Augmenting Out-of-the-Box LLMs with API Calls for Annual Report Auditing

Cathal Murtagh
Dublin City University
Dublin, Ireland
cathaljmurtagh@gmail.com

***Abstract*—Large Language models or LLMs have garnered much attention recently for their huge improvements over older chat bot models. There has been significant discussion about the potential business cases and many companies have seen their stock premiums soar over the potential advantages from using said models such as productivity gains and cost decreases. Despite this, there have been few if any deployments of such models in real situations. Many, in areas such as professional services, have opposed the uptake of these systems largely over issues such as "hallucinations". These are instances where a model will output a response to a query which is incorrect in some way. This leaves them essentially unusable in the eyes of many who otherwise might find great use for them. In this study, we seek to evaluate the current state of LLMs for use in professional work, namely the fields of finance and accounting, that is usually undertaken by highly skilled individuals or groups.**

I. Introduction

A. Background

The emergence of transformer [VSP+17] based Large Language Models has led to substantial advancements and research [RNSS18] into their application in a multitude of fields. Prior to this, Recurrent Neural Networks [RHW86] were the predominant model architecture used in Natural Language Processing and are seen as starting the industry down the path of scaling systems with multiple GPUs, as was first done with AlexNet [KSH12]. Accounting and finance are areas in which there is significant potential for the use of Large Language Models owing to the large amount of text processing that must be done. There are however a number of barriers preventing uptake and more widespread use of them in the field. In particular, there are concerns about potential mistakes or 'hallucinations' made by language models, where the output is not based in actuality or material fact. As accuracy and explainability are two key requirements for all facets of accounting and finance such errors would be unacceptable. Such negative perceptions have limited potential uptake and research of LLMs in these areas of business. The aim of this study is to explore whether state of the art LLMs can produce acceptable outputs for financial auditing while minimising hallucinations. In particular, the study examines the use of LLMs in facilitating the auditing process by speeding up the process of data ingestion and allowing for non trivial analysis of complex documents.

The task that will be addressed relates to the auditing process [KS15]. More specifically, the auditing of a company's financial statements, focusing on the 10k annual report released by all public American companies in accordance with Securities and Exchange Commission (SEC) requirements. This is an arduous and resource intensive process for auditors, as they must comb through vast quantities of company data to formulate their reports and ensure that the company they are auditing is in full compliance with relevant regulations before they sign off on the publication of said financial statements. These financial statements are long documents which contain the operating results for the company over a certain period, including information such as profit and loss accounts, the balance sheet, discussions and challenges for the firm as well as forward earnings guidance. Importantly, these documents are often used by individual and institutional investors to inform their decisions on whether or not to invest in a company.

B. Industry partner EY

This investigation was carried out in partnership with EY Ireland. EY is an international professional services firm that provides assurance services such as auditing and accounting advisory services as well as consultation on other business matters, including transaction advisory and taxation compliance, among others.

One of the many services that firms like EY offer is that of auditing a company's financial accounts, ensuring them to be in compliance with local regulations. As noted earlier, proper accounts, along with skilled and truthful auditing are required to ensure financial health and prevent fraud from taking place inside the firms. Ensuring accurate reporting is of paramount importance for stakeholders as it is their money and often livelihood at stake. Despite this, sometimes individuals or groups within a company may stand to profit from falsifying company data in the short term [Tho02]. For example, managements compensation is often determined by factors such as stock price, revenue or other accepted metrics that often depend on the specific industry. The SEC in America or International Financial Reporting Standards (IFRS) [Int23] elsewhere have a set of standards of practice that must be adhered to. In America, these are known as Generally Accepted

Accounting Practices (GAAP) [Fin23]. Sometimes, a company may decide to set a non-GAAP measure as a key profit metric and tie executive compensation to this metric. This is not against the law, but it often attracts scrutiny from auditors as it is less common and potentially easier to manipulate as the calculations used to generate these figures are not standardized and are subject to change and revision by the company.

C. Overview of task and approach

To evaluate an LLM's viability in the task of auditing the financial reports of companies, we will assess it within the following setting: We take the annual reports of listed companies for 10 years, and, using the workflow of Auditors from EY, evaluate the outputs versus accepted answers. The exact process for this is different for each auditor and company. For the purpose of this investigation, a standardized approach is taken where all companies are subject to the same format. The annual reports in question are taken from a diverse range of companies listed on the S&P 500 index of companies. These are generally companies are that are well respected, while some have been accused of certain accounting irregularities at different points by regulators and whistle blowers. The irregularities themselves are quite complex in nature and thus pose a significant challenge to identify and evaluate. LLMs have previously been investigated for their capabilities in performing the duties of analysts [CLB23] and other task areas, however these tests were rather straightforward and did not involve many layers or comparisons across very large documents.

In the experiments conducted, the individual annual reports range from about 100 to 500 pages in certain cases. This presents a challenge, as the longer a passage gets, the higher the computational requirements to process it are [VSP+17]. Most LLMs have fixed context windows of between 8 thousand and 128 thousand tokens. A token represents a word, punctuation or a part of a word. It is the method used by LLMs to efficiently parse text, with modern LLMs using derivations of the Byte Pair Encoding technique to do so [SHB16]. Some LLMs such as Googles Gemini [TAB+24] offer effective context lengths of over a million tokens, however this does maintain high computational requirements and in our case is still insufficient. A popular method of overcoming this shortfall in memory requirements is that of Retrieval Augmented Generation, or RAG [LPP+21]. At a high level, RAG involves creating a database of the documents you wish to query, which can involve different strategies such as chunking the text and creating vector databases from the chunks.

One aspect that proves challenging is how to format an auditor's workflow into prompts that an LLM can use, while minimizing hallucinations. For this we relied on EY and their expertise to help craft an appropriate method of querying that was usable while not being too leading. E.g. so that the queries did not point directly to the desired result and have the LLM simply parrot from this.

II. Related Work

There have been a number of attempts to evaluate the capabilities of LLMs in numerous fields such as fulfilling the role of an analyst [CLB23], in medicine [TTE+23] and in law [LGW+23]. The task of auditing the statements of a company incorporates many aspects which an LLM on paper should succeed in such as large amounts of knowledge about the world, but also levies some of their direct weaknesses at this point in time, like multi step reasoning and accuracy. Attempts have been made to develop LLMs solely focused on the specific domain of finance [WIL+23] but these have fallen behind as larger and more powerful general or foundation models are released surpassing the abilities of specialized models. Older larger models have also been surpassed on evaluation benchmarks, even in some task specific domains by smaller, newer models that are trained on larger quantities of data [TMH+24]. The trend has moved from pretraining models on specific areas, to pretraining on a wide variety of content and then fine tuning on the specific task [TMS+23]. Other methods have included taking an existing foundational model and performing continual training on specific domain data [XAA23].

Owing to the quadratic nature of the attention mechanism in the transformer architecture [VSP+17], the maximum context length is decided during training and the longer a sequence of text goes on, the less accurate the outputs from an LLM may be. These problems are rooted in the training phase of the model and outside the scope of this experiment, but their effects weigh on the task. Google released a method for massively increasing the context length of models [MFG24] but it remains to be seen whether this impacts reasoning tasks. Even with these advancements, the significant length of passages of text that many use cases present necessitates a different approach. RAG methods offer the potential to limit the effect of context length for certain tasks. First coined by Meta [LPP+21], RAG has become a cornerstone of LLM use, with firms such as Cohere making it the key selling point of their models. Without it, the limited context length would prevent or drastically slow any system utilizing these models. RAG has been used in numerous domains, such as the creation of question answering systems [MHL+20] and is advertised as the key component in many enterprise focused deployments. Differing methods of retrieval methods and how to retrieve relevant information is also a very active area, with many different proposed methods [JXG+23] [ZZY+24] such as hybrid retrievers, re-ranking, fine tuning the retriever and optimization of chunking.

Model hallucination has been a common term associated with LLMs in spaces from business to academia. This term refers to the potential for an LLM to output incorrect or misleading information [TZJ+24]. Hallucinations occur because at their core, LLMs are next token prediction models. There have been numerous high profile gaffs, particularly from Google. The outputs however, can be controlled to some degree with sampling strategies which determine how the next tokens are

selected [SYC+24], with methods being either deterministic, or stochastic. For the purposes of this study, all outputs where possible are sampled in a deterministic manner [LPX+22].

Uses of LLMs to perform cognitive tasks previously reserved for educated professionals have received significant attention. LLMs such as ChatGPT made headlines when the models they used passed reputedly difficult exams [GSH+23] although there is discussion on the validity of these results, owing to some answers being present in training data. Questions remain as to what extent LLMs generalise to tasks and how much is attributable to memorization. A number of tests have also been done on the capabilities of these models to perform real world jobs. In [CLB23] the authors investigate the ability of GPT4 in analysing problems and generating solutions. The results are then compared to human analysts of different skill levels. One limitation of this approach was that the questions posed were rather simple 'toy' problems with no noise and limited 'distraction' for the model. LLMs have been investigated in other areas such as law and medicine [DVCT+24] showing the potential for use in similar knowledge intensive domains. AI has been incorporated in numerous ways in these fields [HT17] for quite some time, however LLMs are a new and somewhat contentious topic. The fields mentioned share similar issues to that of accounting and finance, where they require very high levels of accountability and accuracy.

A major point that similarly gets mentioned when it comes to employing LLMs in spaces where accountability is important is the interpretability of LLM outputs. The outputs can be controlled to an extent in their format and nature, but efforts to actually understand just why the model outputs what it does is still at an early stage [BTB+23] and is an unsolved issue for this domain. Prompt engineering is another key facet of the discussion. The construction of well defined prompts is imperative if an LLM based system is to achieve peak performance. Small, seemingly unimportant details such as being polite [YWH+24] can materially affect outputs.

III. EXPERIMENT

A. Methodology

The task we investigate is whether LLMs utilising RAG can be used to accurately analyse long, complex documents from specialized domains, in this case the field of accounting and finance. The dataset we use for this task is constructed from annual company reports taken from publicly listed American companies. We chose 3 companies and took 10 of these annual reports from each. We kept each set of data separate, leaving us with 3 datasets of between 1 million tokens and 3 million each, too long for even Google's Gemini [TAB+24] model to handle in context, necessitating a RAG approach. The questions we evaluate the LLMs on are provided by EY and are consistent for each company. All constructed databases are kept separate from each other. This minimizes hallucinations from incorrect retrievals as well as allowing for changes or replacements to be more easily made.

A RAG vector database [Tai24] is a store for word embeddings. In present context, these embeddings are representations of the words or information contained in the annual reports with dimensionality of 1024 tokens. For retrieval, the query is converted into dense vectors using the embedding model. A similarity search is performed on the database using the query which returns the top K most similar items, top K referring to the returned items which are attributed the highest relevance scores. Relevance scores are calculated using cosine similarities of the extracted passages. The LLM uses this information as context to generate an output along with the initial query.

A Knowledge Graph [CJX20] or RAG graph is a graph database that stores entities and the relationships between them. These entities can represent people, companies, products, etc [FRZ+20]. The relationships between these stored entities might be "CEO of" and "located in". This allows for more rich information storage and retrieval in comparison to basic vector stores as vector stores only perform a top k similarity search on chunks of text, which can lead to incorrect or overly general results.

B. Setup

We chose to test 3 separate models for this experiment. We took Llama 3 8B instruct as our baseline model and compared it with a fine tuned variant of Llama 3 called finance-Llama3-8B as well as GPT-4o for our closed source model comparison. The Llama models are sourced from the popular online machine learning platform Hugging Face and GPT-4o from OpenAI. In testing however, finance-Llama was found to be incredibly poor for the chosen task. There is also notably a significant size discrepancy between the open weights and closed source models. One consideration in choosing which models to use in a production environment is cost versus quality. For this reason we take some of the smallest but high performing models and pit them against state-of-the-art counterparts to investigate the results.

For each model, we initialised vector databases, created using the annual reports as input and utilising a separate vector database for each company. We did this for each LLM and each embedding model. We then also compare this with a Graph RAG approach, which is a method that uses LLMs to extract meaning and relationships from unstructured text, creating a graph with connected nodes out of the data. In the construction of these nodes, some LLMs outputs resulted in errors as they sometimes produce invalid json. If enough of this happens, it breaks the process and it fails. This is also a computationally intensive task and required 3200 LLM calls for a single company, meaning failures can be expensive in both time and money. In this case, a single LLM call refers to each time the model processes a batch of text. After experimenting with different models and parameters, the model google/gemma-2-9b was chosen to construct the graph, owing to its high performance and low cost [TMH+24]. While not on the performance level of the largest model, it strikes a balance between compute and capability that fits the scope of this investigation.

The knowledge graph approach involves the extraction of entities, such as companies, products or places as well as the relationships between these items. For example, the entity or node (person) "John Doe" might have the relationship of "employee of" to entity "XYZ company Limited". The LLM extracts these relationships and creates structured tables from the unstructured text. On top of this, a key difference in this approach versus more traditional graphs is that the LLM also generates summaries of the text and relates them to the relevant entities with their relationships. This means that at inference time there is significantly more rich data available for the LLM to use in generating its response. It can then include references to these documents in its response, allowing easier fact checking of the response.

To investigate the effect of domain specific knowledge, we also compare the results when using a base embedding model, in this case BAAI/bge-large-en-v1.5 versus a fine-tuned variant Xeolus/fin_embed_large. This is the same base model of BAAI/bge-large-en-v1.5, fine tuned on 50 financial documents including 8ks, 10ks and form 144s which is roughly 7 million tokens. The documents were split into question/answer pairs on summaries of sections of the reports. This was then used as the input for the training run. None of the documents used for this process included the documents used for the main experiment. For the hyperparameters of the training run, we set the learning rate to 2e-5, implemented a WarmupLinear scheduler to start the learning rate at 0 and scale it up to 2e-5 over 88 warm up steps and a liner decay of 0.01 to prevent over fitting. We also cap gradient norm to 1 to improve stability in training.

In order to ascertain the success of the fine-tune, tests on information retrieval are done, comparing the base embedding model with the fine-tuned variant. A held out set of different but similar financial documents is used as the test set. Results can be seen in table 1. For the comparison, accuracy and precision are used where: Accuracy (cos_sim-Acc@1 and cos_sim-Acc@10) refers to how often the top result, or top 10 results, returned by the model are relevant to the query and can be thought of as a measure of how correctly the model identifies the best match.

Precision (cos_sim-Prec@1 and cos_sim-Prec@10) measures the proportion of the retrieved results that are relevant. Specifically, Prec@1 is the relevance of the top result, while Prec@10 considers the relevance of the top 10 results.

The questions used to evaluate models are designed to follow the sample workflow similar to that of an auditor going through the reports. An auditors may vary, but for the subject of this investigation, it was taken that the task in question was to review the company on the subject of its accounting standards and its practices relating to management remuneration. The questions are broken into difficulties of low, medium and high by EY. This relates to how hard it is to conclude or find the answer in the text. For example, some questions may reference a broad metric such as capital expenditure. However, in a certain document, it may not be explicitly mentioned and is instead broken down into purchases of some fixed asset. In this case the LLM must accurately relate the answer in the text to the query term even if the terms are lablled differently. An example question from the set is: "Are the metrics management are remunerated on based on GAAP or non-GAAP metrics?".

For the evaluation stage, the LLM's outputs were taken and scored against the correct answers and given a score between 0-3. Results were also computed for standard machine evaluation methods such as BLEU [PRWZ02] and ROUGE [Lin04]. As much as possible, parameters were kept constant and similar across the experiments.

IV. RESULTS

One part of the problem in dealing with a dataset such as the one used here is that information in one year is often corrected or amended in a later year, meaning that within the set of documents, there are different figures referring to the same item. This was expected going into the experiment, however upon conducting tests, it was found that all methods struggled greatly with retrieving numerical data from across multiple time periods. This was especially the case if inconsistent naming schemes were used across companies. EY indicated the numerical questions to be the easiest out of all, however every method involving an LLM proved inadequate in dealing with questions of this sort. For the vector based methods different chunking sizes, prompts, temperature and retrieval strategies were tried, but none improved performance meaningfully in relation to numerical queries.

The reported results all use similar parameters where possible such as temperature 0.0, top k=2 except in the case of the Fin Llama model where extra effort was taken to export results owing to its poor performance. Different prompt formats, varied temperature and retrieval was used. The final format for Fin Llama being k=50, temperature of 0.75 and a change in the prompt format which instructed the model to ignore any response which did not contain relevant context and responded with "the answer is not contained in text". This was necessary to obtain usable results from the model.

In comparisons between base embeddings and the fine-tuned variant, the fine tune significantly outperformed the base, indicating the impact that a small degree of specialization can have on the construction of the vector stores. In contrast to this, the fine tuned Llama 3 variant performed terribly, indicating a poor fine tune. Its outputs were incoherent for the majority of the queries. When fine tuning a model, there is the potential that there can be performance loss in some generalities as the weights are updated, however it should not demonstrate the loss of ability seen in this experiment.

The clear winner overall was the knowledge graph based approach, scoring higher than all the other methods. There are a number of notable points to be touched upon. First, the simpler vector based approach when combined with domain specific embeddings demonstrated comparatively similar results in many of the queries. It must also be mentioned that the construction of the actual graphs was vastly more computationally intensive than that of the vector stores. A graph was created per company and each involved thousands

Model	cos_sim-Acc@1	cos_sim-Acc@10	cos_sim-Prec@1	cos_sim-Prec@10	cos_sim-@1	cos_sim-@10
BGE Large 1.5	0.513663	0.849879	0.513663	0.084988	0.513663	0.849879
FIN_EMBED	**0.592183**	**0.893462**	0.592183	0.089346	0.592183	0.893462

Table I

RESULTS OF EMBEDDING FINE TUNE. ACC = ACCURACY, PREC = PRECISION AND @1/@10 IS RESULT IN FIRST DOCUMENT/RESULTS IN 10 DOCUMENTS RETRIEVED.

Rating	Description
0	Answer not contained in text or totally incorrect
1	On target, but not all information returned is relevant or accurate
2	Generally correct, but lacking sufficient detail
3	Accurately answered query, referencing the data

Table II

RESPONSE QUALITY RATING SCALE

	Tingo	Newell Brands	Kraft Heinz	Average Score
GPT-4o + BGE	0.44	0.613	0.666	0.574
GPT-4o + FINEMBED	0.518	0.742	0.757	0.672
Llama + BGE	0.518	0.194	0.606	0.439
Llama + FINEMBED	0.703	0.709	0.606	0.673
Fin Llama	0.407	0.290	0.121	0.272
Graph Approach	**0.740**	**0.838**	**0.757**	0.779

Table III

HUMAN EVALUATION RESULTS

of API calls and processed millions of tokens with both the embedding model and LLM which, when using larger models can rapidly lead to rate limits. For this reason, a less performant LLM, but still state of the art for its size, was used in constructing the graph (Gemma 9B). The quality of the graph and thus the results obtained could be improved by using the largest and most performant models available. The quality of the graph must also be compared with the cost to inference it, as even after it is completed, queries referencing it will require on average 10 API calls and 80k tokens. The granularity of searches with the graph can also be modified, allowing more or less specific, and thus expensive, queries to be made.

In relation to explainability and hallucination minimization, the knowledge graph approach showed significant advantages over vector based methods. The knowledge graph provided document sources for each data point returned in responses. These sources were labelled in relation to how the original document was chunked and can be looked at to verify that the LLM is not hallucinating in its response. In the responses, the LLM was prompted to use the context provided by the returned chunks in the case of the vector stores, and the entities and community reports with the knowledge graph method. In each case, the model would indicate whether or not the information was available in the retrieved documents, and if so, used it in constructing its response while also stating where it could be found.

The machine evaluation metrics must be taken with context. Scores are generated by comparing LLM responses with reference answers, with higher scores indicating better results. The LLM responses varied in format, especially between the vector and the graph based methods. Some individual responses received extremely high scores, however the prevalence of long LLM explanations for questions that required only a short response heavily affects the scoring. The BLEU [PRWZ02] scores demonstrate mainly the lower quality of the Finance Llama approach, whereas the Graph RAG scored higher in human evaluation but significantly lower than most vector models. All models scored low across the board which is to be somewhat expected. The ROUGE [Lin04] demonstrate interesting comparisons with the human evaluation, showing a direct correlation between the poorer ROUGE scores in Precision, with higher Recall, correlating with higher human evaluation scores. This is likely the case as the Llama 3 FINEMBED and Graph RAG responses tended to be longer and more divergent from the reference. The outputs from this test can be seen in Table IV.

The nature of the responses from the LLMs merits discussion. The format of the graph based method relies on such a large number of LLM calls, that it allows for a far greater degree of customization in how the graphs are structured and what format the response will take. A more significant investigation into the optimum construction of said graphs would be worthwhile, with changes made to how the reports are constructed and to what depth the communities are queried. These are a few of the many hyperparameters that can be changed and they have a significant effect on the information extracted, as well as the overall computational cost of construction and querying. The hyperparemeters used were chosen as they threaded the needle between performance and cost.

The average score for the human evaluation across all documents and methods was 0.568, with average scores per company being 0.555, 0.564 and 0.568 respectively. This aligns with the general nature of the documents, where the first company listed and used to test the RAG systems had more

	BLEU Score	ROUGE-1 Prec	ROUGE-1 Rec	ROUGE-1 F-m
GPT-4o BGE	0.2078	0.2022	0.3366	0.1883
GPT-4o FINEMBED	0.1991	0.2143	0.3586	0.2017
Llama 3 BGE	0.1906	0.1226	0.3776	0.1309
Llama 3 FINEMBED	0.1936	0.0987	0.4156	0.1302
Finance Llama	0.1041	0.2035	0.2771	0.1556
Graph RAG	0.1547	0.0518	0.4701	0.0864

Table IV
AVERAGE SCORES ACROSS ALL DOCUMENTS FOR BLEU, ROUGE-1 PRECISION, ROUGE-1 RECALL, ROUGE-1 F-MEASURE.

nebulous reporting. Despite this, the other higher scoring companies had significantly more data in the reports, suggesting that document length did not have a material impact on the effectiveness of the system. If we then remove the outlier, that is the Finance Llama model and its outputs, the differences in scoring are more pronounced with the datasets scoring 0.585, 0.619 and 0.679. Anecdotally, this was the impression of the 'quality' or readability and ease of understanding obtained when reading the documents for the human evaluation stage. This implies that the main difficulty faced by the models is the complexity and specificity of the text.

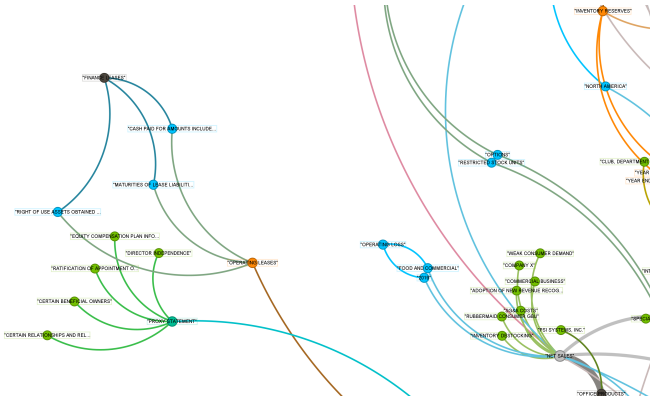

Figure 1. small subsection of one of the graphs created for a company.

V. CONCLUSION

Results from the investigation indicate the abilities of LLMs coupled with different methods of Retrieval Augmented Generation relying on API calls to correctly answer queries on highly specialised areas.

Although vector based methods did not out score the graph variants, there is still room for improvement on the retriever side. The low cost, simplicity and speed at which it can be brought up to a performant level leaves it as a valid first trial for any RAG system. Once greater functionality than that is required, Graph based systems are the optimum choice due to the fine grained customization that can be achieved. The one major drawback being the significant compute required.

The importance of using models with domain specific training is evidenced by the scores of the vector based models. Comparing the base embedding results against the fine tuned financial model presents a clear picture of just how impactful a positive fine tune can be on results. A key facet of the potential for LLMs to be used professionally, in this case in auditing, is not simply a direct correct or incorrect response, but also the degree to which it can direct a users' attention towards specific areas of a document. For this purpose, the Graph based responses offer extensive and customizable information about a query. For numerical examples, retrieving said data from different or inconsistent time periods leaves all methods struggling to return appropriate information. At their current iterations, LLMs will not outmode professionals, but they do present real opportunity to enhance the workflows of users, if the underlying system is well constructed, and queries are crafted with regard to the task and the retrieval process. This is a key facet of the process. Recommendations for improving upon systems such as Knowledge Graphs include further experimentation with the number of LLM API calls, testing different LLM response structures and improving the graph construction process by creating more edges or relationship links between nodes.

LLM based systems present a significant opportunity for use in fields such as accounting and finance. As costs continue to fall, they will become more attractive propositions for firms to use in their workflows. Overall, the processes described here present a significant improvement over baseline systems, demonstrating significant boosts in accuracy and detail. This paper serves to demonstrate the potential specialized systems have for use in these domains. Continued research and refinement of the structure outlined above should lead to higher accuracy and further efficacy, especially in the knowledge intensive fields such as accounting and finance.

REFERENCES

[BTB+23] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html, October 4 2023.

[CJX20] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948, 2020.

[CLB23] Liying Cheng, Xingxuan Li, and Lidong Bing. Is gpt-4 a good data analyst?, 2023.

[DVCT+24] G. D'Anna, S. Van Cauter, M. Thurnher, et al. Can large language models pass official high-grade exams of the european society of neuroradiology courses? a direct comparison between openai chatgpt 3.5, openai gpt4 and google bard. *Neuroradiology*, 66:1245–1250, 2024.

[Fin23] Financial Accounting Standards Board. Generally accepted accounting principles (gaap), 2023. Accessed: 2023-07-22.

[FRZ+20] Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3):bbaa110, 06 2020.

[GSH+23] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*, 9:e45312, Feb 2023.

[HT17] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017. Insights Into the Future of Medicine: Technologies, Concepts, and Integration.

[Int23] International Financial Reporting Standards Foundation. International financial reporting standards (ifrs), 2023. Accessed: 2023-07-22.

[JXG+23] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[KS15] Ravinder Kumar and Virender Sharma. *AUDITING: PRINCIPLES AND PRACTICE*, chapter 1, page 2. PHI Learning Pvt. Ltd., New Delhi, India, 2015.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[LGW+23] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey, 2023.

[Lin04] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[LPP+21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[LPX+22] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc., 2022.

[MFG24] Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention, 2024.

[MHL+20] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, 2002. Association for Computational Linguistics.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, chapter 8, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[RNSS18] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[SHB16] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

[SYC+24] Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of llms, 2024.

[TAB+24] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, and YaGuang Li. Gemini: A family of highly capable multimodal models, 2024.

[Tai24] Toni Taipalus. Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, 85:101216, 2024.

[Tho02] C. William Thomas. The rise and fall of enron: When a company looks too good to be true, it usually is. *Journal of Accountancy*, April 2002. Accessed: 2023-07-04.

[TMH+24] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

[TMS+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[TTE+23] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, et al. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023.

[TZJ+24] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in

large language models, 2024.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[WIL+23] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[XAA23] Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. Efficient continual pre-training for building domain specific large language models, 2023.

[YWH+24] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance, 2024.

[ZZY+24] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.