base model (no fine-tuning)
Ada	$0.0004	$0.0016	$0.0004
Davinci	$0.0300	$0.1200	$0.0200

Cost comparisons are all per 1K tokens.

3 We use the term "naïve" in relation to the concept that the prompt is too simplistic and won't return reliable results. This contrasts with zero-shot and few-shot prompts, which do return reliable results. The terms "zero-shot" and "few-shot" are common terms; we have introduced the term "naïve prompt".

Pre-training

Pre-training is the technique of training a large neural network on a corpus of text data, a technique common to many LLMs. The output of pre-training is a new base model. It allows for the new base model to learn a general representation of the language supplied. Note that when discussing pre-training, we consider only architectures similar to GPT models, not LLMs more generally.

Costs for pre-training

Pre-training incurs a one-off cost when creating the model. Training these models typically takes tens of days, often several months, using virtual machines costing approximately $2,000 per hour. Therefore, the overall approximate cost for compute for pre-training is in the millions.

Section 2: Results for theme and sentiment classification

We report the performance of gpt-4 for theme and sentiment classification, which is a typical use case in the financial industry. Although this work is exploratory only, there is a clear indication that gpt-4 rivals existing models in this task. We also explore using additional information in the prompts, demonstrating an ability to significantly improve misclassification using only a few examples.

Data

Theme and sentiment classification are common problems in the financial industry. LSEG has significant expertise in this field. LSEG's aspect-based sentiment analysis product – SentiMine – offers sentiment scores for financial documents; it is available as part of Workspace and leverages LLMs extensively (15). SentiMine requires optimising the accuracy of sentiment and theme classification and to this end LSEG has performed extensive model tuning and selection experiments, analysing over 100 financially relevant themes in the process. During the development of SentiMine, we observed several challenges when assigning sentiment to financial statements from long-form documents – i.e., transcripts and equity research – for instance, correctly picking up mixed sentiment within a statement, risks expected vs risks realised, sentiment as expressed numerically vs verbally.

The data set chosen for our exploration of GPT models is a set of challenging examples available for theme classification; this choice of difficult examples was intentional as we wished to see how GPT would perform in an area considered suitable for improvement. We wish to emphasise that the examples we have used do not reflect the SentiMine production choices. Importantly, while the results we present are inspired by our knowledge and understanding of LLMs in building SentiMine, they are not suitable as a valid comparison to the SentiMine product. Instead, our results are designed to help build understanding and education on GPT performance and its generalisability to such tasks.

The data used consisted of 2,747 sentences annotated into nine themes and 882 sentences annotated for sentiment. The following tables show the nine themes and three values used for sentiment, along with an example instance of the annotated data. The company name has been redacted from this sentence for this paper.

Task	Categories
Theme classification	“Cloud computing”, “Cost-to-income ratio”, “Customer experience”, “Epidemics” “Marketing and advertising costs”, “Mobile network operator (MNO)” “Mobile virtual network operator (MVNO)”, “Non-interest income”, “Shares buyback”
Sentiment classification	Positive, neutral, negative

Example sentence	Theme label	Sentiment label
Pulling subscribers away from [company] to become more difficult. Having taken the lead in lowering rates, [company] should fare well on the sales front. As such, we see its net subscriber gains improving, including migration from other MNOs and MVNOs.	Mobile network operator (MNO)	Positive

Prompt input

The following figure (Figure 4) shows the text used for zero-shot results for theme and sentiment classification. This was considered the minimum amount of information required to achieve consistent results using GPT models.

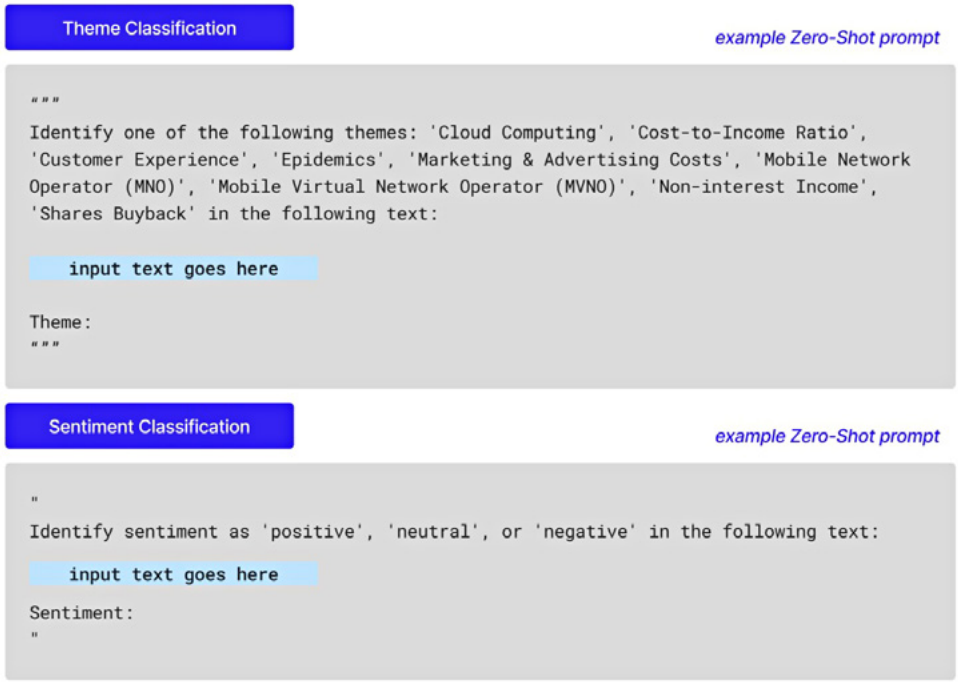

Fig. 4: Structure of zero-shot prompting used in the results section.

Comparison of models

In the table below we summarise the different models used in our experiments. For the GPT family of models, GPT-3, GPT-3.5 and GPT-4 series are used. BART (6) and FinBERT (16) are used to provide benchmarks; BART is used for both theme and sentiment classification, FinBERT is used for sentiment only.

For the GPT results, the temperature parameter was set to 0 to strive for the most deterministic results possible (1).

Model	Description
GPT-4 series: gpt-4	Highest current capabilities and optimised for chat, with processing for even more tokens (choice of 8,000 or 32,000 tokens).
GPT-3.5 series: gpt-3.5-turbo	Optimised for chat, with superior capability to GPT-3 series models. This model is capable of processing 4,000 tokens.
GPT-3 series: davinci	Base model (175B parameters), most capable in the series of base models of generating text.
GPT-3 series: text-ada-001	Fine-tuned ada model (0.75B parameters) capable of small tasks. Fastest model in the series and lowest cost, suited to functional testing. Performance was expected to be poor in comparison to other models.
BART	Bidirectional and Auto-Regressive Transformers (BART) is a sequence-to-sequence model, trained to reconstruct noisy text enabling high-quality text production.
FinBERT	Specifically trained for financial text analysis, using a corpus of regulatory filings and financial news. Includes financial domain vocabulary.
Sentiment classification	Positive, neutral, negative.

Results for zero-shot

The following tables present the results for theme and sentiment classification for zero-shot prompts. Accuracy is used as the primary result to aid in general communications, given it is relatively easier to comprehend. Also included is the number of correct and incorrect predictions, again to aid comprehension of results.

The f1-score is provided as it is the standard metric for classification tasks. The f1-score is calculated as the weighted average of all classes. This represents the aggregation of performance in all prediction categories, with respect to the number of samples in each category – giving importance to imbalanced data.

The ada model is included in the theme results to highlight its poor performance; this is as expected. The ada model routinely didn't predict themes from the set requested.

FinBERT is included only in sentiment, given that it would require fine-tuning with the theme data to provide a valid comparison.

Zero-shot theme classification:

Model	accuracy	f1-score	Correct predictions	Incorrect predictions
BART	0.786	0.784	2158	589
text-ada-001	0.059	0.082	161	2586
davinci	0.847	0.861	2326	421
gpt-3.5-turbo	0.881	0.887	2419	328
gpt-4	**0.941**	**0.945**	**2584**	**163**

Zero-shot sentiment classification:

Model	accuracy	f1-score	Correct predictions	Incorrect predictions
FinBERT	0.450	0.381	397	485
BART	**0.660**	**0.626**	**582**	**300**
davinci	0.413	0.430	364	518
gpt-3.5-turbo	0.536	0.575	473	409
gpt-4	**0.626**	**0.641**	**552**	**330**

Results for few-shot

We wish to demonstrate the value of prompt engineering by comparing few-shot prompts with zero-shot prompts. The few-shot experiments were designed based on the zero-shot results shown above.

We constructed a few-shot prompt containing three examples that had been correctly predicted by GPT models using zero-shot prompts. The few-shot prompt is very similar to Figure 3 in Section 1. We then re-ran the above zero-shot experiments for all sentences except the three examples used in the few-shot prompt. We report these results for sentiment only, given that theme classification is already relatively high using zero-shot prompts. Note that the zero-shot results here are very similar to those above, bar removing the three examples used in the few-shot prompts.

Few-shot sentiment classification:

Model	Prompt	accuracy	f1-score	correct predictions	incorrect predictions
gpt-3.5-turbo	zero-shot	0.536	0.575	473	409
gpt-4	zero-shot	0.626	0.641	552	330
gpt-3.5-turbo	few-shot	0.658	0.651	578	301
gpt-4	few-shot	**0.679**	**0.682**	597	282

Using gpt-4, there were 91 correct classifications using few-shot that zero-shot had classified incorrectly. There were 44 incorrect classifications using few-shot that zero-shot had classified correctly.

Using gpt-3.5-turbo, here were 177 correct classifications using few-shot that zero-shot had classified incorrectly. There were 69 incorrect classifications using few-shot that zero-shot had classified correctly.

Discussion of results

The results appear promising for the use of GPT models and gpt-4 for the specific task of theme and sentiment classification.

In all examples, gpt-4 showed improvement over davinci (GPT-3) and gpt-3.5-turbo. This indicates that there are likely performance advantages in using the latest models in the GPT family. This performance increase also comes at an increase in usage costs.

All GPT models were good at predicting theme using a zero-shot prompt, outperforming the benchmark model (BART). This is potentially related to the data chosen; as we selected the most difficult sentences from a larger set, it is plausible these sentences were biased against the benchmark models for reasons that are not yet studied. Regardless, given the results, GPT models appear to offer value for this specific task and are worth further investigation.

BART outperformed GPT models in sentiment using a zero-shot prompt in accuracy. However, gpt-4 was the top-performing model in both tasks according to its f1-score. Overall, the sentiment values for all models were lower than expected, which again may be due to the selection of challenging data. The performance of gpt-4 in a zero-shot setting is still very encouraging given how similar its results are to BART's.

Of particular interest is the improvement achieved using few-shot over zero-shot for sentiment, with both gpt-4 and gpt-3.5-turbo showing improved accuracy. Overall, the best results were obtained with gpt-4 using few-shot, which is consistent with all other results in indicating that gpt-4 is the most performant model of the GPT family.

Even though our few-shot experiment is a relatively small study, we consider these results to be very promising due to the many variations offered by prompt engineering and the amount of data we can potentially use in the prompt. We see our exploratory results as confirmation that prompt engineering is worthy of further exploration.

A few interesting examples...

Here are a few examples to help enhance understanding of GPT models:

- Occasionally, gpt-4 predicted a theme that was similar, but not identical, to a theme in the set requested. In one such example, for the theme labelled "Epidemics" gpt-4 returned the response "COVID". This may be due to our prompt not being sufficient. Regardless, given the semantic similarity of the results, it is interesting to consider what information could be contained in misclassified results.
- The zero-shot prompt was the simplest prompt with which we could obtain reliable results; more naïve prompts were liable to return additional characters that were erroneous, such as Roman numerals. Using naïve prompts places an additional overhead on post-processing results. It also raises questions as to whether results could be misleading if prompts are not well-formed

– Contained in the data was one sentence used only for unit testing, which consisted of the word “nan”; the response from gpt-4 detailed feedback as to why this input isn’t suitable for theme classification.
  – **Input:** “nan”
  – **Output:** “There is no theme present in the given text as it only contains “nan”, which stands for “not a number” and does not provide any information related to the mentioned themes”

These few examples are indicative that post-processing of output from GPT models is likely a more important consideration compared to traditional machine learning systems. There appears to be information of value for both misclassified results and input data input errors.

Conclusions and future work

The financial industry is only beginning to discover the value of GPT models. Our exploratory analysis using GPT-4 for financial industry tasks shows promise, with the latest GPT models demonstrating clear performance improvements over existing models.

Although we have reported results for the specific tasks of sentiment and theme classification, there are clearly opportunities for using GPT far beyond these tasks. These tasks were natural starting points for experimentation given our experience in these areas using LLMs. It is likely GPT offers new ways for users to interact with models and data given their generative capabilities, which we are keen to explore.

In the foreseeable future, prompt engineering is likely to form a large part of any GPT project, given how accessible and cost-effective it is compared to fine-tuning and pre-training. We seek to harness the increase in available input tokens with the release of GPT-4 series, specifically by using financial data in the prompt.

Finally, the pace of updates to GPT and the performance improvements of GPT-4 are captivating for product implementation; we look forward to discovering new opportunities in this fast-growing field.

Acknowledgements

Authored by LSEG Analytics

David Oliver
Director, Data Science, LSEG Analytics

Aran Batth
Senior Data Scientist, LSEG Analytics

Stanislav Chistyakov
Junior Data Scientist, LSEG Analytics

Mihail Dungarov, CFA
Text Analytics, Product Lead, LSEG Analytics

The authors wish to thank the following contributors:

Will Cruse and Dinesh Kalamegam in LSEG Analytics for infrastructure support for testing GPT

Evgeny Kovalyov and Anna Stief in LSEG Analytics for ongoing contributions

Jingwei Zhang for contributing to the evolution of GPT section

Rachel Sorek, Rani Shlivinski, Lior Gelernter Oryan and others in LSEG Applied NLP

References

1. *GPT-4* Technical Report. **OpenAI**. 2023.
2. *Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.* **Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, Bao Ge.** 2023.
3. **LSEG.** [Online] https://www.lseg.com/content/dam/lseg/en_us/documents/white-papers/discovering-sentiment-in-finances-unstructured-data.pdf. Accessed: April 2023.
4. **LSEG.** [Online] https://www.lseg.com/en/labs/sentimine. Accessed: April 2023.
5. *Language Models are Few-Shot Learners.* **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, Da.** 2020.
6. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation and Comprehension.* **Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer.** 2019.
7. **OpenAI.** [Online] https://platform.openai.com/docs/models/overview. Accessed: April 2023.
8. *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP.* **Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, Samson Tan.** 2021.
9. **OpenAI.** [Online] https://platform.openai.com/tokenizer. Accessed: April 2023.
10. **Reuters.** [Online] https://www.reuters.com/markets/global-markets-view-europe-2023-04-17/. Accessed: April 2023.
11. **OpenAI.** [Online] https://openai.com/pricing. Accessed: April 2023.
12. *A Survey of Large Language Models.* **Wayne Xin Zhao, Kun Zhou*, Junyi Li*, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jia.** 2023.
13. *RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning.* **Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, E. Xing, Zhiting Hu.** 2022.
14. *Prompting GPT-3 To Be Reliable.* **Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, Lijuan Wang.** 2022.
15. **LSEG.** [Online] https://www.refinitiv.com/en/products/refinitiv-workspace. Accessed: April 2023.
16. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.* **Araci, Dogu.** 2019.

Discover more at lseg.com

LSEG

LSG2837757/5-23