Spring 2023

Evaluating Domain Specific LLM Performance Within Economics Using the Novel EconQA Dataset

Tate Van Patten

Evaluating Domain Specific LLM Performance Within Economics Using the Novel EconQA Dataset

Author: Tate Van Patten

Western Washington University

Abstract: This paper describes a novel dataset, EconQA, constructed to assess the performance of large language models within multiple choice economics questions. I present results from 10 experiments, varying prompts and model choices. Results challenge previous findings that prompt choice makes a large impact on quality of response. Using the GPT 3.5 Turbo model, observed performance levels ranged from 70-77% for all prompt choices, with the no prompt baseline scoring 73%. When prompted to use Chain-of-Thought reasoning with examples, performance was highest at 76%. Contrary to previous research, performance on mathematical questions when prompted with Chain-of-Thought was high. This paper closes with an analysis of the types of questions the models performed best on and common errors.

Table of Contents

Evaluating Domain Specific LLM Performance Within Economics Using the Novel EconQA Dataset

Introduction

Large Language Models (LLMs) are rapidly entering mainstream adoption in a wide range of industries. In recent headlines, they have been shown to succeed by many academic measures, including AP exams, the Wharton MBA final exam, and even the Bar exam. These tests often use multiple choice questions for simplicity of grading. Multiple Choice question formats are an ideal choice to evaluate LLMs on accuracy because they are (1) easy to grade, (2) require fewer tokens than short response questions, reducing cost and time, (3) are easily understood and interpreted, and (4) are used in existing standardized settings for human evaluation, making comparisons to human performance simple and quantifiable. However, despite the tremendous potential of Multiple Choice Prompting (MCP) in evaluating LLMs, there has been little exploration of (1) relative strength for domain specific tasks within the social sciences, (2) best prompting practices to elicit strong responses, and (3) the quality of domain specific reasoning under different prompting strategies.

This research paper describes the creation of a novel dataset, EconQA, which is used to test and evaluate the performance of two language models across a broad array of topics within economics. I evaluate overall and question type accuracy across 11 experiments, testing and comparing each prompts performance.

Literature Review

Given the novelty of LLMs, there has been very little published research examining LLM performance on Multiple Choice Prompt (MCP) type questions. The majority of research in this area focuses on common sense reasoning and logical deduction tasks, using benchmarks like CommonsenseQA. Of the existing research evaluating MCP performance, to the best of the author's knowledge, none deeply examine domain specific knowledge within economics, nor any other social science. Existing research in domain specific knowledge is most dense in medicine and science, using benchmarks like PubMedQA or SciQ. OpenAI has published papers on standardized test performance, including performance on AP economics exams, however their sample size was quite small. Generally, research has found the "Chain-of-Thought" (CoT) prompting, which asks the model to 'reason' through each question before answering, sometimes with examples of correct reasoning. Given that this is a rapidly developing line of research, one should expect this body of literature to grow over the next few years. Some selected papers are summarized briefly below.

Chen, J., Chen, L., Huang, H., & Zhou, T. (2023, April 18). *When do you need Chain-of-Thought Prompting for ChatGPT?* ArXiv.org. https://doi.org/10.48550/arXiv.2304.03262

This paper analyzes the use of Chain-of-Thought (CoT) prompting for ChatGPT, a LLM finetuned for conversation. Finetuning is a term that refers to custom training a pretrained model for a particular dataset. CoT prompting is a strategy which appends a prompt such as "Let's think step by step" to an LLM query with hopes of improving response quality. They find that for some tasks which were improved by CoT prompting for previous non conversational models, ChatGPT sees no improvements using CoT. On some datasets, ChatGPT spontaneously generates CoT reasoning steps without requiring CoT prompting. For 4 out of 6 arithmetic datasets, the model performed worse when given a CoT prompt, which is counter to results for GPT 3. Most other non-arithmetic tasks see similar improvements from including a CoT prompt.

Hebenstreit, K., Praas, R., Kiesewetter, L. P., & Samwald, M. (2023, May 4). *An automatically discovered chain-of-thought prompt generalizes to novel models and datasets*. ArXiv.org. https://doi.org/10.48550/arXiv.2305.02897

This paper analyzes and compares 10 different CoT prompts using 6 LLM models on 6 datasets. They find the most robust prompt was "Let's work this out in a step by step way to be sure we have the right answer," a prompt designed by Zhou and inspired by Kojima.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022, May 24). Large Language Models are Zero-Shot Reasoners. *ArXiv:2205.11916 [Cs].* NeurIPS2022, New Orleans. https://arxiv.org/abs/2205.11916

This seminal paper argues zero-shot CoT prompting is often more desirable than few-shot CoT and sometimes yields better results. Zero-shot refers to giving zero examples while few-shot refers to giving one or more examples within a prompt. They find that performance on some tasks, especially 'system-2' tasks like arithmetic which require multi-step reasoning, is significantly improved by appending the phrase "Let's think step by step" to the end of each prompt. Zero-shot prompting differs from few-shot CoT prompting originally proposed by Wei since it does not require solved examples. They find that zero-shot CoT substantially out performs the baseline in 4 out of 6 datasets, with the largest improvement from 17.7% to 78.7% on the MultiArith benchmark. They do not find improvements on commonsense reasoning datasets like CommonSenseQA. They manually investigate reasoning chains for wrong answers, finding that a) reasoning chains are often defensible even when the end answer is not correct, b) sometimes the model chooses multiple answers when it is hard to eliminate, and c) the model will sometimes continue to output unnecessary steps of reasoning after getting the correct prediction.

OpenAI. (2023). GPT-4 Technical Report. In *Arxiv*. https://arxiv.org/pdf/2303.08774.pdf

This paper includes measures of OpenAI's GPT-4's performance on various multiple choice question exams. GPT-4 is a larger and more advanced model than any used in this paper. Of particular interest is performance on the AP Microeconomics and Macroeconomics exams. They found accuracy was 89.7% and 73.1% on each exam, respectively. These results are limited since they only tested on 30 questions, some of which were removed due to contamination concerns. However, these scores would place GPT-4 in the 82-100th percentile or the 84-100th percentile of human test takers, respectively. They used both few-shot prompting and asked for explanations, resulting in a large number of tokens being used for each question.

Robinson, Joshua, Christopher Michael Rytting, and David Wingate. "Leveraging Large Language Models for Multiple Choice Question Answering," 2023.

This paper finds that LLM performance is significantly higher on MCP tasks than Cloze Prompting (CP). Since most existing research evaluates LLMs based on their performance on CP tasks, the authors conclude that LLM capacities have been severely underestimated. MCP tasks are differentiated from cloze tasks in that multiple possible answers are presented as options in the prompt with an associated letter (A, B, C, D, etc.), whereas cloze tasks are more open-ended "fill in the blank" questions. The author also notes that MCP tasks are computationally cheaper, easier to evaluate, and subject to fewer other problems than CP tasks. They note that some models, like GPT3 Davinci, Instruct, and Codex's responses are less sensitive to question order (higher Proportion of Plurality Agreement-PPA), indicating they are better at binding the letter associated with a prompt to their responses, an important prerequisite to successfully complete MCP tasks. They test and find that models with higher PPA also have better accuracy on MCP tasks than on CP tasks. High PPA models also have better accuracy on MCP tasks than low PPA models.

Methodology

Construction of Dataset

Since no preformatted databases of economics questions exist, constructing a novel dataset was necessary. Throughout this paper, I refer to the final structured and formatted database as EconQA. For this, I sourced questions from the McConnell and Brue *Economics 16th edition* test bank. Because a machine readable version of this test bank is not publicly available online, concerns about dataset leakage are reduced. This dataset has the advantage that it is from a widely used textbook, so questions are likely to be similar to what students may encounter in a 200-level college economics exam.

The test bank was converted to text using Adobe Acrobat's built in OCR. I then used a variety of text processing tools in R to neatly structure the information into a usable dataset. I scraped 6 attributes for each question. These attributes were raw_text, question, type, pg_num, topic, and answer. These attributes were each included in the test bank for each question. raw_text is the raw full text for each question block, including all the other attributes. question is the extracted question text which is fed to the LLM. question is the text of the question. type is the type of question based on the format and/or skill required. Examples of types of questions include equation questions, and application of concept questions. A full table describing the coding for type is available in the appendix. Topic is a number which corresponds to a different learning objective for each chapter. For example, in the previous question, topic=1, corresponding to "Economics; economic perspective". Other examples of topics include "Logical pitfalls", and "Positive and normative statements". pg_num is the page number where the primary problem skill is introduced in the McConnel and Brue textbook. Answer refers to the letter of the correct answer. It is important to note this textbook covers both microeconomics and macroeconomics.

Great efforts were made to standardize formatting across questions. Early runs with poor formatting standardization yielded essentially random tokens as a response. This early issue was fixed by removing additional spaces, tabs, line breaks, and unnecessary special symbols. I removed most questions which were labeled 'graphical' or 'tabular' in type since the models available are not multi-modal and formatting problems prevent tabular inputs to the APIs I used. An error caused some graphical questions to remain in the dataset which require graphs which were not supplied. I also removed all questions which involved solving equations or other highly mathematical questions since mathematical shortcomings of LLMs are well documented and I was more interested in performance on more intuitive questions. The test bank originally contained around 6,000 questions, but after making the described omissions, the number of questions fell to 3,623. All 3,623 questions were evaluated for each of the experiments ran[1], making this research the largest scale study on choice of prompt on accuracy within economics found in the literature.

Finetuning

For Exp. 2 and 3, a custom finetuned version of the Babbage model was used. Finetuning is a term that refers to custom training an existing model using a specific dataset. The model was trained using a subset of the EconQA dataset. The dataset was split into two groups, the train set (n=2891) and a test set (n=723). No validation set was used. These datasets were used to finetune the Babbage model with the following hyper parameters:

Hyperparameters for Babbage Finetuning	
Learning rate multiplier	0.1
Prompt Loss Weight	.2
# Epochs	4
Classification # of Classes	6
Batch Size	4

Experimental Design

Using the EconQA database, I ran 11 experiments testing accuracy subject to various prompt and model choice interventions. The purpose of each experiment is summarized in Table 1 and Table 2.

[1] Except Exp. 11, which was terminated early.

Babbage Experiments description

Babbage Experiments		
Experiment Name	**Experiment Description**	**Model Utilized**
Exp. 1-Baseline Babbage	This experiment tested the "out-of-the-box" performance of OpenAI's babbage model. This experiment used no prompting beyond the question verbatim.	text-babbage-001
Exp. 2-Finetuned Babbage	This experiment tested the performance of the custom finetuned version of babbage. This model was custom finetuned on 2911 correct prompt-answer pairs. This experiment used no prompting beyond the question verbatim.	finetuned-text-babbage-001
Exp. 3-Finetuned Babbage with Prompt	This utilized the same finetuned model as experiment 2, but added a prompt before each question. This prompt was designed to discourage the model from attempting to provide any completion other than the letter corresponding the models choice of the best answer. The prompt read: "Please respond to the following multiple choice question. Your response should be exactly one letter long, corresponding to the letter of the most likely answer. Here is the question:". This is the same prompt as used in Chat Experiment 5.	finetuned-text-babbage-001

Table 1

Chats Experiment Description

Chat Experiments		
Exp. 4 Chat "Baseline"	This evaluated the performance of the gpt-3.5-turbo model without any prompting. The model was given the question verbatim. Temperature, a hyperparameter which controls randomness of completions, was set to 0.7 to increase diversity in responses. The results from Exp. 4 are taken from the first completion from Exp. 5 for each prompt.	gpt-3.5-turbo
Exp. 5 Chat "Baseline 3x"	This used the same prompt as Exp. 4 but repeated twice. If the model's output was the same between trials, that output was recorded. If the outputs diverged, the query was run a third time and the third output was recorded. This was to test if self-consistency is a useful heuristic for accuracy and to see if majority voting can improve accuracy. Temperature, a hyperparameter which controls randomness of completions, was set to 0.7 to increase diversity in responses.	gpt-3.5-turbo

Exp. 6 Chat "one letter"	This utilized a similar design to Experiment 3, but using Open AI's gpt-3.5-turbo model. The prompt read: "Please respond to the following multiple choice question. Your response should be exactly one letter long, corresponding to the letter of the most likely answer. Here is the question:", followed by the question. This is the same prompt as used in Experiment 3.	gpt-3.5-turbo
Exp 7- Chat "with Context"	This experiment evaluated if prompting the model to "provide context" about the question before answering improves performance. The purpose was to evaluate if performance can be improved by asking the model to recall and consider other background information not explicit within the prompt.The prompt read in two brief sentences, consider the following question. You do not need to answer the question yet, just provide context on the subject. Here is the question:", followed by the question.	gpt-3.5-turbo
Exp. 8- Chat "with Definitions"	This experiment asked the model to define any relevant terms before answering the question. The prompt read "In 2 brief sentences, consider the following question. You do not need to answer the question yet, just provide relevant information on the subject, including definitions or general knowledge. Here is the question:" Note that all definitions were generated by the model.	gpt-3.5-turbo
Exp. 9- "CoT Few-Shot"	This experiment used a CoT 4-shot prompt. The prompt consisted of four solved examples with no other pre-prompting. This prompt was expected to perform the best given existing literature showing CoT prompts with examples tend to perform better on many multiple choice tasks than any other prompting strategy. Because the prompt is quite long, the full text of the prompt is shown only in the appendix.	gpt-3.5-turbo
Exp. 10 "CoT zero-shot (Zhou mod)"	This experiment modified the zero-shot CoT prompt laid out by Zhou. The prompt was appended to the question rather than preceding the question in other experiments. The prompt read "Answer: Let's work this out in a step by step way, showing all steps before choosing an answer, to be sure we get the right answer."	gpt-3.5-turbo
Exp. 11 "CoT zero-shot (Zhou)"	This experiment used the best performing prompt from the Hebenstreit paper which compared ten prompts on diverse datasets. This prompt was originally suggested by Zhou. Execution of this experiment was abandoned early (n=133) due to a tendency for the model to choose an answer prior to reasoning through the question.	gpt-3.5-turbo

Table 2

For all the Chat experiments (except for Exp. 6 “1 Letter”), evaluation used a two-step approach depicted in Figure 1. The first step queried the LLM with the experiment specific prompt appended with the question. The API returns the LLM’s response, usually some form of verbose reasoning. Then a new query containing the ‘extraction prompt’ is sent to the LLM, along with previous question specific conversation history, to extract the model’s final answer in a machine readable way.

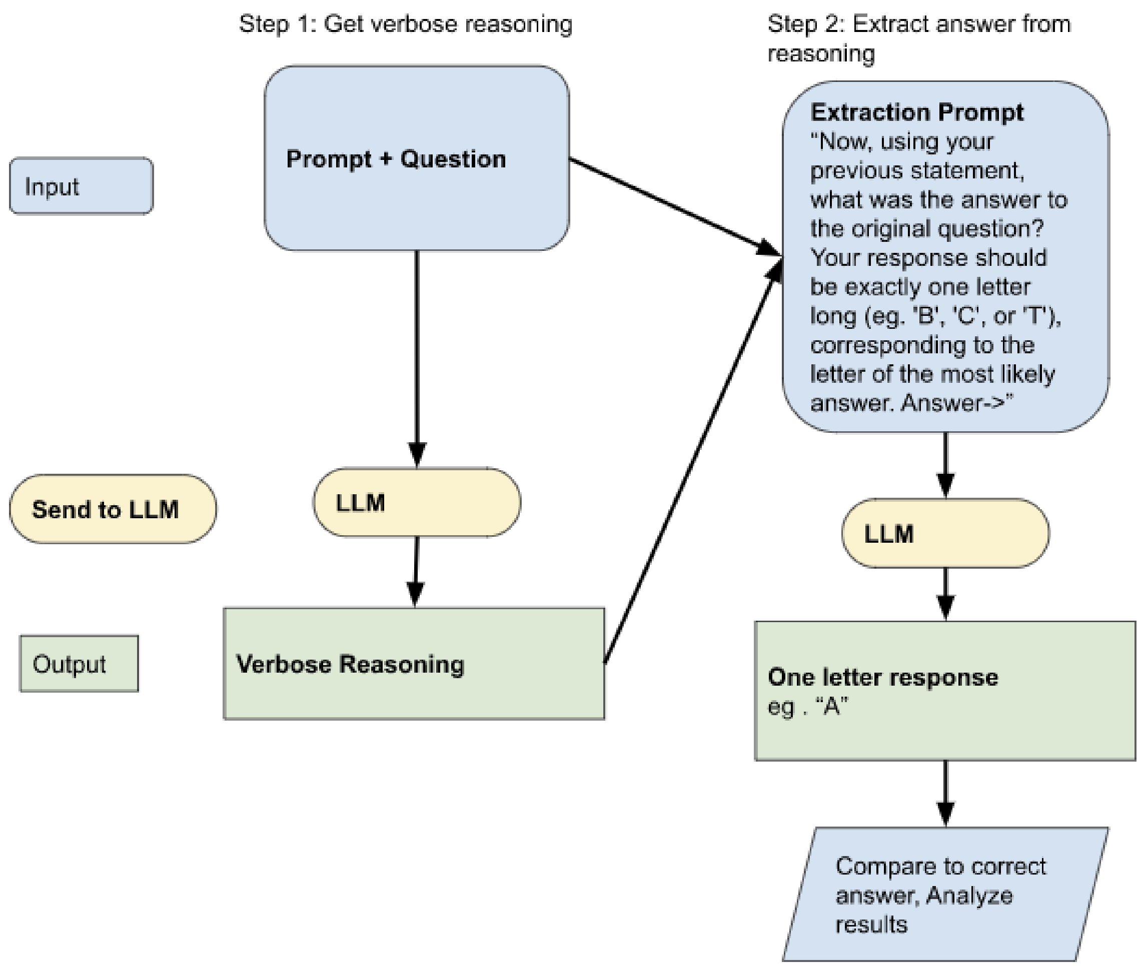


Figure 1- Two Prompt Flowchart

All experiments used similar hyperparameters, with temperature=0 except for experiments 4 and 5, which had temperature set to 0.7. Temperature is a parameter which tells the model how random its completions should be. At temperature=0, completions are entirely deterministic. In experiments 4 and 5, temperature was set to 0.7 to increase randomness and investigate whether results which are repeated between trials increases accuracy when a query is run multiple times is more likely to be accurate than when different trials result in different answers.

Results

In the tables below, I detail a few key metrics regarding each model's performance. When interpreting all the results it is important to remember that these are specific to the EconQA dataset and that results may not generalize to other datasets. Further, results are specific to the specific models used, so results are not guaranteed to generalize. However, based on results from Hebenstreit, relative accuracy of prompts tends to hold across models of sufficient size.

Summary of Accuracy

Overall Accuracy

All models were tested for all 3,623 questions. While computationally costly, large sample sizes give us the best chance to determine how changing models or prompting affects quality of response. This includes the finetuned Babbage models, which were tuned on a subset of the questions from this dataset. For this reason, accuracy in the test subset is reported separately as well. Note that this subset of questions appears to be more difficult than average, despite being chosen randomly, demonstrated by lower performance in Exp.1 on this subset. For this reason, accuracy on this subset is reported for all experiments.

Accuracy for Babbage Experiments			
	Exp. 1-Baseline Babbage	Exp. 2-Fine Tuned Babbage	Exp. 3-Fine Tuned Babbage with Prompt
Accuracy for all questions (n=3623)	0.296	0.522	0.418
Accuracy for questions not used in training (n=732)	0.280	0.337	0.321
Correlation between confidence and accuracy for all questions (Pearson's)	0.057	0.351	

Table 3- Accuracy for Babbage Experiments

Accuracy for Chat models	
Prompt	Accuracy For all questions (n=3623)
Exp. 4- Chat "Baseline"	73.8
Exp. 5- Chat Baseline "Self Consistency"	73.6
Exp. 6- Chat "one letter"	73.1
Exp 7- Chat "Generate Context"	70.2
Exp. 8- Chat "Generate Definitions"	70.8
Exp. 9- Chat "CoT 4-Shot"	76.3
Exp.10- Chat CoT zero-shot “Zhou mod”	74.0
Exp. 11- Chat CoT zero-shot “Zhou”	72.7

Table 4- Accuracy for Chat models

As is seen, accuracy does not vary dramatically between prompts using the chat model. Accuracy is always between 70-77%. However, prompt choice certainly does matter. Accuracy is highest for the CoT few-shot prompt from Exp. 9. Surprisingly, the zero-shot CoT prompts perform similarly to the baseline chats model. The lowest performing model is Exp. 7-Chat “with context”, at 70.2%. Surprisingly Exp. 5’s performance was below Exp. 4’s. The second highest performance

Accuracy by Question Type

Some literature shows that certain prompts improve performance on some tasks but not others. Since the dataset used has many types of labeled tasks, detailed further in the appendix under “question type”, we can assess how finetuning and choice of prompts affects performance for different types of prompts. A code table with examples is presented in appendix (Figure 10 Question Types with Examples) to help interpret the meaning behind each type code.

			Exp. 1 Baseline Babbage *	Exp. 2 Finetuned Babbage*	Exp. 3 Finetuned Babbage "One Letter"*	Exp.4- Baseline Chat	Exp. 5- Baseline Chat "Self-Consistency"	Exp. 6 One letter	Exp. 7 - Chat "Generate Context"	Exp. 8- Chat "Generate Definitions"	Exp. 9 CoT few shot	Exp 10- CoT Zero-Shot "Zhou Mod"
Question type	A	2156	30.0%	33.5%	31.4%	71.0%	71.4%	70.2%	66.5%	67.0%	**74.5%**	71.2%
	C	148	28.0%	34.4%	34.4%	54.7%	55.4%	50.7%	48.0%	45.0%	**63.5%**	62.2%
	D	632	32.0%	36.6%	30.1%	86.1%	86.1%	86.4%	86.1%	85.0%	**88.9%**	86.2%
	E	18	56.0%	40.0%	20.0%	38.9%	38.9%	33.3%	72.2%	39.0%	**77.8%**	61.1%
	F	655	30.0%	33.1%	36.4%	**78.5%**	77.9%	77.9%	74.0%	77.0%	75.6%	76.5%
	G	34	34.0%	0.0%	33.3%	34.4%	31.3%	40.6%	34.4%	**47.0%**	31.3%	28.1%
	Grand Total	**3643**	28.0%	33.7%	32.1%	73.8%	73.9%	73.1%	70.2%	70.8%	**76.4%**	**74.0%**

Table 5- Accuracy by Question Type[2]

We see that type ‘D’ (Definitional) questions have strong performance nearly across the board. For each chat experiment, the strongest question category is D. This is not surprising as these types of questions do not require much successful reasoning and are relatively straightforward. Further, the models are likely to have seen higher quantities of text relevant to definitional questions than other types of questions.

Type ‘G’ represents graphical questions. There is poor performance in this category because some questions involve referencing graphs (which were not provided to the model) to properly answer the questions. An attempt was made to remove questions which referenced graph or other questions, but some were not properly removed. In any future work using EconQA, this category should be excluded from the dataset. However, in some ways, the low performance could be seen as good news, suggesting that the model has not simply ‘memorized’ correct answers learned during its training.

Notice that average performance is highest for Exp. 9 CoT Few-Shot both overall and in 4 out of 6 type groups. Since G (Graphical) is a small sample size with lower quality questions than the others, let’s focus on F (Factual), the only major category which Exp. 9 did not achieve best in class in. Based on these results, I hypothesize that while the CoT mechanism is likely to help with logical reasoning tasks, many facts cannot be reasoned through, and imperfect reasoning may act as a distraction, introducing noise. This distraction results in CoT causing the model to underperform no prompt at all. This is an interesting line of research which should be further explored in future work.

[2] Note the 11th experiment was abandoned early (n=133), so is not included in this table

Assessment of Reasoning

Experiments 4, 5, 7, 8, 9, 10, and 11 used a two-step design. The first query combined the experiment's relevant prompt and question which the LLM responded to with an intermediate response, which I refer to as 'intermediate reasoning'. A second follow up 'extraction prompt' was then sent to the model to the LLM. This process is illustrated in Figure 1- Two Prompt Flowchart.

Generally, for all prompts, reasoning was plausible, even including some reasoning chains which led to wrong answers. The model was more often than not able to provide correct and intelligible definitions of concepts and was able to correctly eliminate some options. Reasoning was of the highest quality in Exp.9.

The most common themes within reasoning chains that resulted in wrong answers include incorrect premises, errant logic, or wrong definitions. Other common themes include rejecting the question (eg. "none of the above"), continuing reasoning after finding a correct answer, or indecisiveness between alternatives.

Wrong answers were consistent between prompts; For most questions, despite differences in reasoning path and prompt, final answers were consistent.

In Exp 8, which asked the model to generate definitions before choosing an answer, most definitions were true, but some were not relevant or sufficient to answer the question. Common errors in the intermediate reasoning step for Exp. 8 included providing overly broad definitions while other errors were due to wrong answers despite correct definitions. Some definitions included the true answer, but the model still answered incorrectly in the final step.

In Exp. 11, which used the original Zhou prompt, reasoning steps were very frequently skipped. For about half of the questions, the model skipped all reasoning steps or answered the question before providing a post hoc justification. This is important because evidence has shown that CoT reasoning only improves performance if reasoning is laid out before an answer is chosen (Hebenstreit). Therefore, the lower performance from Exp. 11 compared to Exp. 10 comports with existing literature, despite Exp. 10 suggesting a different optimal zero-shot prompt than Zhou found.

Biases in Completion Choices

To assess if the model is systemically biased to choose one completion class (A,B,C,D,T,F) over another, I evaluate the distribution of the model's completions compared to the distribution of true values. For each true value, we determine the proportion of model completions for each of the 6 represented classes or a 7th "other" class if the completion is anything other than one of the 6 answer classes represented in EconQA. This analysis allows us to determine if the model expresses a bias in favor of choosing a particular letter. If the model is choosing randomly, we would expect to see an equal proportion of completion classes for each true answer. This analysis is conducted with a table known as a 'confusion matrix'. Full confusion matrices for each prompt and model combination are reported in the appendix.

The Baseline Babbage model from experiment 1 is heavily biased. It has a strong bias to chose C around 80% of the time for ABCD questions, regardless of the true answer. There is also a strong bias towards choosing False, which it choses 100% of the time for TF questions. These biases are somewhat reduced when we look at the finetuned Babbage models and further reduced when using the more powerful chats model.

Across chat prompts and models, raw accuracy for true/false questions is similar to performance for the four choice questions. This is surprising since the accuracy from random guessing for true/false questions is .5 compared to .25 for ABCD questions. This indicates that the model underperforms on TF questions relative to ABCD questions. It is unclear why this may be.

Detailed Experiment Level Results

In this section I present selected findings for each experiment.

Exp. 1- Baseline Babbage

Baseline Babbage's Mean Accuracy is 29.56%. This is the lowest of all the experiments tested.

Babbage's confidence in an answer, measured by the returned probability of the top token, is not a good proxy for propensity for correctness. As can be seen in the below figure, there is not a consistent relationship between Babbage's confidence and how often it gets questions right. In fact, the questions it seems to be most confident (shown in the 80-89% confidence column) about are in fact the questions it gets wrong most often, with just 18% accuracy. If the model is good at judging its likelihood of correctness, mean correctness should fall within the range for each confidence range and we would expect a 'staircase-shaped' graph. The correlation coefficient was low at 0.057.

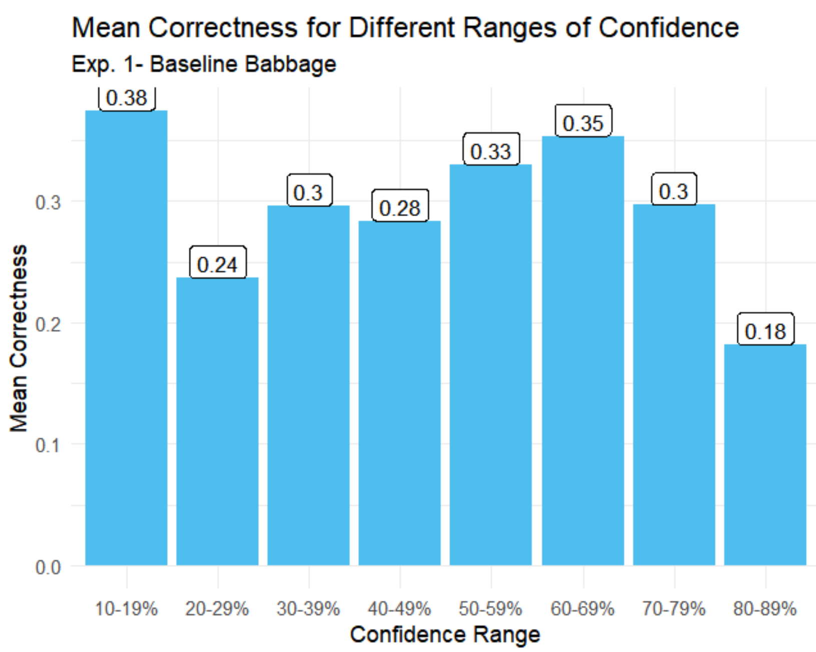


Figure 2

In the confusion matrix in the appendix, we can see that the model understands the format of how it should respond. In all but two responses, the model completes with one of the six possible answers (ABCDTF). However, the distribution of Exp. 1's completions looks very different from the true answers. It seems to have a strong bias to choose C, for whatever reason[3]. All ABCD questions have around 80% of completions being C regardless of the true answer. Interestingly, it never chooses "T" for "true" as its top choice for any question.

[3] Perhaps the popular moniker "when in doubt choose C" was overgeneralized by the model.

In the next graph, I analyze which question formats Babbage does best and worst at. Babbage appears to perform best at equation questions (E), somewhat surprisingly given the body of evidence that LLMs fall short on even some basic arithmetic tasks. However, this question type was poorly represented in the dataset with only 19 type 'E' questions.

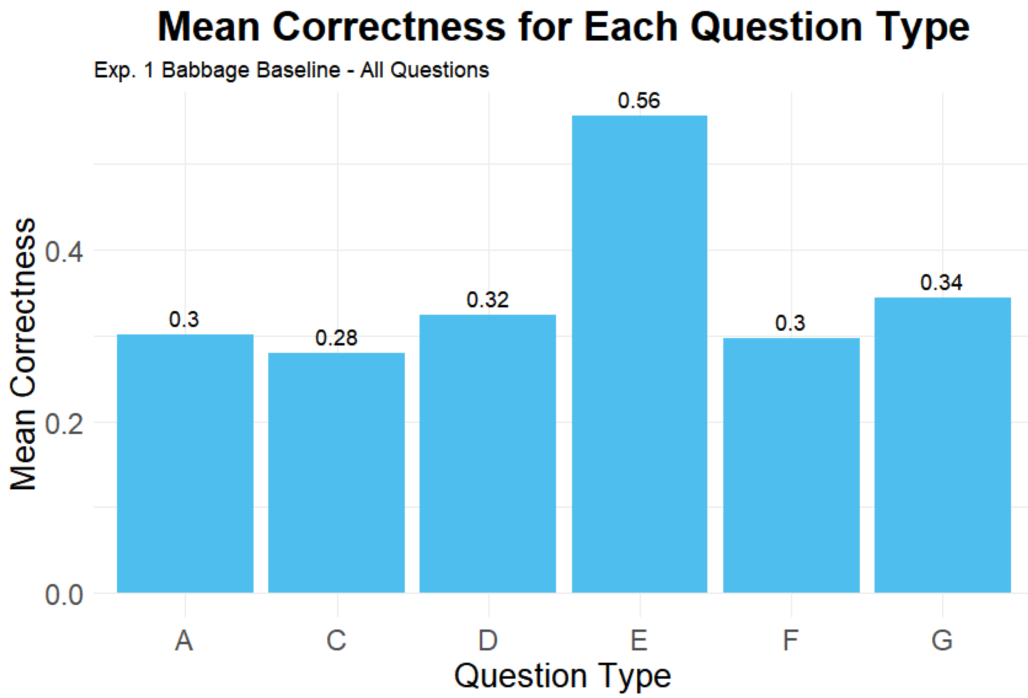


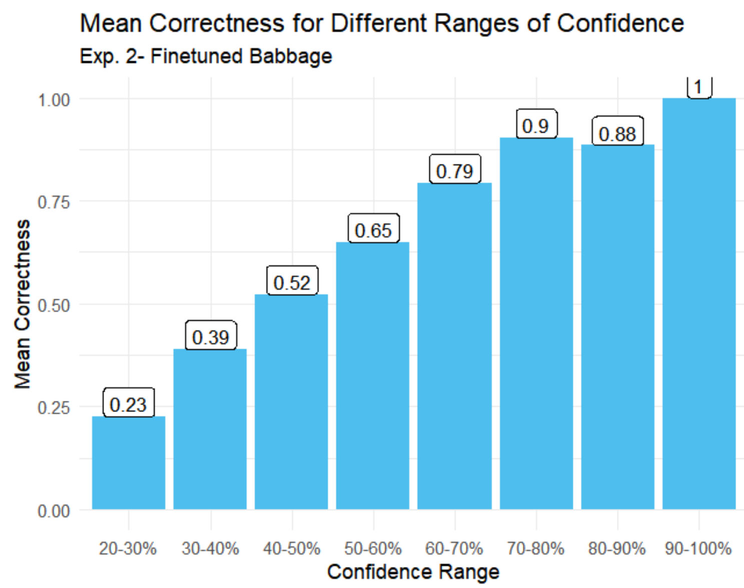
Figure 3

Exp. 2- Finetuned Babbage

Finetuned Babbage's overall accuracy was 52.17%. Accuracy for new questions which were not included in the training set was 33.74%. This represents a modest improvement over Baseline Babbage's accuracy.

The finetuned model's completion distribution is much closer to the distribution of correct answers than Baseline Babbage's. It still guesses C disproportionately often, but to a lesser extent. It also emits no non-sequitur tokens like Baseline Babbage did with 'China' and 'Social'. Note that while Baseline Babbage predicted 'T' zero times, Finetuned Babbage predicted 'T' 160 times, much closer to the underlying true frequency.

For the full set of questions, Finetuned Babbage also does well at evaluating its probability of being correct, shown by the calibration curve below.



However, if we re-examine these results using only questions the model has not been trained on, a more useful metric, the model’s confidence performs less favorably as a proxy for accuracy. This change is not surprising. Models like Babbage use patterns in their training data to make predictions. When presented with unfamiliar data, their performance can drop significantly because they are forced to extrapolate from the patterns they have learned rather than directly apply them. The lack of exposure to similar questions during the training phase impacts the model's ability to predict its own accuracy in these cases. However, the clear trend that higher confidence is associated with greater accuracy remains.

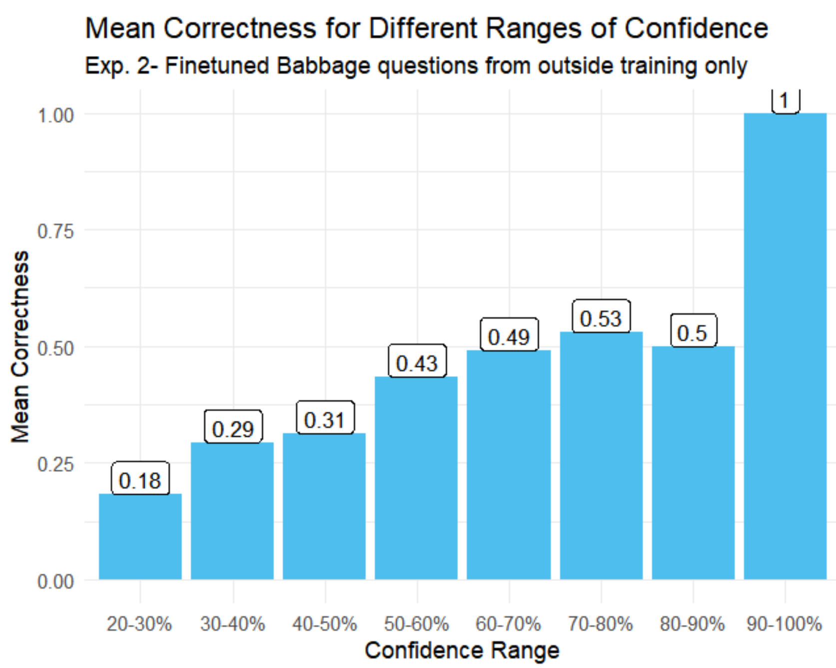


Figure 4

Within the subset of questions from outside the training set, the model also excels most at Equation (E) type questions. Note that the spread in performance between different question types seems to flatten. Also note the fall in correctness for Graphical (G) type questions is primarily due to small sample sizes and some questions which required references to graphs which were not provided.

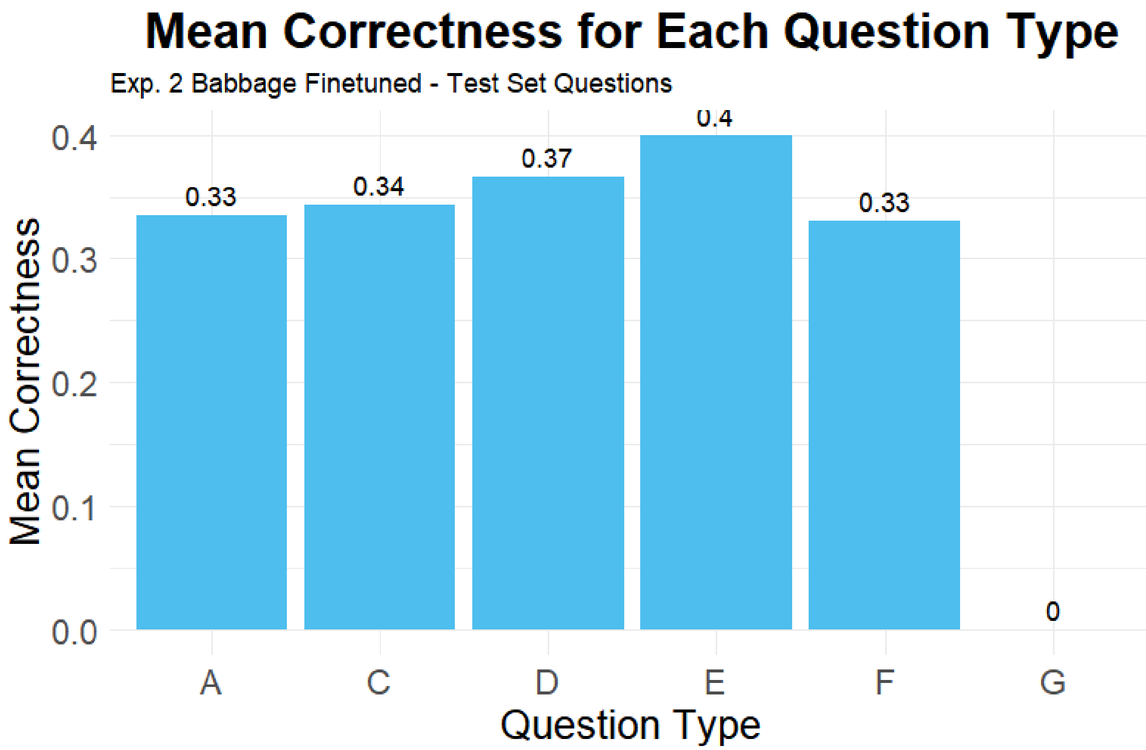


Figure 5

Exp. 3- Finetuned Babbage "one letter prompt"

The model's distribution of T relative to F completions is flipped compared to Experiment 1, where it predicted 'T' much more than the underlying true frequency. This trend can also be seen in the confusion matrix in the appendix.

Exp. 4- Chat Baseline

This experiment featured no prompt, only querying the model with the raw question text. The query to the model in Exp. 4 solely consisted of the verbatim text of the question, succeeded by a follow up extraction prompt. Exp. 4 and Exp. 5 were run together, so the model choices from Exp. 4 are simply the first output from Exp. 5, without any repetitions.

Overall performance was 73%.

Exp. 5- Chat Baseline with Self Consistency

This experiment used the same set up as Exp. 4, but repeated the question multiple times with temperature set to 0.7 to introduce stochasticity into model completions. Each query was run at least twice. If the two results disagreed, the query was run a third time as a 'tie-breaker'.

The purpose of this design is to test if checking consistency can be used as a proxy for how certain the model is, with the goal of both a) measuring the model's certainty, b) evaluating how consistent the model's completions are given a prompt, and c) increasing performance by sampling a wider space of possible answers . The idea is that the model will be more consistent on questions it 'knows' the answer to and on other questions guess more randomly. If the model's choice of the best answer is somewhere between certain and totally random, then consistency between two answers serves as a signal that the model is more confident about a particular answer.

The vast majority of questions had consensus between queries, including consensus where the model was wrong. Answers tend to be the same across prompts as well.

Exp. 5 was expected to perform better, or at least no worse, than Exp. 4. However, accuracy for Exp. 5 was surprisingly slightly lower than Exp. 4. This was most likely a result of random chance. However, the lack of improvement suggests that this is not an effective technique to improve performance on the types of questions represented in in EconQA.

Despite these disappointing results, self-consistency and majority voting have been shown to be powerful tools for improving performance when combined with a CoT reasoning based strategy. While I would have liked to evaluate this for the EconQA dataset using CoT, the computational, time, and financial costs were too large.

Exp. 6- Chat "One Letter"

In Experiment 6, I tested how prompting the model to answer in 1-step instead of the two-step process shown in Figure 1- Two Prompt Flowchart affected performance. The model was given no room to reason through questions and outputs were restricted to one token. The full prompt text read "Please respond to the following multiple choice question. Your response should be exactly one letter long, corresponding to the letter of the most likely answer. Here is the question:". Given that the model in Experiments 4 and 5 did not spontaneously reason, performance for Exp. 6 was expected to be broadly similar.

Overall performance was 73.1%. This was slightly below Exp. 4's performance, which scored 73.8%. Correctness by Question type was also broadly similar.

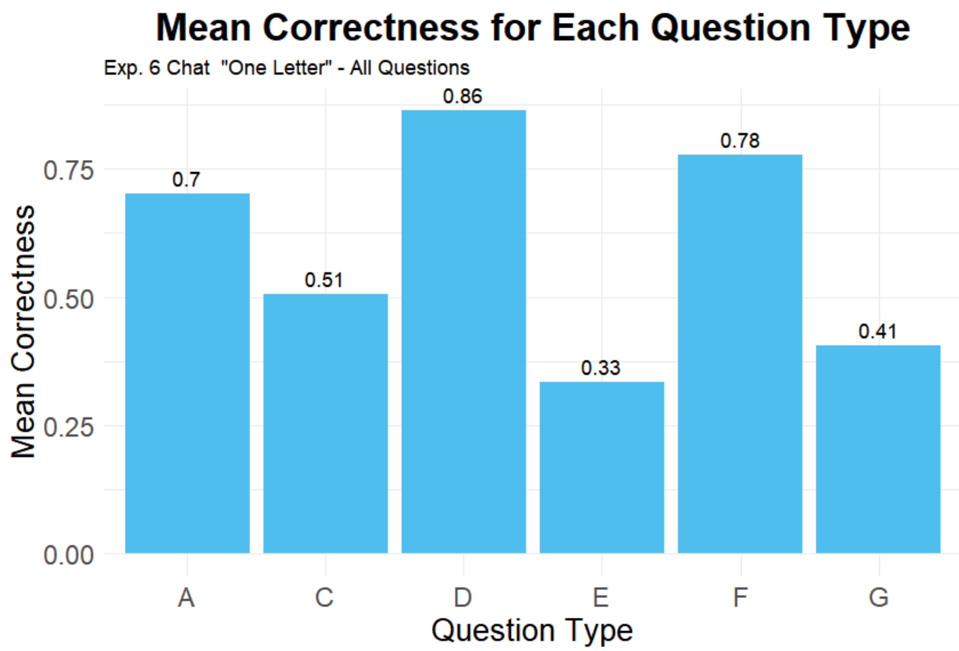


Figure 6- Correctness by type for Exp. 6

Exp. 7- Chat with Context

In Experiment 7, I tested how prompting the model to reflect on the context of the question before deciding on an answer affects accuracy. The prompt read "In 2 brief sentences, consider the following question. You do not need to answer the question yet, just provide context on the subject. Here is the question:" followed by the question.

In contrast to expectations, Exp. 7 performed slightly worse than Exp. 6 Chat "1 letter", suggesting this prompt is not useful to elicit superior outputs. This is somewhat surprising since Experiment 6 was given only one token to answer the question with no room to 'think' through or contextualize the question. This reinforces the notion that the choice of prompt matters; simply giving the model more room to 'think' through a question is not sufficient to improve performance. Exp. 7's inferiority is especially true given the much higher computational costs that were associated with this run, since ~5x more tokens were used compared to Exp. 6.

One problem with this prompt was that it was not adequately specific about what type of context should be provided. While a human reader might interpret 'provide context on the subject' to mean provide context about the ideas and concepts within the question, the model largely restated the question in different words in response to this prompt. A few representative examples of the 'context' which it returned are provided in the appendix.

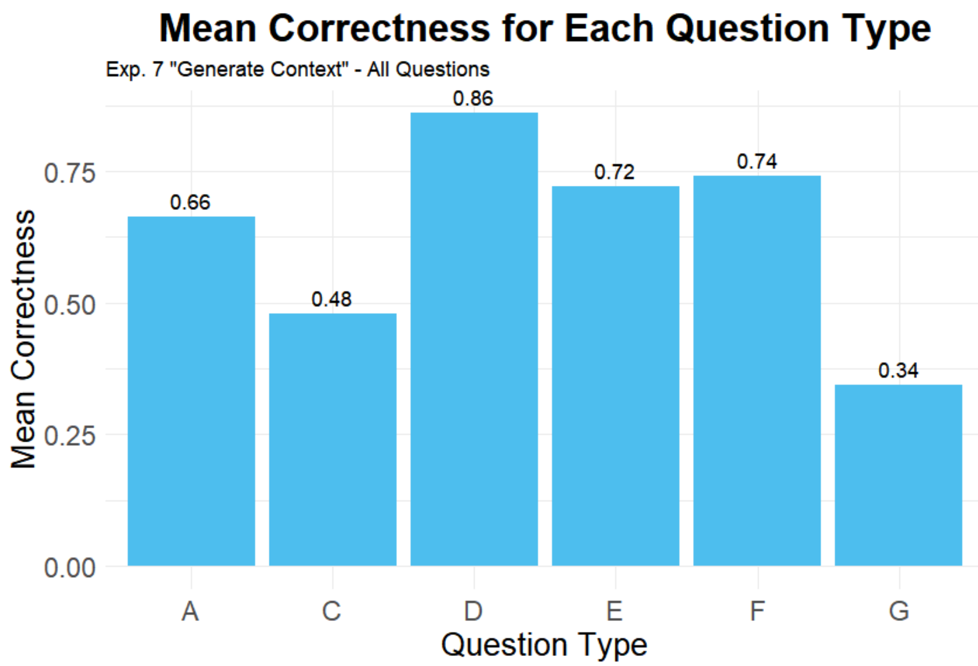


Figure 7

Exp. 8- Chat with Definitions

This Experiment tested if asking the model to define terms and other relevant concepts before answering the final question improves performance. The full prompt read "In 2 brief sentences, consider the following question. You do not need to answer the question yet, just provide relevant information on the subject, including definitions or general knowledge. Here is the question:"

Based on the confusion matrix in Figure 18, there was a bias for the model to choose True when the answer was False that was not present in the opposite direction. There does not appear to be a major bias for ABCD questions.

Overall accuracy was only 70.8%, making this experiment the Chat experiment with the second lowest accuracy rate. Surprisingly, performance on Definitional ("D") questions was the lowest of all the Chat experiments at 85%, despite the prompt being tailored to definitional questions.

Intermediate reasoning steps and generated definitions were generally true. Common errors in the intermediate reasoning step for Exp. 8 included providing overly broad definitions while other errors were due to wrong answers despite correct definitions. Some definitions included the true answer, but the model still answered incorrectly in the final step. However there were oftentimes insufficiently specific definitions to answer the question. Some definitions required a logical inference that was not made in the limited space of the two-sentence response. For example, the third entry in Table 7-Example exchanges for Exp. 8 "Provide Definition" in the appendix shows the model correctly defining the Marginal Rate of Substitution as "the slope of an indifference curve at a given point.", however it was unable to make the inference to get to the final answer "[The marginal rate of substitution] declines as one moves southeast along an indifference curve.".

Exp. 9- Chat CoT Few-shot

Existing research suggests that using Chain-of-Thought (CoT) prompting tends to yield the highest accuracy scores. CoT prompting involves asking or demonstrating to the model a series of intermediate reasoning steps to generate a logical chain of reasoning. Most reported SOTA performance on benchmark datasets use some combination of CoT prompting and other techniques. With this in mind, I tested a 4-shot CoT prompt's performance on all questions. '4-shot' means that 4 examples were given. The prompt was quite long due to the length required to demonstrate four fully reasoned answers. Examples were handcrafted and the same four examples were provided for all questions. Every example solution started with "Thinking step by step," followed by defining important terms. Then the examples proceeded to solve each problem from first principles laid out in definitions and either select the right answer or eliminate wrong answers. Each example was about four sentences long. The full prompt is in the appendix.

Overall accuracy was highest using the prompt from Exp. 9 compared to all other prompts. This finding aligns with existing research which shows that CoT prompts with examples tend to be one of the best ways to increase performance.

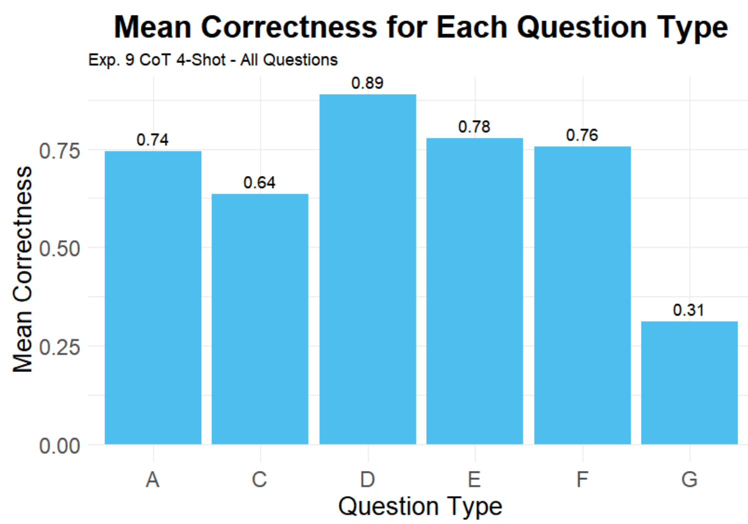


Figure 8

Performance was particularly high on "Equation-Type" questions, which improved from 73.8% in the baseline to 78% in Exp. 9. However, performance on Factual questions was slightly worse than in the baseline case. This aligns with existing research.

Exp. 10- CoT Zero-Shot (Zhou Modified)

This experiment used a modified version of the original zero-shot CoT prompt developed by Zhou and corroborated by Heibenstreit tested in Exp. 11. Zero-shot means that it uses CoT prompt with no examples. The prompt was modified to specify that an answer should only be chosen after fully reasoning through the question. The full prompt read "Answer: Let's work this out in a step by step way, showing all steps before choosing an answer, to be sure we get the right answer."

Overall accuracy was 74.0%. This was the second highest performance prompt, second only to Exp. 9's 4-Shot CoT prompt. This prompt outperformed the Zhou prompt by 1.3%. This suggests that future work using Zero-shot CoT prompts should consider explicitly adding a "think first, answer second" clause, especially if they find the model frequently answers before entering the CoT reasoning.

The question types which performed strongest are shown in Figure 9.

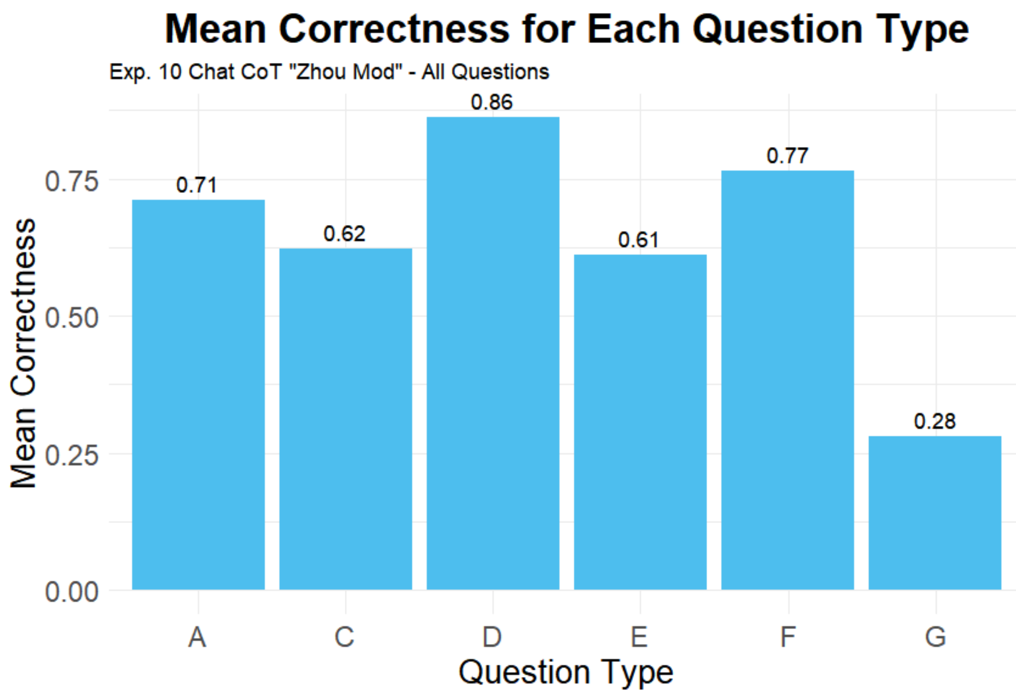


Figure 9

Exp. 11- CoT Zero-Shot (Zhou)

This experiment uses the original zero-shot CoT prompt developed by Zhou and corroborated by Heibenstreit. Zero-shot means that it uses CoT prompt with no examples. The prompt read "Answer: Let's work this out in a step-by-step way to be sure we have the right answer".

Overall performance for the prompt was 72.7%, lower than the baseline or one letter prompts. This was surprising because in tests by Heibenstreit, this prompt far outperformed other prompts.

Testing of this prompt was abandoned early because it frequently chose an answer prior to providing reasoning. For this reason, the prompt was not tested on the full EconQA dataset, but rather a smaller random subset (n=133).

The problem of choosing an answer before reasoning through the question was very common for the Zhou prompt, with this flaw occurring in 66/133 (49.62%) of questions tested. The flaw was particularly noticeable for TF questions, where every single question TF question was answered and then rationalized post-hoc, with many providing no rationalization or explanation at all.

A detailed breakdown of this prompt's accuracy by question type is not presented here because the sample size was too low to reliably generalize any conclusions. However, as in other experiments, Application of Concept ("A"), Definitional ("D"), and Factual ("F") questions had the best performance.

Discussion

Results from this research largely align with existing research. The most central finding is that CoT prompts tend to yield the best accuracy metrics.

Model performance was significantly improved by finetuning, with the Finetuned Babbage achieving a notable increase in accuracy over the baseline Babbage model. This reinforces the effectiveness of task-specific finetuning in enhancing model performance. However, the absolute increase in performance was limited, highlighting the inherent constraints of the model.

Of note is the disparity between model performance on questions from the training set versus new, unseen questions. Models trained on a particular dataset can overfit to its characteristics and therefore underperform when exposed to new data. This observation underscores the importance of generalizability in the development and evaluation of large language models.

Regarding prompting strategies, the Chat models generally outperformed the Babbage models across different prompts, showing the superiority of the larger and more powerful GPT-3.5-turbo model both on EconQA tasks and generally. Interestingly, overall model accuracy didn't vary dramatically across the different prompt strategies. This could imply that the model's inherent capabilities are relatively prompt-agnostic for most tasks represented in the EconQA dataset.

The study found that CoT few-shot prompts (Exp. 9) yielded the highest overall accuracy, outperforming other prompts across various question types. This aligns with previous research suggesting that the Chain-of-Thought approach is beneficial for many multiple-choice tasks. However, it underperformed on Factual questions, which might suggest that the CoT mechanism may not be an effective way to improve accuracy for fact retrieval tasks where attempts at logical reasoning might be a distraction.

Further, different types of questions yielded varying accuracy rates, suggesting that certain prompts are better suited to specific types of questions. Definitional questions had consistently high performance across all models and prompts. Conversely, graphical questions demonstrated poor performance, highlighting the limitations of text-based models in interpreting and responding to questions that require visual or spatial reasoning.

The results also revealed biases in completion choices, particularly with the Baseline Babbage model, which demonstrated strong biases towards certain responses. Finetuning somewhat mitigated these biases, indicating that carefully designed training can help manage model bias. The Chat models performed better than the Babbage models in this regard, but there remains a notable underperformance on true/false questions relative to ABCD questions.

Questions remain why performance on TF questions seems to be worse than expected. In Exp. 1, the model outputs 'F' for False for every single TF question. It is not clear if there is a bias towards choosing F generally built into Babbage's weights or if it was specific to this dataset. However, if this bias is systematic, it is not well documented.

Conclusion

The results from this exploration support existing stylized facts that choice of prompting does affect output quality. Existing literature indicates that using CoT prompting can improve performance over direct prompting, which this evidence supports. The two best performing experiments both used a CoT prompt. However, effects were relatively small. All Experiments using Chats scored between 70-76%.

In alignment with expectations, finetuning did change performance on accuracy dramatically, with a statistically significant and positive effect. One interesting finding is Babbage seems to perform worse with TF questions than even random chance would predict. All experiments had relatively worse performance for TF questions than ABCD questions when accounting for a higher base rate from random chance alone.

Generally, Intermediate reasoning steps were at least partially, if not fully correct. The best intermediate reasoning, based on a qualitative survey of randomly selected questions, was given by Exp. 9, which used CoT few-shot. Common pitfalls included insufficient specificity in reasoning, false assumptions or definitions, or logical non-sequiturs. Other reasons include reasoning to the correct answer but choosing a wrong one anyways.

Across the board, definitional questions were the question type with the highest performance. Equational questions were greatly improved by using CoT, with performance approximately doubling from the baseline Exp. 4 to CoT 4-shot Exp. 9. Using CoT lowered performance on factual questions for both Exp. 9 and Exp. 10, probably because these questions do not require complex reasoning. Suggest that CoT is most likely to be useful for ‘System 2’ tasks, while ‘System 1’ tasks often have equal or superior results with no prompting.

Further research should further investigate the strengths and weaknesses of LLMs in domain specific applications, like economics. Evaluation should focus on different skills and utilize multiple types of questions. Using textbook test banks is likely to be a useful data source for future domain specific research.

Appendix

Question Types with Examples

To explain and illustrate the different types of questions represented in the dataset, a sample of some questions are provided in the table below. This table is a useful reference for interpreting some of the figures and comparisons presented in the results section and is also useful to understand the breadth of questions represented in EconQA.

Figure 10 Question Types with Examples

Letter Code	Meaning	Description (from test bank)	Example Questions
A	Application of Concept	A question that tests student understanding of a definition or concept.	Which of the following arguments comes closest to constituting a legitimate economic exception to the case for free trade? A) the increase-domestic-employment argument C) the diversification-for-stability argument B) the cheap-foreign-labor argument D) the infant-industry argument
			Suppose the supply of product X is perfectly inelastic. If there is an increase in the demand for this product, equilibrium price: A) will decrease but equilibrium quantity will increase. B) and quantity will both decrease. C) will increase but equilibrium quantity will decline. D) will increase but equilibrium quantity will be unchanged.
C	Complex Analysis	A question that calls for students to apply more than one concept in order to reach the correct answer.	Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or decrease. D) will not change.
			If investment increases by $10 billion and the economy's MPC is .8, the aggregate demand curve will shift: A) leftward by $40 billion at each price level. C) rightward by $50 billion at each price level. B) rightward by $10 billion at each price level. D) leftward by $20 billion at each price level.
D	Definition	A straightforward question that tests recognition of a definition.	Nonprice competition refers to: A) low barriers to entry. B) product development, advertising, and product packaging. C) the differences in information which consumers have regarding various products. D) an industry or firm in long-run equilibrium.

			Normal profit is: A) a cost because any excess of total receipts over total costs will go to the businessperson. B) a cost because they represent payments made for the resources which the businessperson owns and supplies in his or her own enterprise. C) not a cost because a firm can avoid this payment by temporarily closing down. D) not a cost of production because it need not be realized for a firm to retain entrepreneurial ability.
E	Equation	A question that calls for solving one or more equations. Many of these are designated as advanced analysis questions.	If M is $300, P is $4, and Q is 200, then V must be: A) 4. B) 22/ 3. C) 1. D) 3.
			The multiplier can be calculated as: A) 1/(MPS + MPC) B) MPC/MPS C) 1/(1 - MPC) D) 1 - MPC = MPS
F	Fact	A problem that tests ability to recall a fact or some data.	In 2001, some 41 million Americans did not have health insurance. T) True F) False.
			Between 1999 and 2002, real GDP in Russia increased by an average annual rate of: A) 6 percent. B) 4 percent. C) 2 percent. D) zero percent.
G	Graphical	A question that requires students to demonstrate working facility with graphs.	Landowners will not receive any rent so long as: A) there is any tax on land. B) the supply and demand curves for land intersect. C) the supply curve of land is perfectly inelastic. D) the supply curve lies entirely to the right of the demand curve.
			The monopolistically competitive firm shown in the above figure: A) will realize allocative efficiency at its profit-maximizing output. B) cannot operate at a loss. C) is in long-run equilibrium. D) is realizing an economic profit.[4]

[4] Note this question is not answerable without the graph. While I attempted to remove all questions which referenced other questions or data, some graphical questions were errantly included in EconQA. This ends up being somewhat useful since the poor performance on these questions can be used to evaluate leakage (Chen et.al Apr 18 2023 “when do you need CoT”)

T	Table	A question that requires students to interpret or manipulate tabular material.	Removed from dataset

Full Text of Prompts

The Unabridged full text of each prompt is displayed below to promote transparency and reproducibility. All text formatting is preserved.

Exp. 1- Babbage Baseline Prompt

No prompt used.

Exp. 2- Babbage Finetuned Baseline Prompt

No prompt used.

Exp. 3- Babbage Finetuned "One Letter" Prompt

```
Please respond to the following multiple choice question.
Your response should be exactly one letter long,
corresponding to the letter of the most likely answer. Here
is the question:
```

Exp. 4- Chat Baseline Prompt

No prompt used.

Exp. 5- Chat Baseline "Self-Consistency" Prompt

No prompt used.

Exp. 6- Chat "One Letter" Prompt

```
"Please respond to the following multiple choice question.
Your response should be exactly one letter long,
corresponding to the letter of the most likely answer. Here
is the question:”
```

Exp. 7- Chat "Provide Context" Prompt

```
In 2 brief sentences, consider the following question. You
do not need to answer the question yet, just provide
context on the subject. Here is the question:
```

Exp. 8- Chat "Generate Definition" Prompt

```
In 2 brief sentences, consider the following question. You
do not need to answer the question yet, just provide
relevant information on the subject, including definitions
or general knowledge. Here is the question:
```

Exp. 9- CoT 4 shot Prompt

Q: Suppose we find that the price elasticity of demand for a product is 3.5 when its price is increased by 2 percent. We can conclude that quantity demanded: A) increased by 7 percent. B) decreased by 7 percent. C) decreased by 9 percent. D) decreased by 12 percent.

A: Thinking step by step, we know when price increases, quantity demanded must go down. That is, PE_d is always negative, so PE_d=-3.5. The formula for elasticity of demand is PE_d=(% change Q)/(% change P). Using the numbers from the problem, we have -3.5=(% change Q)/2. Multiplying both sides by 2, we get 7=(% change Q). Therefore, quantity demanded decreased by 7 percent. So the answer is B.

Q: The GDP tends to: A) overstate economic welfare because it does not include certain nonmarket activities such as the productive work of housewives. B) understate economic welfare because it includes expenditures undertaken to offset or correct pollution. C) understate economic welfare because it does not take into account increases in leisure. D) overstate economic welfare because it does not reflect improvements in product quality.

A: Thinking step by step, we know GDP measures the market value of all the final goods and services produced and sold in a given country or area. We should now evaluate if any options can be eliminated. While GDP does not include domestic work, this would understate, not overstate GDP, so we can eliminate option A. While expenditures taken to offset pollution are included in GDP, this would not lead to an underestimate in GDP, so we can eliminate option B. While GDP does not account for improvements in product quality, this implies GDP understates, not overstates economic welfare, so we can eliminate D. It is correct that GDP does not take into account increases in leisure and it is also true that this would cause GDP to understate economic welfare. So the answer is C.

```
Q: If a pure monopolist is producing more output than the
MR = MC output: A) the firm may, or may not, be maximizing
profits. B) it will be in the interest of the firm, but not
necessarily of society, to reduce output. C) it will be in
the interest of the firm and society to increase output. D)
it will be in the interest of the firm and society to
reduce output.


A: Thinking step by step, remember that MR=MC is the
optimality condition under perfect competition and the
societally optimal condition under normal assumptions.
Remember monopolists maximize profits when MR=MC and P>=MC.
The question states the monopolist is producing more than
the MR=MC output, so they are producing where MR<MC, which
is to the right of equilibrium on a supply and demand
curve. Thus, reducing output to MR=MC could raise both firm
profits and social welfare. So the answer is D.


Q: Noncash gift giving involves value loss when the
marginal utility of the gift to the receiver is less than
the product price. T) True F) False.


A: Thinking step by step, if the marginal utility of the
gift to the receiver is less than the product price, then
the giver has paid a certain amount for the gift, but the
receiver does not value it as much. This means there is a
loss of value. So the Answer is T.


Q:
```

Exp. 10- Chat CoT Zero-Shot "Zhou Mod"[5]

```
Answer: Let's work this out in a step by step way, showing
all steps before choosing an answer, to be sure we get the
right answer.
```

Exp. 11- Chat CoT Zero-Shot "Zhou"[5]

```
Answer: Let's work this out in a step-by-step way to be
sure we have the right answer
```

Confusion Matrices for Each Experiment

Confusion matrices are a common tool used in machine learning research to visualize the performance of an algorithm by comparing its predicted classifications against the actual, true classes. In a confusion matrix, the rows typically represent the actual, or true classes, while the columns represent the predicted classes by the model. The utility of a confusion matrix is in its ability to reveal potential biases in the model's classification. If the model shows a disproportionate preference for a particular class that doesn't align with its actual occurrence rate, this can indicate a systematic bias.

In this case, where the classes are A, B, C, D, T, and F, each row in the confusion matrix represents the actual class, and each column represents the predicted class. Each cell within the matrix shows the percentage of times the model predicted a particular class (column) for the given true class (row). By analyzing the distribution across each row, we can assess how well the model is performing in predicting each class, and where its predictions may be skewed.

Exp. 1- Babbage Baseline

		Model Output						
		A	B	C	D	F	Other	Grand Total
True Answer	A	3.68%	8.54%	78.65%	8.69%	0.00%	0.44%	679
	B	4.45%	6.96%	81.41%	5.93%	0.00%	1.25%	877
	C	4.27%	8.18%	79.03%	7.70%	0.00%	0.83%	844
	D	4.54%	6.88%	80.82%	6.88%	0.00%	0.88%	683
	F	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	305
	T	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	255
	Grand Total	**131**	**235**	**2467**	**223**	**560**	**27**	**3643**

Figure 11

[5] For Exp. 10 and 11 the prompt was appended to the end of the question rather than preceding the question in other trials. This was to match the conditions the prompt was designed by Zhou for.

Exp. 2- Babbage Finetuned

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answers	A	21.35%	30.34%	32.70%	15.61%	0.00%	0.00%	0.00%	679
	B	6.04%	59.06%	25.31%	9.46%	0.00%	0.00%	0.11%	877
	C	6.16%	18.48%	63.74%	11.37%	0.00%	0.00%	0.24%	844
	D	7.47%	22.40%	25.77%	44.36%	0.00%	0.00%	0.00%	683
	F	0.00%	0.00%	0.00%	0.00%	88.52%	11.48%	0.00%	305
	T	0.00%	0.00%	0.00%	0.00%	50.98%	49.02%	0.00%	255
	Grand Total	**301**	**1033**	**1158**	**588**	**400**	**160**	**3**	**3643**

Figure 12

Exp. 3- Babbage Finetuned One Letter

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	11.93%	19.44%	38.88%	29.75%	0.00%	0.00%	0.00%	679
	B	2.74%	33.18%	41.28%	22.69%	0.00%	0.00%	0.11%	877
	C	3.44%	11.97%	62.44%	21.92%	0.00%	0.00%	0.24%	844
	D	4.25%	13.18%	33.53%	49.05%	0.00%	0.00%	0.00%	683
	F	0.00%	0.00%	0.00%	0.00%	13.11%	86.89%	0.00%	305
	T	0.00%	0.00%	0.00%	0.00%	3.53%	96.47%	0.00%	255
	Grand Total	**163**	**614**	**1382**	**921**	**49**	**511**	**3**	**3643**

Figure 13

Exp. 4- Baseline Chats

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	68.48%	13.40%	7.51%	10.01%	0.00%	0.44%	0.15%	679
	B	6.27%	77.77%	6.04%	9.46%	0.00%	0.23%	0.23%	877
	C	6.64%	11.14%	72.63%	8.89%	0.00%	0.12%	0.59%	844
	D	6.15%	9.22%	9.96%	74.23%	0.00%	0.00%	0.44%	683
	F	0.00%	0.33%	0.66%	0.00%	73.11%	25.90%	0.00%	305
	T	0.00%	0.39%	0.78%	0.00%	21.96%	76.86%	0.00%	255
	Grand Total	**618**	**932**	**789**	**733**	**279**	**281**	**0**	**3643**

Figure 14

Exp. 5- Baseline Chats 3x

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	69.51%	12.81%	7.51%	9.87%	0.00%	0.00%	0.29%	679
	B	6.85%	77.51%	6.28%	9.13%	0.00%	0.11%	0.11%	876
	C	7.01%	11.28%	73.16%	7.96%	0.00%	0.12%	0.48%	842
	D	6.00%	9.22%	9.81%	74.52%	0.00%	0.15%	0.29%	683
	F	0.00%	0.00%	0.33%	0.00%	72.46%	26.89%	0.33%	305
	T	0.00%	0.39%	1.18%	0.00%	22.75%	75.69%	0.00%	255
	Grand Total	**632**	**925**	**793**	**723**	**279**	**278**	**10**	**3640**

Figure 15

Exp. 6- One Word

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	71.72%	14.73%	7.22%	6.33%	0.00%	0.00%	0.00%	679
	B	7.41%	78.56%	6.39%	7.41%	0.00%	0.00%	0.23%	877
	C	7.70%	13.27%	72.39%	6.40%	0.00%	0.00%	0.24%	844
	D	9.52%	10.83%	9.37%	70.28%	0.00%	0.00%	0.00%	683
	F	0.00%	0.00%	0.00%	0.00%	56.39%	43.61%	0.00%	305
	T	0.00%	0.00%	0.00%	0.00%	12.94%	87.06%	0.00%	255
	Grand Total	**682**	**975**	**780**	**642**	**205**	**355**	**1422**	**3643**

Figure 16

Exp. 7- Provide Context

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answers	A	68.78%	17.08%	6.92%	7.07%	0.00%	0.00%	0.15%	679
	B	6.96%	77.51%	8.22%	6.85%	0.00%	0.00%	0.46%	876
	C	9.26%	15.56%	67.22%	7.48%	0.00%	0.00%	0.48%	842
	D	10.25%	13.76%	9.52%	66.03%	0.00%	0.00%	0.44%	683
	F	0.00%	0.00%	0.00%	0.00%	64.59%	34.10%	1.31%	305
	T	0.00%	0.00%	0.00%	0.00%	23.14%	76.86%	0.00%	255
	Grand Total	**676**	**1020**	**750**	**622**	**256**	**300**	**4676**	**3640**

Figure 17

Exp. 8- Generate Definition

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	66.86%	15.76%	7.51%	9.87%	0.00%	0.00%	0.00%	679
	B	7.76%	76.60%	7.53%	7.99%	0.00%	0.00%	0.11%	876
	C	7.60%	13.42%	70.78%	7.84%	0.00%	0.00%	0.36%	842
	D	8.64%	12.01%	11.42%	67.64%	0.00%	0.00%	0.29%	683
	F	0.00%	0.00%	0.00%	0.00%	58.36%	41.64%	0.00%	305
	T	0.00%	0.00%	0.00%	0.00%	15.69%	84.31%	0.00%	255
	Grand Total	**645**	**973**	**791**	**665**	**218**	**342**	**6**	**3640**

Figure 18

Exp. 9- CoT 4 shot

		Model output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	75.26%	7.07%	5.15%	7.66%	0.00%	1.33%	3.53%	679
	B	7.41%	75.26%	6.27%	6.61%	0.00%	1.71%	2.74%	877
	C	5.69%	6.52%	76.07%	7.46%	0.00%	1.42%	2.84%	844
	D	6.73%	5.12%	6.73%	78.18%	0.00%	1.02%	2.20%	683
	F	0.66%	0.66%	0.66%	0.00%	80.33%	16.39%	1.31%	305
	T	0.00%	0.00%	0.78%	0.00%	23.14%	73.73%	2.35%	255
	Grand Total	**672**	**800**	**782**	**707**	**304**	**281**	**97**	**3643**

Figure 19

Exp 10- Chat CoT Zero-Shot "Zhou Mod"

		Model Output								
		A	B	C	D	E	F	T	Other	Grand Total
True Answer	A	72.5%	10.6%	6.3%	7.7%	0.0%	0.0%	0.6%	2.4%	679
	B	8.3%	72.3%	7.8%	9.9%	0.3%	0.0%	0.2%	1.1%	877
	C	6.8%	8.3%	75.4%	6.8%	0.2%	0.0%	0.1%	2.5%	844
	D	7.8%	7.9%	8.9%	72.3%	0.4%	0.0%	0.1%	2.5%	683
	F	0.0%	0.3%	0.0%	0.0%	0.0%	76.1%	21.6%	2.0%	305
	T	0.0%	0.0%	0.4%	0.0%	0.0%	22.7%	75.7%	1.2%	255
	Grand Total	**675**	**831**	**809**	**690**	**8**	**290**	**267**	**73**	**3643**

Figure 20

Exp 11- Chat CoT Zero-Shot “Zhou”

		Model Output							
		A	B	C	D	F	T	Other	Grand Total
True Answer	A	56.0%	16.0%	20.0%	8.0%	0.0%	0.0%	0.0%	25
	B	3.4%	79.3%	0.0%	13.8%	0.0%	0.0%	3.4%	29
	C	9.7%	12.9%	74.2%	3.2%	0.0%	0.0%	0.0%	31
	D	3.4%	6.9%	6.9%	82.8%	0.0%	0.0%	0.0%	29
	F	0.0%	0.0%	0.0%	0.0%	60.0%	40.0%	0.0%	10
	T	0.0%	0.0%	0.0%	0.0%	25.0%	75.0%	0.0%	8
	Grand Total	**19**	**33**	**30**	**31**	**8**	**10**	**1**	**132**

Figure 21

Example Exchanges with Each Chat Prompt with reasoning steps

In order to demonstrate typical patterns of reasoning, selected examples of intermediate reasoning chains are presented below. The below exchanges exemplify typical patterns in the responses for each prompt.

Example Exchanges for Exp. 7 "Provide Context"

Example exchanges for Exp. 6 "Provide context"			
Question	Model provided reasoning	Model's top completion	Correct answer
Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or decrease. D) will not change.	This question pertains to the relationship between productivity of labor and the price of capital, and how it affects the demand for labor. The answer options suggest different possible outcomes.	C	C
Some economists advocate taxes on land because such taxes: A) do not affect the supply of land. B) increase the supply of land. C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. D) have a positive incentive function.	The question is related to the use of taxes on land as a policy tool in economics. It asks about the potential benefits of such taxes.	C	A

Table 6- Example Exchanges for Exp. 7 "Provide Context"

Example exchanges for Exp. 8 "Provide Definition"

Example exchanges for Exp. 8 "Provide Definition"			
Question	Model provided reasoning	Model's top completion	Correct answer

Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or decrease. D) will not change.	Productivity of labor refers to the amount of output produced per unit of labor input. Price of capital refers to the cost of using capital goods, such as machinery or equipment, in production.	C	C
Some economists advocate taxes on land because such taxes: A) do not affect the supply of land. B) increase the supply of land. C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. D) have a positive incentive function.	Land taxes are a type of tax that is levied on the value of land, rather than on the buildings or other structures that are built on it. The idea behind land taxes is that they can help to encourage more efficient use of land, by discouraging land speculation and encouraging development.	C	A
The marginal rate of substitution: A) may increase or decrease on a given indifference curve, depending on whether the substitution or the income effect is dominant. B) increases as one moves southeast along an indifference curve. C) is constant at all points on the budget line. D) declines as one moves southeast along an indifference curve.	The marginal rate of substitution (MRS) measures the rate at which a consumer is willing to trade one good for another while maintaining the same level of utility. It is the slope of an indifference curve at a given point.	A	D

Table 7-Example exchanges for Exp. 8 “Provide Definition”

Example Exchanges for Exp. 9 "CoT 4-shot"

Example exchanges for Exp. 9 "CoT 4-shot"			
Question	Model provided reasoning	Model's top completion	Correct answer
Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or decrease. D) will not change.	A: Thinking step by step, an increase in labor productivity would increase the demand for labor, as firms would want to hire more workers to take advantage of the increased productivity. However, if the price of capital, which is complementary to labor, also increases, firms may substitute capital for labor, reducing the demand for labor. Therefore, the answer is C, as the change in demand for labor could go either way depending on the magnitude of the two changes.	C	C
Some economists advocate taxes on land because such taxes: A) do not affect the supply of land. B) increase the supply of land. C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. D) have a positive incentive function.	A: Thinking step by step, a tax on land is a tax on a fixed resource, so it does not affect the supply of land. Therefore, we can eliminate option B. The purpose of a tax on land is to encourage efficient use of land by making it more expensive to hold onto unused or underused land. This would improve the allocation of land by shifting it from low-productivity to high-productivity uses, so the answer is C. Option D is incorrect because taxes are generally seen as having a negative incentive function, as they discourage the taxed activity.[6]	C	A

[6] Note here the model seems to get the right answer in its reasoning and then continues, skipping over A, which is the correct answer.

Example Exchanges for Exp. 10 CoT Zero-Shot "Zhou Mod"

Example exchanges for Exp. 10 "CoT Zero-Shot Zhou Mod"			
Question	Model provided reasoning	Model's top completion	Correct answer
Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or decrease. D) will not change.	Step 1: When the productivity of labor increases, the marginal product of labor increases. This means that each worker can produce more output, which increases the demand for labor. Step 2: When the price of capital increases, the cost of using capital increases. This makes labor relatively cheaper compared to capital, which increases the demand for labor. Step 3: Combining these two effects, we can see that the increase in productivity of labor and the increase in the price of capital both lead to an increase in the demand for labor. Therefore, the answer is A) will increase.	A	C
Some economists advocate taxes on land because such taxes: A) do not affect the supply of land. B) increase the supply of land. C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. D) have a positive incentive function.	Step 1: What is a tax on land? A tax on land is a tax levied on the value of land, regardless of any buildings or improvements on it. Step 2: How does a tax on land affect the supply of land? A tax on land does not affect the supply of land because the supply of land is fixed. Land cannot be created or destroyed, so the amount of land available is constant.[7] Step 3: How does a tax on land	C	A

[7]Note that this is correct! If the model stopped here and chose A, it would be right.

	improve the allocation of land? A tax on land can improve the allocation of land by shifting it from low-productivity to high-productivity uses. When land is taxed, landowners have an incentive to use their land more efficiently and productively. This can lead to a more efficient use of land resources and a better allocation of land to its most productive uses. Step 4: What is the incentive function of a tax on land? A tax on land has a positive incentive function because it encourages landowners to use their land more efficiently and productively. This can lead to a more efficient use of land resources and a better allocation of land to its most productive uses. Step 5: What is the answer? The correct answer is C) improve the allocation of land by shifting it from low-productivity to high-productivity uses.		

Example Exchanges for Exp. 11 CoT Zero-Shot "Zhou Mod"

Example Exchanges for Exp. 11 CoT Zero-Shot Zhou Mod"			
Question	Model provided reasoning	Model's top completion	Correct answer
Suppose the productivity of labor increases and at the same time the price of capital, which is complementary to labor, increases. As a result, the demand for labor: A) will increase. B) will decrease. C) may either increase or	Step 1: Identify the relationship between productivity of labor and demand for labor. When productivity of labor increases, it means that each worker can produce more output per unit of time. This leads to an increase in the marginal product	A	C

decrease. D) will not change.	of labor, which in turn increases the demand for labor. Step 2: Identify the relationship between price of capital and demand for labor. Capital and labor are complementary inputs, which means that they are used together to produce goods and services. An increase in the price of capital means that it becomes relatively more expensive compared to labor. This leads to a substitution effect, where firms may choose to use less capital and more labor to produce the same level of output. This increases the demand for labor. Step 3: Combine the two effects to determine the overall effect on demand for labor. The increase in productivity of labor leads to an increase in demand for labor, while the increase in price of capital leads to an increase in demand for labor. Both effects are positive, which means that they reinforce each other. Therefore, the overall effect on demand for labor is an increase. Answer: A) will increase.		
Some economists advocate taxes on land because such taxes: A) do not affect the supply of land. B) increase the supply of land. C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. D) have a positive incentive function.	Step 1: Read the question and the answer choices carefully. Step 2: Identify the key information in the question. The question is asking why some economists advocate taxes on land. Step 3: Examine each answer choice and eliminate any that are clearly incorrect. - A) do not affect the supply of land. This answer choice may be	C	A

	true, but it does not explain why some economists advocate taxes on land. Eliminate. - B) increase the supply of land. This answer choice is incorrect. Taxes on land do not increase the supply of land. Eliminate. - C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. This answer choice is a possible reason why some economists advocate taxes on land. Keep. - D) have a positive incentive function. This answer choice may be true, but it does not explain why some economists advocate taxes on land. Eliminate. Step 4: Review the remaining answer choice and make sure it answers the question. - C) improve the allocation of land by shifting it from low-productivity to high-productivity uses. This answer choice explains that some economists advocate taxes on land because they believe it will improve the allocation of land by encouraging it to be used for high-productivity purposes. Step 5: Select the correct answer. The correct answer is C) improve the allocation of land by shifting it from low-productivity to high-productivity uses.		

(Consider This) In 1996 McDonald's introduced its Arch Deluxe hamburger, which failed to catch on with the public and was subsequently dropped from the menu. This failure illustrates the idea of: A) consumer sovereignty. B) technological change. C) downsloping demand. D) specialization.	C) downsloping demand.[8]	C	A

Table of Figures

[8] Note that the model does not provide any reasoning through the problem. In about half of all completions in Exp 11, the model provides no reasoning or only provides reasoning after choosing an answer.

Table of Tables