EconLogicQA: A Question-Answering Benchmark for Evaluating Large Language Models in Economic Sequential Reasoning

Yinzhu Quan[*]
Georgia Institute of Technology
Atlanta, GA 30332, USA
yquan9@gatech.edu

Zefang Liu[*]
Georgia Institute of Technology
Atlanta, GA 30332, USA
liuzefang@gatech.edu

Abstract

In this paper, we introduce EconLogicQA, a rigorous benchmark designed to assess the sequential reasoning capabilities of large language models (LLMs) within the intricate realms of economics, business, and supply chain management. Diverging from traditional benchmarks that predict subsequent events individually, EconLogicQA poses a more challenging task: it requires models to discern and sequence multiple interconnected events, capturing the complexity of economic logics. EconLogicQA comprises an array of multi-event scenarios derived from economic articles, which necessitate an insightful understanding of both temporal and logical event relationships. Through comprehensive evaluations, we exhibit that EconLogicQA effectively gauges a LLM's proficiency in navigating the sequential complexities inherent in economic contexts. We provide a detailed description of EconLogicQA dataset and shows the outcomes from evaluating the benchmark across various leading-edge LLMs, thereby offering a thorough perspective on their sequential reasoning potential in economic contexts. Our benchmark dataset is available at https://huggingface.co/datasets/yinzhu-quan/econ_logic_qa.

1 Introduction

Logical reasoning is a pivotal skill in many professional and academic domains, enabling individuals to make informed decisions by understanding relationships between sequential events or pieces of information. In practice, the reasoning capabilities of large language models (LLMs) are frequently utilized across various applications, yet their effectiveness in logical reasoning remains underexplored. Despite its importance, there is an evident gap in the literature regarding the capability of LLMs to perform logical reasoning at a high level. This paper addresses this gap by introducing EconLogicQA, a new benchmark designed to rigorously assess the logical reasoning capabilities of LLMs specifically within the contexts of economics, business, and supply chain management.

EconLogicQA distinguishes itself from existing benchmarks by challenging LLMs to not only identify but also logically sequence multiple interconnected events from realistic economic scenarios. This approach aims to reflect the intricate decision-making processes required in these fields, going beyond mere fact recall or simple event prediction. By focusing on the sequencing of events based on logical rather than simply chronological order, EconLogicQA inspects the LLMs' ability to engage with and understand the underlying mechanics of economic phenomena.

The benchmark utilizes a curated dataset derived from a wide range of business news articles, guiding GPT-4 to generate multi-choice questions that demand an intelligent understanding of logical connections. A rigorous human review process ensures the accuracy and appropriateness of the content, refining the dataset to enhance its practical value. Through comprehensive testing across various state-of-the-art LLMs, this paper not only demonstrates EconLogicQA's effectiveness in evaluating logical reasoning but also provides insights into the potential improvements and applications of LLMs in complex reasoning tasks.

Our contributions of this paper are as follows:

1. We propose a novel benchmark, EconLogicQA, which rigorously assesses LLMs' logical reasoning capabilities within economics, business, and supply chain management.
2. We utilize GPT-4 to generate questions and answers from business articles, ensuring high-quality, well-crafted multiple-choice questions through meticulous human review.
3. We conduct a comprehensive evaluation of both open and proprietary LLMs to assess

[*] These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2273–2282
November 12-16, 2024 ©2024 Association for Computational Linguistics

their performance on this benchmark.

2 Related Work

Sequential Reasoning Benchmarks. In the realm of assessing complex reasoning abilities, Jin et al. (2023) introduce the CLADDER dataset, exploring capacities of large language models (LLMs) for causal reasoning, differentiating itself by focusing on formal rules-based causal inference instead of the typical evaluation of commonsense causality (Talmor et al., 2018; Huang et al., 2019) in Natural Language Processing (NLP). Wang et al. (2023) present STEPS, a rigorous benchmark designed to assess models' understanding of action sequence order in sequential tasks such as cooking and manufacturing, which highlights challenges of current LLMs in performing order reasoning without specific tuning. In examining adjacent domains, Guha et al. (2024) launch LEGALBENCH, emerging as a notable benchmark that evaluates LLMs in legal reasoning, having been developed collaboratively with legal experts to cover various facets of practical and theoretical legal analysis. Yang et al. (2024) establish AQA-Bench, serving as an interactive benchmark that evaluates LLMs' sequential reasoning abilities across various algorithmic tasks, including Depth-First Search (DFS), Breadth-First Search (BFS), and binary search, by requiring models to dynamically interact with the task environment, and thereby uncovering notable performance disparities among different LLMs. Valmeekam et al. (2024) create PlanBench as an extensible benchmark focused on evaluating LLMs' planning and reasoning capabilities, particularly in action and change, where diverse scenarios are used from the automated planning community to discern between genuine planning abilities and mere retrieval from pre-trained knowledge.

Economic Benchmarks. In the finance domain, Shah et al. (2022) launch the Financial Language Understanding Evaluation (FLUE) benchmark alongside the Financial LANGuage (FLANG) model, offering a comprehensive suite of evaluations focused on economic and financial domains, significantly outperforming existing models on various NLP tasks. Hendrycks et al. (2020) compile the Massive Multitask Language Understanding (MMLU) benchmark of 57 diverse tasks, including the economics subject, designed to evaluate the multitask accuracy of language models, revealing that even the largest models still struggle with expert-level performance and have inconsistent accuracy across subjects. Lu et al. (2023) propose the BBT-CFLEB benchmark, supporting advanced understanding and generation tasks in the financial domain and fostering significant research and development in this specialized area. Zhang et al. (2023) present the FinEval, a specialized benchmark for assessing financial knowledge in LLMs, demonstrating significant potential through GPT-4's high performance across diverse prompt types. Van Patten (2023) introduce the EconQA, a novel dataset for assessing LLMs on multiple-choice economics questions, reveals that Chain-of-Thought reasoning improves performance, particularly in mathematical queries, while prompt variations have a moderate effect on accuracy.

3 EconLogicQA

In this section, we detail the dataset generation and human review processes for creating the EconLogicQA benchmark and provide illustrative examples from it.

3.1 Dataset Generation

To streamline the question-generation process and reduce the subjectivity, labor-intensiveness, and randomness of manual creation, we utilize the GPT-4 to automatically generate questions by extracting key points from news articles. We specifically select economics-related articles from the 2011 to 2022 news dataset available on Kaggle[1], which is under the CC0 Public Domain license. This cleaned dataset provides a comprehensive range of economic news articles, and we further narrow our focus to those categorized under business to align with our research scope in economics.

In the data generation process, instructional prompts are developed to guide GPT-4 in creating multi-choice questions that challenge models to logically sequence events within the framework of business-related scenarios. These questions start with a brief scenario description and involve four events that must be ordered based on their logical or chronological sequence rather than their appearances in the source articles. The selected events pertain to typical business or economic situations, necessitating a deep understanding of business practices and economic principles for accurate sequencing. The prompts specify that the generated content

[1] https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning/

should be original, concise, and crafted without referencing the original news articles and unnecessary details. Each question is designed to be completed independently, making it suitable for evaluation. The formatted output includes a scenario description followed by four choices labeled A, B, C, and D, concluding with the correct sequence and a brief explanation to ensure that the reasoning behind the sequence is clear and deducible solely from the information presented in the question and choices. This structure is intended to enhance comprehension and application of business concepts. See Appendix A Figure 1 for an example of GPT-4 response with the prompt.

3.2 Review Process

In order to maintain the integrity and quality of the dataset, human verification is incorporated into the workflow. This manual review is essential as some generated responses exhibit errors in the correct sequencing of events. Each question undergoes meticulous examination, and adjustments are made to ensure accuracy and clarity in the logical sequence provided. Furthermore, the dataset undergoes a rigorous review to identify and exclude sensitive news articles that could be inappropriate. In total, 204 questions are removed from the initial pool of 854 questions. The criteria for removal include scenarios that yield multiple valid sequences and instances where a logical sequence cannot be clearly established. This comprehensive vetting process significantly enhances the evaluation quality. The final dataset consists of **650 questions**, divided into training, validation, and test sets containing 390, 130, and 130 questions, respectively.

3.3 Dataset Examples

To provide a clear depiction of EconLogicQA's contents, we present two examples from the dataset in Table 1. The first example details a sequence of decisions by Costco to manage its chicken supply chain effectively, while the second outlines steps taken by the Federal Reserve to navigate fiscal challenges. These examples illustrate the dataset's primary objective: to evaluate the capability of large language models in sequencing economic events logically, not just chronologically. Each question is meticulously designed to challenge models to demonstrate their understanding of complex economic interactions and to apply logical reasoning within real-world business contexts.

EconLogicQA Example 1

Question: Costco, a major retailer, is taking steps to control its chicken supply chain to maintain the price of its popular rotisserie chickens. Arrange the following events in the logical sequence based on Costco's strategy.
Choices:
A. Costco recruits farmers to grow chickens for them.
B. Costco opens a $450 million poultry complex in Nebraska.
C. Costco faces supply challenges and cost pressures in the poultry industry.
D. Costco decides to keep its rotisserie chickens at $4.99 despite competitors increasing their prices.
Correct Order: D, C, A, B

EconLogicQA Example 2

Question: Arrange the following events related to the Federal Reserve's monetary policy in their correct order.
Choices:
A. The Federal Reserve announces the end of its pandemic-era stimulus and accelerates the roll-back of its asset purchases.
B. The Federal Reserve signals it would hike interest rates multiple times throughout 2022.
C. The Federal Reserve reduces its massive balance sheet.
D. The Federal Reserve raises interest rates due to inflation well above 2% and a strong labor market.
Correct Order: A, B, D, C

Table 1: Question and answer examples from EconLogicQA.

4 Experiments

This section outlines experiments with the EconLogicQA dataset, assessing the sequential reasoning of multiple open and proprietary large language models (LLMs) in economic scenarios.

4.1 Experiment Setup

We implement experiments on various LLMs using the EconLogicQA dataset to assess its sequential reasoning capabilities within the intricate realms of economics, business, and supply chain management. We select the current mainstream open and proprietary LLMs in our study, including Llama-2 (Touvron et al., 2023a,b), Llama 3 (Meta AI, 2024), Gemma (Team et al., 2024), Mistral (Jiang et al., 2023), Yi (Young et al., 2024), Zephyr (Tunstall et al., 2023), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (Achiam et al., 2023). Each model is evaluated in both 1-shot and 5-shot settings. We do not include 0-shot in this section because the results are unsatisfactory due to lack of examples. Therefore, we choose to using a few-shot approach for sorting problems. Accuracy is the primary metric used, offering a direct measure of each model's understanding of the concepts within the EconLogicQA dataset.

All experiments are conducted on NVIDIA A100 GPUs. Each open LLM use in this paper is sourced from the Huggingface Transformers library (Wolf et al., 2019). Language Model Evaluation Harness (Gao et al., 2023) is used to test open LLMs on sequential reasoning evaluation tasks. The YAML configuration file is used to set key parameters in our scenario, such as terminating outputs, extracting answers, evaluating results, and so on. LLMs are configured to stop generating responses once it begins to pose new questions after answering the question in the prompt, setting the temperature to 0 without sampling. Then, we extract the answer from the response generated by LLMs using regular expression. Finally, we verify the answer through exact matching and use accuracy as the evaluation metric.

4.2 Experiment Results

The results from our experiments shown in Table 2 indicate diverse performances among the models, with significant disparities observed in their ability to handle the sequencing of economic events. GPT-4-Turbo exhibits the highest accuracy, achieving 56.92% in the 1-shot scenario and 56.15% in the 5-shot scenario, making it the best-performing model in our tests. GPT-4 follows closely, demonstrating the second-highest performance with 55.38% in the 1-shot and 53.85% in the 5-shot settings. Remarkably, the 1-shot scenario generally results in better performance than the 5-shot scenario for these two models, which could be attributed to the models' ability to leverage their pre-trained knowledge effectively without the potential confusion introduced by additional context in the 5-shot scenario.

Open LLMs are evaluated, showing varied performances as outlined in Table 2. Notably, Llama-3-8B-Instruct and Llama-3.1-8B-Instruct demonstrates significant improvements when fine-tuned with instructions compared with Llama-3-8B and Llama-3.1-8B respectively, achieving 34.62% and 36.15% accuracy in the 1-shot setting and 37.69% and 36.92% in the 5-shot setting, which highlights the substantial impact of instruction tuning on enhancing performance for question answering tasks. Similarly, Mistral-7B-Instruct-v0.3 exhibits promising results, with the accuracy of 39.23% in the 1-shot setting and 35.38% in the 5-shot setting, underscoring its adaptability to complex reasoning tasks, though it still lags behind GPT-4's overall performance.

Model	**1-Shot**	**5-Shot**
Llama-2-7B	0.77%	1.54%
Llama-2-7B-Chat	9.23%	10.00%
Llama-2-13B	9.23%	1.54%
Llama-2-13B-Chat	14.62%	8.46%
Llama-3-8B	23.85%	23.85%
Llama-3-8B-Instruct	34.62%	37.69%
Llama-3.1-8B	31.54%	27.69%
Llama-3.1-8B-Instruct	36.15%	36.92%
Gemma-2B-IT	7.69%	7.69%
Gemma-1.1-2B-IT	8.46%	6.92%
Gemma-7B-IT	2.31%	4.62%
Gemma-1.1-7B-IT	0.77%	3.85%
Mistral-7B-v0.1	26.15%	30.00%
Mistral-7B-v0.2	26.15%	32.31%
Mistral-7B-v0.3	26.15%	33.08%
Mistral-7B-Instruct-v0.1	15.38%	20.77%
Mistral-7B-Instruct-v0.2	31.54%	32.31%
Mistral-7B-Instruct-v0.3	39.23%	35.38%
Yi-6B	3.85%	29.23%
Yi-6B-Chat	20.77%	30.77%
Zephyr-7B-Alpha	23.08%	23.08%
Zephyr-7B-Beta	17.69%	14.62%
GPT-3.5-Turbo	37.69%	38.46%
GPT-4	55.38%	53.85%
GPT-4-Turbo	**56.92%**	**56.15%**
GPT-4o	42.31%	45.38%
GPT-4o-mini	47.69%	46.15%

Table 2: Comparison of the accuracy of multiple large language models on the EconLogicQA dataset under 1-shot and 5-shot learning scenarios.

These experiments collectively demonstrate the varying degrees of proficiency in applying LLMs to economic sequential reasoning, reflecting the current landscape of LLMs capabilities in this domain. There is still a clear gap in the ability of current LLMs to accurately handle many economic scenarios and correctly sequence events, especially open LLMs. This limitation points to significant challenges that persist in the field, emphasizing the need for targeted improvements and innovations in future research. Addressing these shortcomings could lead to more robust models that are better equipped to navigate the complexity of economic reasoning.

4.3 Sensitivity Analyses

This section examines how changes in event order, the number of examples (shots), and the use of Chain-of-Thought (CoT) reasoning (Wei et al., 2022) influence the performance of GPT-4 and GPT-4-Turbo, revealing their varying sensitivities to these factors. As shown in Table 3, GPT-4 maintains steady performance even when event order is altered, especially in the 5-shot setting, while GPT-4-Turbo experiences a decline in accuracy when the order is permuted in the 1-shot scenario. This highlights GPT-4's robustness in maintaining accuracy under varying conditions, particularly when provided with more context.

Model	Permuted	1-Shot	5-Shot
GPT-4	✗	55.38%	53.85%
GPT-4	✓	52.31%	53.08%
GPT-4-Turbo	✗	56.92%	56.15%
GPT-4-Turbo	✓	47.69%	50.77%

Table 3: Accuracy of GPT-4 and GPT-4-Turbo on the EconLogicQA dataset with and without permuted order under 1-shot and 5-shot learning scenarios.

Table 4 and Table 5 further explore the impact of example quantity and CoT reasoning. Performance peaks at the 1-shot and 5-shot settings, with accuracy dropping as more than five examples are introduced, suggesting that excessive context may introduce confusion rather than clarity. CoT reasoning enhances performance in the 5-shot setting by promoting better coherence and logical flow, but in the 1-shot scenario, it has minimal impact, indicating that the model's pre-trained knowledge is often sufficient for simpler tasks without needing explicit reasoning chains.

5 Conclusion

This study introduces EconLogicQA, a novel benchmark specifically designed to assess the logical reasoning capabilities of large language models (LLMs) in the domains of economics, business, and supply chain management. The benchmark challenges LLMs with complex realistic economic scenarios. Utilizing GPT-4, high-quality, well-crafted multiple-choice questions are generated from business articles and refined through meticulous human review. A comprehensive evaluation of both open and proprietary LLMs is conducted, providing deep insights into their capabilities and limitations within this specialized context. The error analysis in Appendix B reveals that the models often struggle with sequencing strategic decisions, interpreting policy changes, and distinguishing between stages of implementation and development, indicating a need for improvement in handling complex economic reasoning.

Future improvements to LLMs in economic reasoning could include refining prompt engineering to enhance accuracy in complex scenarios and applying fine-tuning with the EconLogicQA dataset for optimized responses. Additionally, developing specialized LLMs designed for economics, business, and supply chain management offers a promising opportunity.

Model	# of Shots	Accuracy
GPT-4-Turbo	0	48.46%
GPT-4-Turbo	1	56.92%
GPT-4-Turbo	2	51.54%
GPT-4-Turbo	3	53.08%
GPT-4-Turbo	4	51.54%
GPT-4-Turbo	5	56.15%
GPT-4-Turbo	8	50.00%
GPT-4-Turbo	10	50.00%

Table 4: Accuracy of GPT-4-Turbo on the EconLogicQA dataset with different numbers of shots.

Model	CoT	1-Shot	5-Shot
GPT-4	✗	55.38%	53.85%
GPT-4	✓	55.38%	54.62%
GPT-4-Turbo	✗	56.92%	56.15%
GPT-4-Turbo	✓	55.38%	59.23%

Table 5: Accuracy of GPT-4 and GPT-4-Turbo on the EconLogicQA dataset with and without Chain-of-Thought (CoT) Reasoning under 1-shot and 5-shot learning scenarios.

Limitations

Scope of Data. The effectiveness of the EconLogicQA benchmark is currently validated using a specific dataset of economic news articles. This reliance on a single data source limits the generalizability of our findings to other datasets in the domain, which may have distinct characteristics and diverse compositions that could influence the performance of LLMs.

Temporal Coverage. The dataset spans articles from 2011 to 2022, potentially missing recent economic developments and trends. This temporal limitation could affect the benchmark's relevance and the models' performance in current economic contexts.

Ethical Considerations

Our research emphasizes transparency in methodology, reporting, and data utilization. We adhered to the principles of responsible AI research throughout the study. The data employed in this research is sourced from public domains, ensuring no private user data was involved. We incorporated a stringent human review process to maintain dataset accuracy and integrity, excluding any sensitive or inappropriate content. These measures reflect our commitment to ethical standards and research integrity.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Meta AI. 2024. Meta llama 3. Accessed: 2024-05-09.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Tate Van Patten. 2023. Evaluating domain specific llm performance within economics using the novel econqa dataset.

Weizhi Wang, Hong Wang, and Xifeng Yan. 2023. Steps: A benchmark for order reasoning in sequential tasks. *arXiv preprint arXiv:2306.04441*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Siwei Yang, Bingchen Zhao, and Cihang Xie. 2024. Aqa-bench: An interactive benchmark for evaluating llms' sequential reasoning ability. *arXiv preprint arXiv:2402.09404*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

A Prompt and Response Example

Figure 1 illustrates an example of the prompt used to generate the sorting question and the corresponding response from GPT-4.

B Error Analyses

To systematically understand the errors made by the LLMs, we implement a series of automated analyses using a set of specially designed prompts. These prompts guide GPT-4o in identifying and categorizing errors based on their types and underlying causes. The instruction prompts are divided into four parts: Overview, Provided Materials, Steps to Follow, and Error Types and Criteria. The Overview provides a general introduction, the Provided Materials section lists the necessary resources, and the Steps to Follow outlines the process for using the prompts. The Error Types and Criteria part details the specific error categories, which include cause and effect misinterpretation, strategic and decision-making confusion, implementation and development confusion, policy and regulatory misunderstanding, impact and outcome misinterpretation, and growth and expansion misinterpretation. Each error type is matched with specific indicators to ensure precise classification, enhancing the clarity and accuracy of the analysis. For more specific details, please refer to Appendix C.

According to Figure 2, the error analysis reveals that the most common error type is strategic and decision-making confusion, which accounts for 38.3% of the total errors. This indicates that the model often struggles with understanding the correct sequence of strategic decisions and their logical order. The second most frequent error type is policy and regulatory misunderstanding, making up 31.7% of errors, highlighting the model's difficulty in correctly interpreting the sequence and impact of policy changes. Implementation and development confusion constitutes 11.7% of the errors, reflecting challenges in distinguishing between various stages of project execution. Impact and outcome misinterpretation account for 10% of the errors, showing some issues in linking actions with their results. Lastly, growth and expansion misinterpretation, comprising 8.3% of the errors, indicates some misunderstanding in the sequence of growth-related activities. These insights suggest that while the model performs reasonably well in some areas, significant improvements are needed in understand-

Prompt:

Generate a concise multi-choice question without referencing the original news article, starting with a brief scenario description. Include four events from this scenario that need sequencing based on their logical connection, not just the order of appearance in the news. These events should be relevant to typical business or economic situations, requiring knowledge of business practices and economics for correct sequencing. Ensure a single correct sequence, free from ambiguity. The question should be crafted for independent completion, suitable for student evaluation. Create original, concise content for both the question and choices, avoiding plagiarism and limiting excessive details or numerical data. Format the output with the scenario description, followed by choices labeled A, B, C, D, and finish with the correct sequence and a brief explanation, ensuring the sequence logic is deducible solely from the question and choices presented.

Here is the output format:

```
Question: ...
Choices:
A. ...
B. ...
C. ...
D. ...
Correct Order: ...
Explanation: ...
```

Here is the business news article:

```
{news_article}
```

GPT-4 Response:

Question: Arrange the following events related to the development and launch of Apple's Siri voice assistant in their correct chronological order.

Choices:
A. Apple acquired Siri in April 2010.
B. Siri was spun out as a standalone business in 2008.
C. Siri was integrated into the iPhone 4S, which was announced to be sold in stores on October 14.
D. Work on Siri began in the summer of 2007 at SRI International.

Correct Order: D, B, A, C

Explanation: The development of Siri started in the summer of 2007 at SRI International. The project was then spun out as a standalone business in 2008. Apple acquired Siri in April 2010. Finally, Siri was integrated into the iPhone 4S, which was announced to be sold in stores on October 14.

Figure 1: The prompt structure for generating sorting questions in EconLogicQA. It involves filling the contents of news articles into prompt templates. We also give an example of GPT-4 response to a specific prompt constructed from this information.

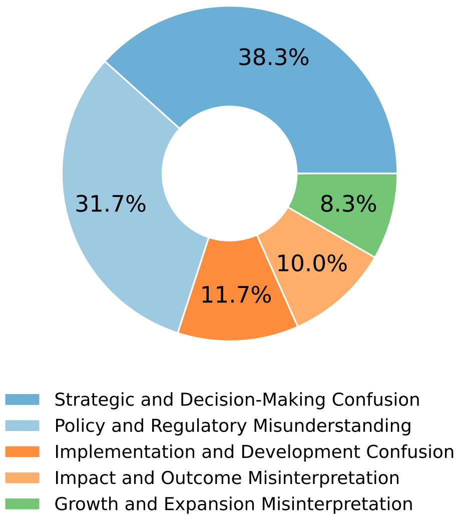

Figure 2: Error types of GPT-4 responses on the testing set.

ing strategic decisions and policy sequences.

C Instruction Sheet for Exam Review

Overview

This instruction sheet provides guidance on how to review and classify responses in an economics question set. The exam requires sorting choices correctly, and the provided responses contain various errors. The goal is to classify each response into its respective error type using the detailed criteria provided.

Provided Materials

1. Exam Questions and Choices: Each question includes multiple choices that need to be sorted in the correct sequence.
2. Correct Answers: The correct order for each question is provided.
3. Responses: The responses that need to be reviewed and classified according to the error types.

Steps to Follow

1. Read the Response: Carefully read the response to the question.
2. Identify Misunderstandings: Look for signs of misunderstanding or confusion in the response.
3. Match Indicators: Compare the mistakes to the indicators listed in each error type.
4. Classify the Error: Assign the mistake to the appropriate error type based on the matching indicators.
5. Provide Feedback: Offer specific feedback based on the identified error type to help understand and correct the mistake.

Error Types and Criteria

1. Cause and Effect Misinterpretation
   - Description: Errors where the relationships between causes and their effects, including immediate versus delayed reactions, are misunderstood.
   - Indicators:
     - An effect is placed before its cause.
     - Confusion between actions and their immediate or delayed outcomes.
   - Examples:
     - Misinterpreting which actions lead to specific results.
     - Placing the result of an action before the action itself.
2. Strategic and Decision-Making Confusion
   - Description: Errors involving misunderstanding strategic decisions and the correct sequence of these decisions.
   - Indicators:
     - Confusion about the order of strategic steps.
     - Misordering company actions or milestone events.
   - Examples:
     - Placing a company's strategic milestone after subsequent actions.
     - Misunderstanding the sequence of decision-making processes.
3. Implementation and Development Confusion
   - Description: Errors related to confusion between different stages of implementation, development, and testing processes.
   - Indicators:
     - Mixing up the order of development stages.
     - Placing testing phases incorrectly.
   - Examples:
     - Placing the implementation stage before the development stage.
     - Misordering feature rollouts and testing phases.
4. Policy and Regulatory Misunderstanding
   - Description: Errors involving misunderstandings of policy changes, regulatory actions, and economic indicators.
   - Indicators:
     - Misinterpreting the sequence of policy changes.

    - Confusion about the order of regulatory actions.
  - Examples:
    - Misplacing the impact of a policy change.
    - Misunderstanding the sequence of regulatory steps.

5. Impact and Outcome Misinterpretation
   - Description: Errors related to misinterpreting the impact or outcomes of actions and decisions.
   - Indicators:
     - Confusion between actions and their impacts.
     - Misplacing outcomes in the sequence of events.
   - Examples:
     - Placing an outcome before the action that caused it.
     - Misunderstanding the effect of a specific action or decision.
6. Growth and Expansion Misinterpretation
   - Description: Errors related to misunderstanding growth metrics and expansion steps.
   - Indicators:
     - Misordering steps related to company growth.
     - Confusion about the sequence of expansion activities.
   - Examples:
     - Placing growth milestones out of order.
     - Misunderstanding the sequence of steps for company expansion.

Example Application

Question: The Federal Reserve is considering the creation of a digital currency. What steps did they take in sequence?

Choices:

A: The Federal Reserve collects public feedback on the digital currency.

B: The Federal Reserve releases a discussion paper on the digital currency.

C: The Federal Reserve receives clear support from the public.

D: The Federal Reserve addresses major policy questions.

Correct Answer: B, A, C, D

Response: B, A, D, C

Error Type: Policy and Regulatory Misunderstanding

Indicators Matched: Misinterprets the sequence of policy changes (placing policy questions before public support).

Feedback: The response indicates a misunderstanding of the sequence in which the Federal Reserve addressed policy questions and gathered public support. It's important to understand that the Federal Reserve first gathers public feedback before addressing policy questions.