Retrieval-Augmented Chain-of-Thought in Semi-structured Domains

Vaibhav Mavi
New York University
vm2241@nyu.edu

Abulhair Saparov
New York University
as17582@nyu.edu

Chen Zhao
New York University
cz1285@nyu.edu

Abstract

Applying existing question answering (QA) systems to specialized domains like law and finance presents challenges that necessitate domain expertise. Although large language models (LLMs) have shown impressive language comprehension and in-context learning capabilities, their inability to handle very long inputs/contexts is well known. Tasks specific to these domains need significant background knowledge, leading to contexts that can often exceed the maximum length that existing LLMs can process. This study explores leveraging the semi-structured nature of legal and financial data to efficiently retrieve relevant context, enabling the use of LLMs for domain-specialized QA. The resulting system outperforms contemporary models and also provides useful explanations for the answers, encouraging the integration of LLMs into legal and financial NLP systems for future research.

1 Introduction

Building NLP systems for answering questions in the legal and financial domains could save time and resources, ensure compliance, and enhance the overall accuracy and effectiveness of legal and financial operations (Nay et al., 2023; Yang et al., 2023). Applying QA systems to such domains poses unique challenges. These domains feature complex jargon, nuanced phrasing, and contextual dependencies that require specialized knowledge and expertise (Katz et al., 2023; Wu et al., 2023a). A system tailored to these domains should be able to efficiently process and analyze large volumes of legal, financial, or regulatory documents, extracting relevant insights and answering targeted queries.

Large language models (LLMs) have shown impressive performance on several NLP tasks (Zhao et al., 2023). de Padua et al. (2023) show that LLMs trained on large amounts of data are able to obtain the necessary domain knowledge through in-context learning (ICL) (Brown et al., 2020). However, a major limitation of LLMs is the limit on the input size. There are many attempts to address this limitation (Press et al., 2022; Haviv et al., 2022; Zhu et al., 2023) and multiple transformer models are able to handle longer contexts (OpenAI, 2023; Rozière et al., 2023; Dai et al., 2019; Sun et al., 2023b). However, (Liu et al., 2023) show that model performance on certain parts of the input decreases with input size. Further, the cost and latency of LLMs increases with the input size.

The context required for legal and financial questions is often large and may not fit within the token limit, requiring more efficient retrieval. Financial and legal documents are often semi-structured. For example, Figure 1 shows a section from the US Internal Revenue Code. The text is organized into subsections, paragraphs and bullet points, which we leverage for better information retrieval. Further, financial reports often contain quantities in tabular format. We exploit these structures in a prompting approach that incorporates retrieval to workaround the context token limit.

We evaluate the proposed method on two datasets: FinQA (Chen et al., 2021) and SARA (Holzenberger et al., 2020). These datasets feature complex questions which require multiple steps of reasoning and arithmetic computations, which is challenging for language systems. We adopt chain-of-thought (CoT) prompting (Wei et al., 2023) for generating the answers since it is well suited for performing reasoning in a step-by-step manner. A chain of thought is a coherent sequence of reasoning steps that lead to the correct answer in a step-by-step manner. Providing examples of question-answer pairs along with their CoTs prepended to the test question causes GPT-3 to likewise output a CoT along with the answer for the test question, and improve its overall reasoning accuracy. CoT prompting is especially useful for complex tasks which require multiple steps of reasoning over the given input.

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 178–191

§7703. Determination of marital status	s7703
(a) General rule	s7703(a)
(1) the determination of whether an individual is married shall be made as of the close of his taxable year …	s7703(a)(1)
(2) an individual legally separated from his spouse under a decree of divorce or of separate maintenance …	s7703(a)(2)
(b) Certain married individuals living apart	s7703(b)
For purposes of those provisions of this title which refer to this subsection, if-	s7703(b)
(1) an individual who is married (within the meaning of subsection (a)) and who files a separate return …	s7703(b)(1)
(2) such individual furnishes over one-half of the cost of maintaining such household during the year, and	s7703(b)(2)
(3) during the last 6 months of the taxable year, such individual's spouse is not a member of such household,	s7703(b)(3)
such individual shall not be considered as married.	s7704(b)

Figure 1: An example of a statute from US Internal Revenue Code (left) and the subsection name assigned to each sentence after parsing as described in section 4.1.1 (right).

The results demonstrate that this simple and efficient approach outperforms state-of-the-art models in these domains. Training LLMs on financial and legal data may not be feasible as they may contain sensitive information. The use of ICL circumvents the problem and avoids expensive and tedious process of data collection and training. This makes the proposed approach a practical solution in scenarios where labeled data is limited or expensive to obtain. Additionally, CoT prompting offers the advantage of generating explanations and facilitating interpretability in critical domains where it is a key obstacle for the adoption of AI systems (Danilevsky et al., 2020). However, a major drawback of the approach is that it is task-specific. In particular, the retrieval relies on the structure within in the data and needs to be adapted to data from different sources[1].

We hope our work fosters research on coupling LLMs with retrieval in domains such as finance and law, where the ability to extract insights and answer questions about vast amounts of domain-specific data has many practical applications.

2 Related work

Previous work has proposed training specialized LLMs for financial and legal domains (Wu et al., 2023a; Huang et al., 2023; Nguyen, 2023; Yang et al., 2023). However, doing so requires a large amount of data, compute and cost.

Sun et al. (2023a) evaluate GPT-2 on FinQA (Chen et al., 2021). Blair-Stanek et al. (2023) evaluate GPT-3 with different prompting techniques on SARA where the context includes all the sections from the statutes. Since the input size of GPT-3 is limited, the prompts only included a subset of sections, which may not contain the required information. Further, fewer in-context examples were used for CoT as compared to few-shot learning. Li et al. (2023) and Wu et al. 2023b observe better performance with more in-context examples.

Nay et al. (2023) test various GPT models with ICL to answer multiple choice questions over tax laws. A retriever augmented setting is tested where a dense passage retriver, GTR (Ni et al., 2021), retrieves the top 4 relevant sections to the questions. Since entire sections are passed to the LLMs, the text has to be truncated.

This study extends past work by complementing LLMs with a retriever that extracts the relevant text from within the statutes, allowing for larger contexts and more in-context examples in the prompt.

3 Data

We use two datasets containing questions that involve multi-step logical and arithmetic reasoning from the legal and financial domains respectively.

3.1 SARA

StAtutory Reasoning Assessment dataset (SARA) (Holzenberger et al., 2020) is designed to evaluate statutory reasoning over a set of sections extracted from the US Internal Revenue Code (IRC). For each of the subsections contained in the selected sections, there are two hand-written case scenarios. Correctly solving these cases requires multiple steps of arithmetic as well as logical reasoning. For instance, some cases require computing the amount of tax owed according to a given section, only if the section applies to the given case. Thus, this dataset serves as a challenging task for an AI system, requiring domain expertise and reasoning abilities.

3.2 FinQA

FinQA (Chen et al., 2021) is a financial QA dataset. It comprises of 8,281 examples where each question is accompanied by a financial report, containing text as well as a table. The report contains the necessary information to correctly answer the question. FinQA poses many challenges for a QA system. The questions require retrieval, arithmetic and logical reasoning simultaneously over tables

[1]The code for this work is publicly available at github.com/vaibhavg152/Retrieval-Augmented-Chain-of-Thought-in-Semi-structured-Domains

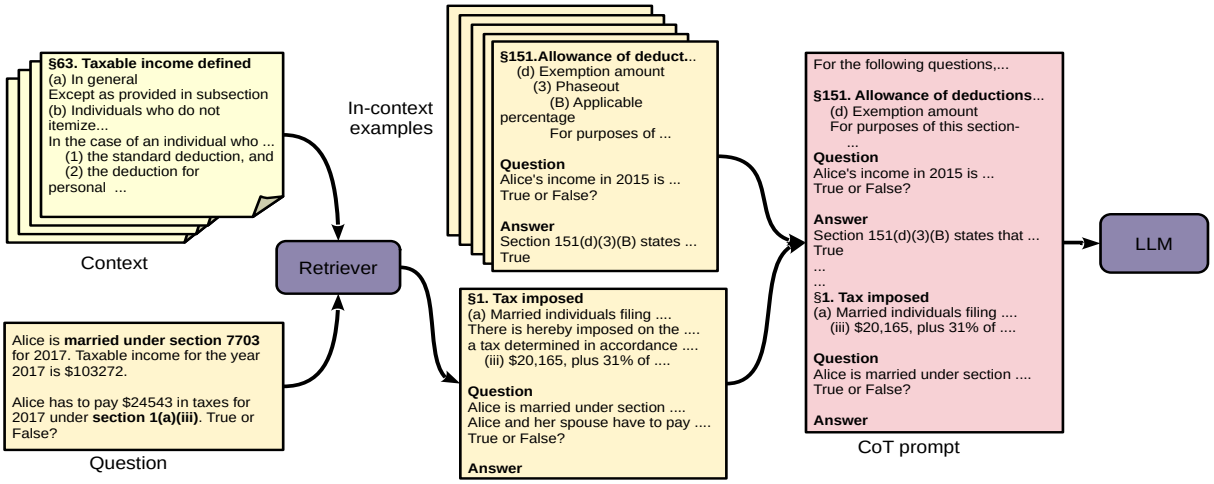

Figure 2: An overview of the proposed system on a sample input from SARA. The retriever extracts the relevant information from the context and combines it with the question. In-context examples are appended with the retrieval output to construct a prompt which is used for querying LLMs to generate an answer along the chain-of-thought.

and text. The questions also require an understanding of financial jargon. Finally, multiple reasoning steps are required to derive the answer.

4 Methodology

Figure 2 shows an overview of our proposed approach, which consists of two main components: retrieval and answering. The retrieval step involves filtering paragraphs from text and rows from tables that are relevant to the question. The retrieved information is then passed to the answering model.

4.1 Retrieval

Retrieval is essential for fully leveraging ICL and CoT reasoning abilities of LLMs. It can help to prevent the required context from exceeding the token limit, while also allowing the prompt to include enough in-context examples along with their CoT explanations. It can also help to reduce the time and cost of inference.

We propose to leverage the structure present in the data to retrieve the relevant context from the legal statutes and financial reports. This structure is specific to the data source and the retriever needs to be designed accordingly. In our analysis, we explore datasets with two different sources: SARA, where a template-based algorithm can be used for effective retrieval; and FinQA, where a more sophisticated pre-trained retrieval model is required.

4.1.1 SARA

As shown in Figure 1, the statutes in SARA are organized in a hierarchical structure with sections, sub-sections, paragraphs, and bullets. This hierarchical structure offers valuable information for efficient and accurate retrieval.

Figure 3 in the Appendix shows an example of a question from SARA. The questions contain references to the specific sub-sections they pertain to. Firstly, a simple regular expression-based extractor scans the question text to identify the relevant section name.

Next, a rule-based statute parser extracts the mentioned sub-section. The parser reads each sentence in the given statutes and assigns it to the most specific sub-section to which the sentence belongs. Figure 1 shows an example of a parsed statute section. We explore three retrieval strategies:

1. `mentioned-only`: The retriever returns all the sentences that are assigned to sub-sections containing the queried sub-section as a prefix. For Figure 1, a query for sub-section $7703(a)(1)$ will result in sentences assigned to $s7703$, $s7703(a)$ and $s7703(a)(1)$.
2. `entire-section`: Retriever returns the entire sub-section. In Figure 1, a query for sub-section $7703(a)(1)$ will result in sentences assigned to $s7703$, $s7703(a)$, $s7703(a)(1)$, as well as $s7703(a)(2)$.
3. `references`: Retriever returns sub-sections mentioned in the question along with those that are referenced in these retrieved subsections[2].

4.1.2 FinQA

The absence of a hierarchical structure in FinQA reports makes it impractical to adopt a rule-based approach for retrieval. Chen et al. (2021) convert the tables into text and then use BERT for retrieving relevant sentences from the report.

[2]It is intuitive to consider an approach that recursively retrieves text from sections mentioned in the sections retrieved in the previous step. However, this recursive approach proves to be impractical as it generates excessively large contexts.

However, using templates to convert tables into text leads to very long contexts. These templates can also introduce grammatical and logical errors, leading to a loss in the performance of the answering module. Thus, we use a tabular format during the answering step in order to exploit the structure (see Figure 5 in Appendix).

We also evaluate the system with gold retrieved sentences (GPT3-Gold, LLaMA2-70B-Gold).

4.2 Answering

In this study, we test GPT-3 (`text-davinci-003`) (Brown et al., 2020) and LLaMA-2 (Touvron et al., 2023) to answer the queries. We experiment with different prompting techniques, namely *zero-shot*, *few-shot* and *chain-of-thought* prompting. CoT prompting has been shown to improve the ICL abilities of sufficiently large LLMs (Zhao et al., 2023) and is especially useful for tasks that require multiple steps of reasoning.

In the zero-shot setting, the model is given the retrieved context and the question and is expected to output just the answer without any explanation.

In the few-shot setting, we further include in-context examples of question-answer pairs (8 examples for SARA and 12 examples for FinQA[3]).

In the CoT setting, we use the same in-context examples as used for the few-shot setting but each example also includes a CoT explanation. These explanations are manually written for each example. The model is expected to generate the answer along with the CoT explanations for the test cases.

For all questions in a dataset, we use the same prompt containing the same in-context examples which are selected using prompt tuning as described in Appendix section A.1.

Figures 4 and 5 in the Appendix show the CoT prompts used for SARA and FinQA respectively.

5 Experimental setup

5.1 Evaluation

For SARA, the task is formulated as an entailment task and is evaluated as a binary classification task.

For FinQA, Chen et al. (2021) propose **program accuracy** where the model is expected to generate a 'program' along with the answer. A program is a sequence of mathematical operations that leads to the final answer. The evaluation thus compares the output program with the gold standard program and checks if the two evaluate to the same answer.

We also measure the **answer accuracy** by ignoring errors only in units, prefix, suffix, precision digits or rounding errors.

[3]Tabular data in FinQA leads to shorter retrieved context and allows more examples per prompt.

Model name	Accuracy
Majority baseline [11]	50.0 ± 8.22
Feed-forward [11]	54.0 ± 8.20
Legal-BERT [11]	49.0 ± 8.22
BERT [12]	59.0 ± 8.09
GPT-3 (0-shot) [3]	71.0 ± 7.46
GPT-3 (CoT) [3]	57.0 ± 8.14
GPT-3 (dynamic) [3]	60.0 ± 8.06
GPT-3 + Ret	$\mathbf{81.6 \pm 4.22}$
LLaMA2-7B + Ret	53.5 ± 5.43
LLaMA2-7B_chat + Ret	54.4 ± 5.43
LLaMA2-13B + Ret	57.5 ± 5.39
LLaMA2-13B_chat + Ret	66.7 ± 5.13
LLaMA2-70B + Ret	71.1 ± 4.94

Table 1: Comparison of proposed system's performance on SARA with the existing baselines. The top section shows non-LLM based methods. The middle section shows the evaluation results from Blair-Stanek et al. (2023). The bottom section shows the results of our proposed system with 'Ret' representing the proposed retrieval. Results are shown with the 90% confidence interval.

5.2 Comparison with existing methods

Tables 1 and 2 show results on SARA and FinQA respectively[4]. Descriptions of the baselines are provided in Appendix Section B.

On SARA, both GPT-3 and LLaMA2-70B surpass the existing methods by a significant margin. We also observe the expected trend of the performance improving with the increase in the model size, with GPT-3 (175B) performing significantly better with LLaMA-2 models (Kaplan et al., 2020)[5].

On the other hand, the performance on FinQA with GPT-3 is comparable with baselines in terms of program accuracy but lags behind in answer accuracy. We believe this behavior is due to arithmetic errors made by LLMs (Qian et al., 2023), resulting in cases with correct programs but incorrect answers. Our approach with LLaMA2-13B/70B and GPT-3 outperforms general crowd workers

[4]For testing on FinQA, we randomly sample 200 examples from the public test set due to the high cost of LLM queries.

[5]Although Touvron et al. (2023) report similar performance of LLaMA-2 to GPT-3 on benchmark datasets, the task addressed here is domain-specific and requires more complex mathematical and logical reasoning than the benchmarks they use for evaluation.

Model	Program acc	Answer acc
Longformer [2]	21.90 ± 2.01	20.48 ± 1.96
ELASTIC [35]	57.54 ± 2.40	62.16 ± 2.36
DyRRen [17]	61.29 ± 2.37	63.30 ± 2.34
TabT5 [1]	**68.00 ± 2.27**	**70.79 ± 2.21**
APOLLO [28]	65.60 ± 2.31	67.99 ± 2.27
FinQANet-BERT [6]	58.86 ± 2.39	61.24 ± 2.37
GPT-3-BERT	**68.00 ± 5.43**	52.50 ± 5.81
LLaMA2-7B-BERT	25.50 ± 5.07	16.50 ± 4.32
LLaMA2-7B_chat-BERT	34.50 ± 5.53	14.50 ± 4.10
LLaMA2-13B-BERT	52.50 ± 5.81	33.00 ± 5.47
LLaMA2-13B_chat-BERT	50.50 ± 5.82	26.50 ± 5.13
LLaMA2-70B-BERT	60.50 ± 5.69	51.00 ± 5.81
Human non-expert [6]	48.17 ± 2.43	50.68 ± 2.43
Human expert [6]	87.49 ± 1.61	91.16 ± 1.38
FinQANet-Gold [6]	68.76 ± 2.25	**70.00 ± 2.23**
GPT-3-Gold	**72.50 ± 5.19**	56.50 ± 5.77
LLaMA2-70B-Gold	63.00 ± 5.62	54.50 ± 5.79

Table 2: Comparison with state of the art and baselines methods on FinQA. Results are presented with a 90% confidence interval.

Retrieval strategy	Accuracy
`entire-section`	52.50 ± 12.99
`references`	75.00 ± 11.26
`mentioned-only`	**77.50 ± 10.86**

Table 3: Comparison of the three retrieval strategies used with GPT-3 on SARA validation set.

who lack domain expertise in finance, whereas it falls short compared to financial experts.

The bottom section of Table 2 highlights the effectiveness of GPT-3 over FinQANet (Chen et al., 2021) when provided with the gold retrieved results. However, LLaMA-2 shows sub-par performance.

5.3 Ablation studies

Comparison of prompting techniques: Table 4 in the Appendix shows the evaluation results for zero-shot, few-shot and CoT prompting. CoT prompting leads to significantly better results across all models.

Comparing retrieval strategies: As outlined in section 4, we test three different retrieval strategies for SARA. Table 3 reveals that `mentioned-only` and `references` perform significantly better than `entire-section`. The questions in SARA are designed in a way where additional context apart from the mentioned sub-sections is not required. The difference in accuracy indicates the benefit of more targeted retrieval for model performance, since over-retrieval may dilute the signal provided by more directly relevant context.

Case analysis We perform manual qualitative inspection of the generated CoT explanations and report the analysis in Appendix section D.2.

6 Discussion

This study aims to utilize LLMs for challenging domain-specific QA tasks by using ICL along with retrieval techniques that leverage the semi-structured nature of financial and legal data. The proposed approach is simple and performs well compared to existing systems. It exploits ICL which avoids the costly and time-consuming processes of data collection and training. Since the proposed system produces a chain-of-thought with each output, it is easily interpretable and errors can be identified and rectified by human supervision (Danilevsky et al., 2020).

We hope this work will encourage researchers to delve deeper into the analysis and development of LLM-integrated NLP systems and retrieval-augmented LLMs.

7 Limitations

The retrieval algorithms in our study are specifically tailored to each dataset. Despite good reasoning abilities, the evaluation reveals that arithmetic errors are common. Further, inference with LLMs can be costly with latency higher than traditional approaches, making it sub-optimal for handling large volumes of data efficiently.

These limitations point to interesting future directions such as using arithmetic tools as plug-ins (Schick et al., 2023) for better performance and more generalizable retrieval algorithms. Further, several domain-specific LLMs can be tested (Huang et al., 2023; Yang et al., 2023; Wu et al., 2023a).

8 Acknowledgements

We sincerely thank the anonymous reviewers for their valuable feedback. For the experiments with LLaMA2, we thank the NYU IT High Performance Computing resources, services, and staff expertise.

References

Ewa Andrejczuk, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, and Yasemin Altun. 2022. Table-to-text generation and pre-training with tabt5.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning?

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Raul Salles de Padua, Imran Qureshi, and Mustafa U. Karakaplan. 2023. Gpt-3 models are few-shot financial reasoners.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2742–2758, Online. Association for Computational Linguistics.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. 2023. Natural language processing in the legal domain.

Mukai Li, Shansan Gong, Jiangtao Feng, Yiheng Xu, Jun Zhang, Zhiyong Wu, and Lingpeng Kong. 2023. In-context learning with many demonstration examples.

Xiao Li, Yin Zhu, Sichen Liu, Jiangzhou Ju, Yuzhong Qu, and Gong Cheng. 2022. Dyrren: A dynamic retriever-reranker-generator model for numerical reasoning over tabular and textual data.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2023. Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9298, Toronto, Canada. Association for Computational Linguistics.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Jiashuo Sun, Hang Zhang, Chen Lin, Yeyun Gong, Jian Guo, and Nan Duan. 2023a. Apollo: An optimized training approach for long-form numerical reasoning.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2023b. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14590–14604, Toronto, Canada. Association for Computational Linguistics.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

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

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023b. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

Jiaxin Zhang and Yashar Moshfeghi. 2022. Elastic: Numerical reasoning with adaptive symbolic compiler.

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, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training.

A Appendix

A.1 Prompt tuning

We iteratively refine the prompt using the validation sets of 40 samples for each dataset, with the aim of finding a prompt that encompasses a diverse range of cases while avoiding an overabundance of trivial or similar examples.

B Baselines

On SARA, we evaluate our system against the following baselines:

- **Majority baseline:** A trivial baseline that predicts the majority class for all the samples.
- **Feed-forward:** The feed-forward networks evaluated by Holzenberger et al. (2020).
- **Legal-BERT:** A BERT model trained specifically on legal domain (Chalkidis et al., 2020) and adapted for SARA by Holzenberger et al. (2020).
- **BERT:** A BERT model adopted for SARA by Holzenberger and Van Durme (2021)
- **GPT-3 (0-shot):** GPT-3 evaluated with a 0-shot prompt and without retrieval (Blair-Stanek et al., 2023).

- **GPT-3 (CoT):** GPT-3 evaluated with a CoT prompt and without retrieval (Blair-Stanek et al., 2023).
- **GPT-3 (dynamic):** GPT-3 evaluated with a dynamic few-shot prompt and without retrieval(Blair-Stanek et al., 2023). The prompt includes different in-context examples for different questions.

On FinQA, we compare our system with the following baselines in Table 2:

- **Pre-trained Longformer:** Longformer (Beltagy et al., 2020) is a model designed to take long input documents in one step. The model can be seen as a representative of one-step approaches.
- **FinQANet:** (Chen et al., 2021) use an LSTM decoder with attention for implementing the program generator and different models of BERT for retrieval.
- **ELASTIC:** (Zhang and Moshfeghi, 2022) use an adaptive symbolic compiler to generate the program.
- **DyRRen:** (Li et al., 2022) employ dynamic reranking of retrieved facts in every step.
- **TabT5:** (Andrejczuk et al., 2022) use a T5 model pre-trained on the Wikipedia tables.
- **APOLLO:** (Sun et al., 2023a) The retriever is based on the sequence-pair classification following (Nogueira and Cho, 2020). The program generator leverages a BERT encoder and an LSTM decoder with attention mechanism along with consistency-based reinforcement learning.

Text
Alice and Bob got married on April 5th, 2012.
Bob died September 16th, 2017.

Question
Section 7703(a)(1) applies to Alice for the year 2012.
Answer
Entailment

Figure 3: A case in SARA for section 7703.

Model	Zero-shot	Few-shot	CoT
LLaMA2-7B	50.4 ± 5.4	58.8 ± 5.3	53.5 ± 5.4
LLaMA2-7B_chat	42.1 ± 5.3	51.8 ± 5.4	54.4 ± 5.4
LLaMA2-13B	43.9 ± 5.4	53.1 ± 5.4	57.5 ± 5.4
LLaMA2-13B_chat	60.1 ± 5.3	53.9 ± 5.4	66.7 ± 5.1
LLaMA2-70B	49.6 ± 5.4	67.5 ± 5.1	71.1 ± 4.9
GPT-3	64.9 ± 5.2	74.6 ± 4.7	81.6 ± 4.2

Table 4: Comparison of different prompting techniques used on SARA. The prompts used are as described in section 4.2 and shown in Appendix section E. Note that the results shown here are from our proposed retrieval-augmented method, and not the same as the baselines shown in Table 1, which come from Blair-Stanek et al. (2023).

	Correct CoT	Incorrect CoT
Correct ans	23	8
Incorrect ans	3	6

Table 5: Results of the manual analysis performed on the validation set using GPT-3.

C Data examples

Figure 3 shows a question from SARA.

D Ablation studies

D.1 Zero-shot, few-shot and CoT prompting

Table 4 shows the performance of different models with different prompting techniques.

D.2 Case analysis

SARA We conducted a manual analysis of the model's output on the validation set. Table 5 presents the results of this analysis on SARA, indicating the number of examples where both the answer and the chain-of-thought reasoning provided by the model were correct, both were incorrect and cases where one of them was incorrect. We found that in 58.5% of the examples, the model accurately predicted both the output and the reasoning. For the remaining cases, we categorized the errors into four distinct categories, shown in Table 6.

Reasoning Error type	# of cases
Arithmetic errors	8
Logical errors	6
Context too long	2
Retrieval error	1

Table 6: Error analysis on the validation set of SARA.

Error type	# of cases
Arithmetic error	3
Logical error	2
Annotation error	2
Retrieval error	3

Table 7: Error analysis of the incorrect examples on 40 samples from FinQA.

FinQA On the constructed validation set comprising 40 samples, we observe that 30 samples have correct answers as well as corresponding programs. For the remaining 10 samples, we manually classify the errors into different categories, as shown in Table 7.

E Prompts

Figures 4 and 5 show the prompts used for SARA and FinQA respectively.

For the following questions, answer true or false based on the given context. For example:

§151. Allowance of deductions for personal exemptions
(d) Exemption amount
For purposes of this section-
(3) Phaseout
(B) Applicable percentage
For purposes of subparagraph (A), the term "applicable percentage" means 2 percentage points for each $2,500 (or fraction thereof) by which the taxpayer's adjusted gross income for the taxable year exceeds the applicable amount in effect under section 68(b). In the case of a married individual filing a separate return, the preceding sentence shall be applied by substituting "$1,250" for "$2,500". In no event shall the applicable percentage exceed 100 percent.

Question
Alice's income in 2015 is $276932. Alice is not married. The applicable amount according to section 68(b) is $250000.
Under section 151(d)(3)(B), the applicable percentage for Alice for 2015 is equal to 22. True or False?

Answer
Section 151(d)(3)(B) states that the applicable percentage means 2 percentage points for each $2,500 (or fraction thereof) by which the adjusted gross income exceeds the applicable amount in effect under section 68(b). The question states that the applicable amount according to section 68(b) is $250000 and Alice's income is $276932 for the year 2015. Therefore, the amount by which the income exceeds the applicable amount according to section 68(b) is 276932-250000 = 26932. There are 26932/2500 = 10.78 multiples of 2500 in 26932. 2 percentage points for each multiple of $2,500 is 2*10.78 = 21.56 which rounds to 22. Thus, the applicable percentage is equal to 22 and the hypothesis is correct.
True

§152. Dependent defined
(c) Qualifying child
For purposes of this section-
(1) In general
The term "qualifying child" means, with respect to any taxpayer for any taxable year, an individual-
(B) who has the same principal place of abode as the taxpayer for more than one-half of such taxable year,

Question
Alice has a son, Bob. From September 1st, 2015 to November 3rd, 2019, Alice and Bob lived in the same home.
Section 152(c)(1)(B) applies to Bob with Alice as the taxpayer for the year 2016. True or False?

Answer
Section 152(c)(1)(B) only applies to someone who has the same principal place of abode as the taxpayer for more than one-half of the taxable year. The question states that Bob is Alice's son who lived in the same house from September 1st, 2015 to November 3rd 2019. The whole of 2016 falls within September 1st, 2015 to November 3rd 2019. Therefore, Bob had the same principal place of abode as Alice for more than one-half of 2016. Thus, section 152(c)(1)(B) applies to Bob with Alice as the taxpayer for the year 2016.
True

§3306. Definitions
(a) Employer
(3) Domestic service
In the case of domestic service in a private home, local college club, or local chapter of a college fraternity or sorority, the term "employer" means, with respect to any calendar year, any person who during the calendar year or the preceding calendar year paid wages in cash of $1,000 or more for such service.

Question
Alice has paid $3200 in cash to Bob for agricultural labor done from Feb 1st, 2017 to Sep 2nd, 2017. Bob has paid $4200 in cash to Alice for domestic service in his home, done from Apr 1st, 2017 to Sep 2nd, 2018.
Section 3306(a)(3) applies to Bob for the year 2018. True or False?

Answer
Section 3306(a)(3) states that in case of domestic service in a private home, the term "employer" means any person who during the preceding calendar year paid wages in cash of $1,000 or more for such service. The question states that Bob has paid $4200 in cash to Alice for domestic service in his home, done from Apr 1st, 2017 to Sep 2nd, 2018. 4200 is greater than 1000 and 2017 is the preceding year of 2018. Thus, section 3306(a)(3) applies to Bob for the year 2018.
True

§63. Taxable income defined
(f) Aged or blind additional amounts
(1) Additional amounts for the aged
The taxpayer shall be entitled to an additional amount of $600-
(A) for himself if he has attained age 65 before the close of his taxable year, and

Question
In 2017, Alice was paid $33200. Alice and Bob have been married since Feb 3rd, 2017. Alice was born March 2nd, 1950 and Bob was born March 3rd, 1955.
Section 63(f)(1)(A) applies to Alice in 2017. True or False?

Answer
Section 63(f)(1)(A) applies to a person only if he has attained age 65 before the close of his taxable year. Alice was born March 2nd, 1950. Alice turned 65 on March 2nd, 2015. March 2nd, 2015 is before the close of the taxable year 2017. Thus, section 63(f)(1)(A) applies to Alice in 2017.
True

§151. Allowance of deductions for personal exemptions
(d) Exemption amount
For purposes of this section-
(3) Phaseout
(B) Applicable percentage
For purposes of subparagraph (A), the term "applicable percentage" means 2 percentage points for each $2,500 (or fraction thereof) by which the taxpayer's adjusted gross income for the taxable year exceeds the applicable amount in effect under section 68(b). In the case of a married individual filing a separate return, the preceding sentence shall be applied by substituting "$1,250" for "$2,500". In no event shall the applicable percentage exceed 100 percent.

Question
Alice's income in 2015 is $395276. The applicable amount according to section 68(b) is $250000.
Under section 151(d)(3)(B), the applicable percentage for Alice for 2015 is equal to 118. True or False?

Answer
Section 151(d)(3)(B) states that in no event shall the applicable percentage exceed 100. 118 exceeds 100. Therefore, the applicable percentage cannot be equal to 118 and the hypothesis is incorrect.
False

§152. Dependent defined
(c) Qualifying child
For purposes of this section-
(1) In general
The term "qualifying child" means, with respect to any taxpayer for any taxable year, an individual-
(B) who has the same principal place of abode as the taxpayer for more than one-half of such taxable year,

Question
Alice has a son, Bob. From September 1st, 2015 to November 3rd, 2019, Alice and Bob lived in the same home.
Section 152(c)(1)(B) applies to Bob with Alice as the taxpayer for the year 2015. True or False?

Answer
Section 152(c)(1)(B) states that the qualifying child has the same principal place of abode as the taxpayer for more than one-half of such a taxable year. The question states that Bob is Alice's son who lived in the same house from September 1st, 2015 to November 3rd 2019. November 3rd 2019 is after the end of 2015. The time between September 1st,2015 and the end of 2015 is 4 months. Half a year is 6 months. 4 is less than 6. Therefore, Bob had the same principal place of abode as Alice for less than one-half of 2015. Thus, section 152(c)(1)(B) does not apply to Bob with Alice as the taxpayer for the year 2015.
False

§3306. Definitions
(b) Wages
For purposes of this chapter, the term "wages" means all remuneration for employment, including the cash value of all remuneration (including benefits) paid in any medium other than cash; except that such term shall not include-
(15) any payment made by an employer to a survivor or the estate of a former employee after the calendar year in which such employee died;

Question
Alice employed Bob for agricultural labor from Feb 1st, 2011 to November 19th, 2019. On November 25th, Bob died from a heart attack. On December 20th, 2019, Alice paid Charlie, Bob's surviving spouse, Bob's outstanding wages of $1200.
Section 3306(b)(15) applies to the payment that Alice made to Charlie in 2019. True or False?

Answer
Section 3306(b)(15) applies only after the calendar year in which the employee died. The question states that Alice employed Bob for agricultural labor from Feb 1st, 2011 to November 19th, 2019. Bob died on November 25th and on December 20th, Alice paid Charlie. Thus, the payment is made in the same calendar year in which Bob died. Thus, section 3306(b)(15) does not apply to the payment that Alice made to Charlie in 2019.
False

§1. Tax imposed
(b) Heads of households
There is hereby imposed on the taxable income of every head of a household (as defined in section 2(b)) a tax determined in accordance with the following:
(i) 15% of taxable income if the taxable income is not over $29,600;

Question
Alice is a head of household for the year 2017. Alice's taxable income for the year 2017 is $1172980.
Alice has to pay $442985 in taxes for the year 2017 under section 1(b)(i). True or False?

Answer
Section 1(b)(i) applies only if the taxable income is not over $29,600. The question states that Alice's taxable income for the year 2017 is $1172980. 1172980 is greater than 29600. Therefore section 1(b)(i) does not apply to Alice in 2017 and the amount of taxes to be paid by Alice is unknown.
False

§151. Allowance of deductions for personal exemptions
(a) Allowance of deductions
In the case of an individual, the exemptions provided by this section shall be allowed as deductions in computing taxable income.

Question
Alice's income in 2015 is $100000. She gets one exemption of $2000 for the year 2015 under section 151(c). Alice is not married.
Alice's total exemption for 2015 under section 151(a) is equal to $6000. True or False?

Answer

Figure 4: Chain-of-thought prompt for SARA for a sample. The complete prompt contains 8 in-context examples with CoT explanations followed by the question that the model is supposed to answer. The in-context examples and explanations remain the same for all questions in the dataset. The text highlighted in yellow are the CoT explanations that we hand-crafted, while the test question is shown in blue.

The balance as of December 31 2009 of $540 is 604. The balance as of December 31 2010 of $540 is $2063.

Question
What was the percentage change in warranty reserve between 2009 and 2010?

Answer
The change in reserve between 2009 and 2010 is equal to 2063 - 604. The percentage change is equal to 100 times the previous result divided by 604.

Program: subtract(2063, 604), divide(#0, 604)
Final answer: 242%

In fiscal 2018, net cash used for financing activities of $755.1 million consisted primarily of cash dividends paid to stockholders of $440.9 million and purchases of common stock of $195.1 million and net repayments of debt of $120.1 million.

Question
In 2018 what was the percent of the net cash used for financing activities used for the purpose of purchases of common stock?

Answer
The net cash used for financing activities was 755.1 million and the cash dividends paid for the purchases of common stock was 195.1 million. Therefore, the percent of the net cash used for financing activities used for the purpose of purchases of common stock is 100 times 195.1 divided by 755.1.

Program: divide(195.1, 755.1)
Final answer: 25.84%

The capital purchase obligations of payments due in total is 12068 million. The capital purchase obligations of payments due by a period of less than 1 year is 9689 million. The capital purchase obligations of payments due by a period of 1-3 years is 2266 million. The capital purchase obligations of payments due by a period of 3-5 years is 113 million. The capital purchase obligations of payments due by a period of more than 5 years is 2014 million. The total of all the payments due is $65947 million. The total of payments due by a period of less than 1 year is $14265 million. The total of payments due by a period of 1-3 years is $10432 million. The total of payments due by a period of 3-5 years is $10067 million. The total of payments due by a period of more than 5 years is $31183 million.

Question
What percentage of total contractual obligations do capital purchase obligations make up as of December 30, 2017?

Answer
Total of contractual obligations payments due is $65947 million. Total of capital purchase obligations is $12068 million. The percentage that $12068 million is of $65497 million is equal to 100*(12068/65947).

Program: divide(12068, 65947)
Final answer: 18%

The impact on net income for the year that ended on December 31, 2006 is $-3680 thousand. The impact on net income for the year that ended on December 31, 2005 is $-1016 thousand. The impact on net income for the year that ended on December 31, 2004 is $-403 thousand.

Question
What was the difference, in thousands, in impact on the net income due to compensation expense for stock options and restricted stock between 2004 and 2005?

Answer
The impact on the net income for the year 2004 was $-403 thousand and the impact on the net income for the year 2005 is $-1016 thousand. The difference, in thousands, is equal to 1016-403.

Program: subtract(1016, 403)
Final answer: 613

The following is a reconciliation of basic shares to diluted shares: The diluted weighted-average shares for the year that ended on December 31, 2014 is 172.8 million. The diluted weighted-average shares for the year that ended on December 31, 2013 is 158.7 million. The diluted weighted-average shares for the year that ended on December 31, 2012 is 145.8 million.

Question
What was the average, in millions, of weighted-average diluted shares from 2012-2014?

Answer
The average of weighted-average diluted shares from 2012-2014 in millions, is equal to the average of 172.8, 158.7 and 145.8 which is equal to (172.8 + 158.7 + 145.8)/3.

Program: add(172.8, 158.7), add(#0, 145.8), divide(#1, const_3)
Final answer: 158.8

The future minimum lease payments under noncancellable operating leases for office space in effect at December 31, 2008 are $8.8 million in 2009, $6.6 million in 2010, $3.0 million in 2011 , $1.8 million in 2012 and $1.1 million in 2013.

Question
What was the average future minimum lease payments under noncancellable operating leases for office space from 2009 to 2013 in millions.

Answer
The average of the future minimum lease payments under noncancellable operating leases for office space from 2009 to 2013 in millions, is the average of 8.8, 6.6, 3.0, 1.8 and 1.1 which is equal to (8.8+6.6+3.0+1.8+1.1)/5.

Program: add(8.8, 6.6), add(#0, 3.0), add(#1, 1.8), add(#2, 1.1), divide(#3, const_5)
Final answer: 4.26

The amount per share for the payment date of 2018 is $1.90 and the total amount is $262 million. On November 2, 2018, the board declared a cash dividend of $0.50 per share that was paid on January 25, 2019 to the stockholders of record on December 31, 2018, for an aggregate amount of $68 million.

Question
Considering the year 2018, what is the percentage of the cash dividend paid per share concerning the total amount paid per share?

Answer
The cash dividend that was paid to the stockholders of record for 2018 was $0.50 per share. The total amount paid for 2018 is $1.90 per share. The percentage that $0.50 is of $1.90 is equal to 100*(0.50/1.90).

Program: divide(0.50, 1.90)
Final answer: 26.31%

The additional collateral or termination payments for a one-notch downgrade, as of December 2014 is $1072 million. The additional collateral or termination payments for a one-notch downgrade, as of December 2013 is $911 million. The additional collateral or termination payments for a two-notch downgrade, as of December 2014 is $2815 million. The additional collateral or termination payments for a two-notch downgrade, as of December 2013 is $2989 million.

Question
What is the difference in the required additional collateral or termination payments for a two-notch downgrade and additional collateral or termination payments for a one-notch downgrade in millions in 2014?

Answer
The required additional collateral or termination payments for a two-notch downgrade in 2014 is $2815 million and for a one-notch downgrade in 2014 is $1072 million. The difference in millions is 2815-1072.

Program: subtract(2815, 1072)
Final answer: 1743

The proved undeveloped reserves as of December 31, 2012 of U.S. is $407. The proved undeveloped reserves as of December 31, 2012 of Canada is $433. The total proved undeveloped reserves as of December 31, 2012 is $840. Devon Energy Corporation and subsidiaries notes to consolidated financial statements 2013 (continued) proved undeveloped reserves. The following table presents the changes in Devon 2019s total proved undeveloped reserves during 2012 (in mmboe). On December 31, 2012, Devon had 840 mmboe of proved undeveloped reserves. This represents a 7 percent increase as compared to 2011 and represents 28 percent of its total proved reserves.

Question
What is the approximate total amount of proved reserves?

Answer
On December 31, 2012, Devon had 840 mmboe of proved undeveloped reserves which represents 28 percent of its total proved reserves. Therefore, the total proved reserves are 840 divided by 28% which is equal to 840*100/28.

Program: divide(const_100, 28), multiply(#0, 840)
Final answer: 2998.8

The interest payments in 2010, 2009 and 2008 totaled $189 million, $201 million and $228 million, respectively.

Question
What would 2011 interest payments be based on the rate of change in 2009 to 2010?

Answer
The rate of change of interest payments from 2009 to 2010 is equal to 189 divided by 201. Using the same rate of change, the interest payment of 2011 would be equal to 189 times 189 divided by 201.

Program: divide(189, 201), multiply(189, #0)
Final answer: 177.7

The revenue of the year that ended on December 31, 2015 is $7426 million. The revenue of the year that ended on December 31, 2014 is $7834 million. The revenue of the year that ended on December 31, 2013 is $7789 million. The trustees of the plan have certain rights to request that our U.K. working capital increased by $77 million from $809 million on December 31, 2014 to $886 million on December 31, 2015.

Question
What is the working capital turnover in 2015?

Answer
The working capital turnover is calculated by dividing the company's net annual sales by its average working capital. The working capital was $809 million in 2014 and $886 in 2015. Therefore, the average working capital is equal to (809 + 886)/2. The revenue of the year 2015 us $7426. Therefore, the working capital turnover is equal to $7426 million divided by (809 + 886)/2.

Program: add(809, 886), divide(#0, const_2), divide(7426, #1)
Final answer: 8.8

The jpmorgan chase on December 31, 2013 is $100.00. The jpmorgan chase on December 31, 2014 is $109.88. The jpmorgan chase on December 31, 2015 is $119.07. The jpmorgan chase on December 31, 2016 is $160.23 ; the jpmorgan chase on December 31, 2017 is $203.07. The jpmorgan chase on December 31, 2018 is $189.57. The kbw bank index on December 31, 2013 is 100.00. The kbw bank index on December 31, 2014 is $109.36. The kbw bank index on December 31, 2015 is $109.90. The kbw bank index on December 31, 2016 is $141.23. The kbw bank index on December 31, 2017 is $167.49. The kbw bank index on December 31, 2018 is $137.82.

Question
Did jpmorgan chase outperform the kbw bank index?

Answer
The jpmorgan chase index increased from $100.00 in 2013 to $189.57 in 2018, while the kbw bank index increased from $100.00 in 2013 to $137.82 in 2018. jpmorgan chase outperformed the kbw bank if 189.57 is greater than 137.82.

Program: greater(189.57, 137.82)
Final answer: yes

The following table reconciles cash provided by operating activities (gaap measure) to free cash flow (non-gaap measure) : millions of dollars 2008 2007 2006 .

millions of dollars | 2008 | 2007 | 2006 ||
cash provided by operating activities | $ 4070 | $ 3277 | $ 2880 ||
cash used in investing activities | -2764 (2764) | -2426 (2426) | -2042 (2042) ||
free cash flow | $ 825 | $ 487 | $ 516 ||

we plan to continue implementation of total safety culture (tsc) throughout our operations .

Question
What was the percent of the cash provided by operating activities?

Program:

Figure 5: Chain-of-thought prompt for FinQA for a sample. The complete prompt contains 12 in-context examples with CoT explanations followed by the question that the model is supposed to answer. The in-context examples and explanations remain the same for all questions in the dataset. The text highlighted in yellow are the CoT explanations that we hand-crafted, while the test question is shown in blue.