Scaling Open-ended Survey Responses Using LLM-Paired Comparisons

Matthew DiGiuseppe
Associate Professor
Leiden University
mdigiuseppe@gmail.com

Michael Flynn
Professor
Kansas State University
meflynn@ksu.edu

January 25, 2025

Abstract

Survey researchers rely heavily on closed-ended questions to measure latent respondent characteristics like knowledge, policy positions, emotions, ideology, and various other traits. While closed-ended questions ease analysis and data collection, they necessarily limit the depth and variability of responses. Open-ended responses allow for greater depth and variability in responses, but are labor-intensive to code. Large Language Models (LLMs) can solve some of these problems, but existing approaches to using LLMs have a number of limitations. In this paper, we propose and test a pairwise comparison method to scale open-ended survey responses on a continuous scale. The approach relies on LLMs to make pairwise comparisons of statements that identify which statement "wins" and "loses". With this information, we employ a Bayesian Bradley-Terry model to recover a 'score' on a the relevant latent dimension for each statement. This approach allows for finer discrimination between items, better measures of uncertainty, reduces anchoring bias, and is more flexible than methods relying on Maximum Likelihood Estimation techniques. We demonstrate the utility of this approach on an open-ended question probing knowledge of interest rates in the US economy. A comparison of 6 LLMs of various sizes reveals that pairwise comparisons show greater consistency than zero-shot 0-10 ratings with larger models (> 9-billion parameters). Further, comparison of pairwise decisions are consistent with high-knowledge crowd source workers.[1]

[1]Corresponding Author: Matthew DiGiuseppe - mdigiuseppe@gmail.com

Public opinion scholars and survey researchers are often interested the latent traits of individual respondents like political knowledge, literacy, comprehension, engagement, ideology, emotional reactions, psychological characteristics, and values. Due to convenience, most scholars use closed-ended questions independently or in a scale to measure these traits. However, closed-end responses come with several undesirable properties. They introduce ceiling and floor effects. They can also introduce measurement error from false or inadvertent responses, and force an assumption of linearity on scales. Most importantly, they reduce the richness of responses and often introduce concepts that would otherwise be apparent to respondents.

Alternatively, open-ended responses offer an unstructured and more flexible alternative that allows for in-depth and detailed responses that captures substantive uncertainty and important qualifications. However, open-ended responses have traditionally required costly human coders before they are usable in statistical analysis (Lazarsfeld, 1944; Geer, 1991; Converse, 1984; Haaland et al., 2024; Andre et al., 2024; Roberts et al., 2014). Given a large number of potentially lengthy responses, coding the various dimensions for all respondents can be labor intensive. More often than not, this process also reduces these high dimension data to linear discrete scales that reintroduce some issues of closed-ended questions.[2]

Advances in text as data methods (Roberts et al., 2014) in the past 10-15 years have opened to door to more automated text analysis. However, these methods are only useful in certain applications that do not require background knowledge to make judgments. However they are particularly useful in identifying differences in sophistication or word use among groups (Kraft, 2024; Zollinger, 2024). Advances in Large Language Models (LLMs) have significantly reduced the cost of annotating and scaling open-ended questions further (Rathje et al., 2024; Heseltine and Clemm von Hohenberg, 2024; Mens and Gallego, 2023). Notably, frontier LLMs come with the added benefit of strong domain knowledge often exceeding the abilities of crowd workers or undergraduate research assistants (Gilardi et al., 2023; Bermejo

[2]Beyond cost concerns, scholars have broader methodological concerns. Roberts et al. (2014) provides a nice discussion of the benefits and limitations of open-ended responses.

et al., 2024; Ludwig et al., 2024; Ornstein et al., 2025). However, using LLMs in place of research assistants for scaling responses typically generates unanchored scores, which also lack corresponding estimates of uncertainty. Further, LLMs vary in their output both across different LLMs and even across different versions of the same LLMs in classification and annotation tasks (Barrie et al., 2024). As such, there is still substantial uncertainty as to the suitability of LLMs for helping survey researchers.

In this paper, we introduce a framework for using LLMs to scale latent dimensions in open-ended responses with *pairwise comparisons*. First, researchers prompt an LLM to make zero-shot, independent pairwise comparisons of two random responses and indicate which is more closely aligned with the latent concept, or if they are too similar to distinguish. After collecting $\boldsymbol{N}$ comparisons, researchers use the results of the pairwise comparisons to fit a Bayesian Bradley-Terry (BT) model to generate a latent variable for respondent knowledge that we then use to scale and rank the individual respondents (Bradley and Terry, 1952; Davidson, 1970). The estimate of this latent dimension and the error surrounding the estimate can then be used in subsequent analyses.

Just as in using pairwise comparisons with human raters (Carlson and Montgomery, 2017), LLM pairwise comparisons have several advantages over placement on an ordinal or discrete scale. Importantly, the approach produces an estimate that can be interpreted as a position on a latent scale relative to other observations in the dataset rather than the unanchored responses generated by human and LLM coders prompted to place a statement on a scale. Next, because of the large-number of pairwise comparisons possible, it is easier to recover an estimate closer to the true parameter, with smaller errors, even if there is considerable noise in initial pairwise comparisons. This subsequently allows researchers to recover more nuanced differences among observations. The credible intervals around the latent variable estimates can also be used to incorporate the inherent uncertainty associated with any given observation or measure into subsequent analyses. Further, by forcing a binary ordering, rather than a ranking, unobserved biases that do not influence rank order of

pairwise comparisons washes out. Further, as we show below, this approach allows for the generation of latent estimates of respondent knowledge even when we have sparse comparisons, using readily available software, reducing dependence upon customized packages for estimating pairwise comparison models.

Others have demonstrated the utility of using pairwise comparisons made by crowd source workers to scale latent concepts in text. Unlike crowd source workers used to make pairwise comparisons (Carlson and Montgomery, 2017), using LLMs can increase the potential application of this approach. LLMs enable pair-wise comparisons on large datasets (that often exceed N=1000) that would require many coder-hours. Additionally, frontier LLMs have strong domain knowledge across a variety of subjects—including economics and finance—which we use in our illustration (Yang et al., 2024; Hultberg et al., 2024). LLMs combined with pairwise comparisons enable researchers to scale tasks that previously required expert programmers.

Notably, Wu et al. (2023a) are the first to identify the utility of pairing LLMs with pairwise comparisons in the evaluation of sentiment in 'tweets' in the context of a chain of thought framework. However, their work does little to validate this approach in comparison to other prompting strategies, assess performance against a human benchmark or against closed-ended questions. Further, their approach does little to test the domain knowledge of LLMs needed to code many concepts embedded in open-ended questions.

Our contribution is to demonstrate the utility of this approach within the realm of open-ended survey responses and with zero-shot prompting. To do so, we illustrate how pairwise comparisons can create an indicator of 'interest rate knowledge' from an original open-ended survey question asking respondents how interest rates are set in the United States economy. We first show that zero-shot pairwise comparisons of LLMs are consistent with those made by 'close-to-expert' crowd-source workers. We then demonstrate the benefits of the approach by comparing the BT estimates to closed-ended responses to show they have strong face validity. We then demonstrate that the pairwise comparison estimates are not sensitive to the choice

of LLM, above a certain parameter threshold, and that they produce more consistent results than pure unanchored rankings from LLMs.

Our application demonstrates the usefulness of the approach and how it performs relative to zero-shot ratings. It also shows that LLMs are useful beyond sentiment analysis and classification. The embedded domain knowledge in LLMs can be used, in some cases, in place of close-to-expert coders. As such, it allows for scaling of concepts that are outside the reach of student or crowd sourced workers. Importantly, we show that this advantage grows with the capabilities of LLMs. We end with a discussion of the limitations of the approach and a warning that survey respondents are also growing more reliant on LLMs.

LLMs & Pairwise Comparisons

The utility of using the embedded knowledge in LLMs to annotate, classify, and scale text has been widely demonstrated in a variety of social science fields (Heseltine and Clemm von Hohenberg, 2024; Mellon et al., 2024; Lincan Li, 2024; Gilardi et al., 2023; Törnberg, 2024). Applications are diverse. Mellon et al. (2024) uses LLMs to code the most important problem identified in open-ended responses. A number of studies use LLMs to code the sentiment or other characteristics of tweets or news reports (Gilardi et al., 2023; Törnberg, 2024; Heseltine and Clemm von Hohenberg, 2024; Ornstein et al., 2025). Others Rathje et al. (2024) use LLMs to scale emotions in text data (tweets and reddit comments) for psychological research. They focus largely on identifying sentiment, discrete emotions, and moral foundations.

These contributions demonstrate that LLMs often meet or exceed crowd-source workers on annotation and scaling tasks. However, these measures still have several well-known limitations that also apply to human-coded data. First, in scaling tasks, the responses of LLMs are unanchored. As such, it is not apparent what differentiates the maximum and the minimum values, or various other items on a scale. Second, LLMs, like the human mind, are black boxes. There is likely to be unobserved bias that influences scale placement.

Third, the LLM generated responses do not produce uncertainty estimates. Consequently, subsequent models can't discern between differences on a scale that are in fact distinct or differences that appear distinct but are in fact statistically indistinguishable. Beyond these issues, the enthusiasm of reducing the cost of dimension reduction is further dampened by concerns about replicability across LLM models and within models overtime. In a series of tests, Barrie et al. (2024) show that the variance across and within models across time is 'unacceptably high.'

Given these issues, can survey researchers feel confident in using LLMs to scale open-ended questions? LLMs are likely to improve and grow more consistent and their biases will become more transparent, but until such time researchers are faced with the task of finding alternative methods for dealing with these issues. Some of the challenges of using existing models can be improved by using LLMs to make pairwise comparisons of text and then using these pairwise judgments to generate estimates of the desired latent traits with a Bradley-Terry model.

Pairwise comparisons are not new to social science research. They have been employed to measure political sophistication (Benoit et al., 2019), persuasiveness of political arguments (Loewen et al., 2012), and the dimensions of government actors (Zucco Jr et al., 2019). In fact, pairwise comparisons are particularly useful for measuring subjective constructs. Narimanzadeh et al. (2023) demonstrate that, in scaling subjective constructs, using crowd-sourced pairwise comparisons and than estimating individual scores outperforms majority voting methods to scaling these concepts.

The approach yields several benefits for survey researchers over traditional scaling techniques. The first is that pairwise comparisons produce a relative assessment of each response. A well-known measurement issue with scaling text is that the responses are unanchored.[3] This means that there is bound to be a lack of clarity on what each value on a scale means relative to the subjective construct the researcher wants to scale. While training coders can

[3]By unanchored we mean that coders lack a explicit reference points, or benchmarks that allow them to differentiate between labels on a scale.

go some way to alleviating this concern, the problem can not be fully resolved. When using human coders, biases may also emerge emerging from differences in the order of appearance or variability across coders that have different understandings of the scale. It is still unclear what problem this poses for LLMs. Conceivably, while prompt instructions remain constant across one-shot calls, the actual prompt may change based on the added human written response that is unique to each respondent. As such, it is difficult to know how this impacts how the LLM places items on a scale across numerous calls. Further, human coders can eventually converge on an anchor by completing multiple responses assuming that they consistently apply coding rules. For LLMs this convergence is not easily achieved. Each zero-shot or even multiple shot calls relies on a new call of the model which entails a 'clean slate' requiring a repetition of instructions. Further, if a long-chain of thought could be achieved (at increasing cost in input tokens), some worry that LLMs exhibit recency bias (Peysakhovich and Lerer, 2023). Further, efforts at multi-shot prompting relies on picking examples in an attempt to anchor scaling. Yet, it remains unclear how the choices of these examples impact the final dataset given the black box nature of LLMs. In sum, the subjectivity inherent in placing items on scale likely generates error that is not directly observable to the researcher.

As Carlson and Montgomery (2017) argue, pairwise comparisons help ameliorate this unobserved bias when using human raters. If one rater tends to rate higher or lower on a scale, this is irrelevant because the forced comparison requires a single cut point. Where on the scale an item is placed is not relevant for the final estimate unless it changes which item "wins" or "loses" a comparison. Similarly, there is concern that LLMs are inconsistent (Barrie et al., 2024) and thus a similar logic should apply. If different LLMs or differences in the prompt language (or language within a piped in response) lead to different placement on a scale based on some characteristic, this should only be relevant when the separate individual ratings move in opposite directions on the scale and far enough to change the outcome. As long as any bias moves ratings in the same direction, the bias is irrelevant

when employing pairwise comparisons. For example, let's assume that a human coder or an LLM rates a knowledge response higher (lower) because the grammar and spelling are flawless (sloppy). This may lead to an higher (lower) score for responses with perfect (flawed) writing. On a 0-10 scale this might result in real differences in responses that have the same knowledge but are presented in different ways. In the pairwise comparison framework, this bias is only relevant if it results in a flip of rank-ordering of the pair changing the winning response to the a losing response. The bias, therefore, has a higher threshold to meet to influence inferences in a pairwise comparison outcome relative to simple scaling.

The second benefit of the pairwise comparison approach is that it naturally incorporates uncertainty into the estimates, allowing us to recognize when observed differences may not be statistically meaningful. While some items can be clearly distinguished from others because of real differences, in many cases, the difference between two statements is ambiguous and should be treated as such. In contrast, a simple rating-based approach may produce distinctions that are seemingly meaningful as a one-point difference on a 10-point scale—that downstream models might treat as reliable information. By explicitly incorporating uncertainty into subsequent analyses (Blackwell et al., 2017), we reduce the risk of over-interpreting these minor, potentially irrelevant differences, by incorporating information of uncertainty.

A third benefit of pairwise comparisons approach is that it makes use of fine grained and nuance differences in a way that are difficult to implement with a fixed-scale. Nuanced differences are difficult to map on a scale without placing large cognitive demands on human raters and potentially exaggerates measurement bias with LLMs (Benoit et al., 2019). Consider that the cognitive demands needed to determine the difference between a 65 and 66 on a 100-point scale with out a clear reference point or, from the perspective of a human rate, multiple cases that were coded previously. Then consider the difficulty of determining which statement is 'better' than the other even if the difference is nuanced. The later is inherently easier and quicker to assess. Similarly, LLMs, while frequently impressive, may have a sim-

ilarly hard time making consistent judgments on a large scale given it would be difficult to have clear instructions for differences on a scale that would allow for fine grained differences. By reverting to multiple pairwise comparisons, LLMs, like humans coders, can be utilized to produce fine grained differences among observations with pairwise comparisons.

Until recently, pairwise comparisons have drawn on crowd workers, students, or experts to make comparisons. While the benefits of pairwise comparisons are clear from a measurement perspective (Carlson and Montgomery, 2017), the cost of employing crowd-workers or RAs to engage in numerous pairwise comparisons is likely responsible for its infrequent adoption. LLMs easily remedy this concern. Scholars have demonstrated that the embedded knowledge of LLMs is sufficient to carry out these comparisons with significantly lower cost. For example, Wu et al. (2023b) use LLMs pairwise comparisons to recreate latent ideology scores of US Senators. They provide LLMs with only the Senators names and prompting LLMs to indicate which is more liberal, conservative, supportive of gun rights, etc without providing any additional text. The estimates generated from these comparisons map nicely on to the often used DW-NOMINATE scores (Poole and Rosenthal, 2000). Di Leo et al. (2024) use a similar approach to estimate the ideology of European parties. The expertise of LLMs extends beyond political judgments. LLMs have consistently demonstrated a strong embedded knowledge of economic and psychological concepts.

In sum, LLMs have the embedded knowledge to replace expert coders in classification and scaling tasks. However, they are often used in a way that fails to address the subjectivity of rankings and lacks a measure of corresponding uncertainty. Below, we show that pairwise comparisons offer a better alternative to scaling approaches and can be done with reasonable costs. However, we also show where they may fail to produce acceptable data.

Pairwise Comparison Work Flow

Figure 1 outlines the workflow to take individual open-ended responses and create estimates to be used in subsequent analyses. After collecting open-ended survey responses, researchers first create pairs of responses. In our analysis, we started with each response and randomly paired it with 20 responses from the dataset without replacement. Once the pairs have been assigned, researchers must develop a prompt that a) clearly outlines the task the LLM will perform, provides each of the open ended responses, and requests a response to identify which response best aligns with the concept or if they are indistinguishable. The prompt can then be used in calls to an LLM API or, for smaller models, run locally. Once the LLM returns the judgments, a Bradley-Terry (BT) model can be fitted to the responses.[4] Following the estimation of the BT model, we retrieve the median estimate and the errors. If the researcher is interested in using the estimates in downstream analyses either as an outcome or predictor then the posterior distributions from the Bradley-Terry models can be incorporated into subsequent analyses as needed, thereby allowing the researcher to incorporate the uncertainty from the Bradley-Terry models directly into additional models.

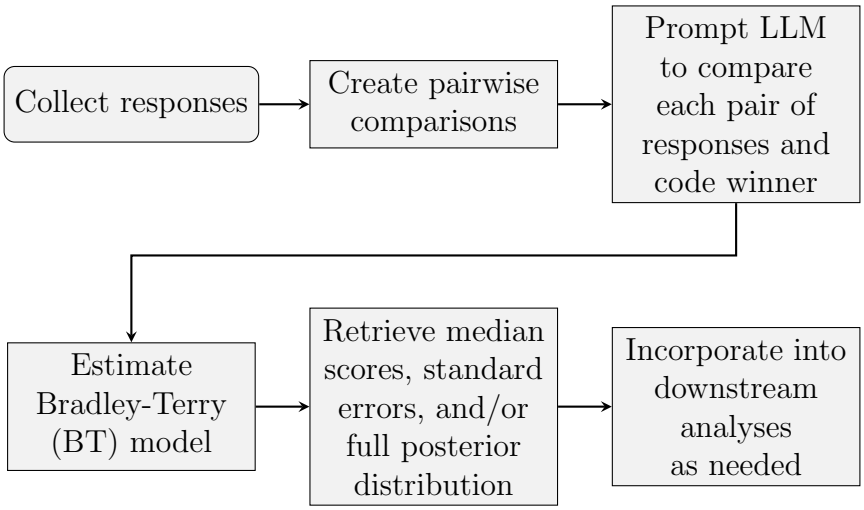

Figure 1: Workflow Diagram of the Estimation Process

[4]If you ask the LLM to code ties, you can simply recode a winner randomly (as we do here) or instead fit a Bradley-Terry-Davidson model (Davidson, 1970) that accommodates ties.

MLE or Bayesian Estimation?

Once we have obtained the paired responses the LLM will make judgments about the domain knowledge contained within each response. The instructions ask the LLM to decide if 1) response #1 wins, if response #2 wins, or if the two responses are equivalent (i.e. a tie). Once we have the scores for each of the pairwise comparisons we can then fit a Bradley-Terry model to generate a latent variable capturing the underlying trait the researcher cares about (e.g. knowledge, emotion, etc.).

Traditionally Bradley-Terry models have been fit with Maximum Likelihood Estimation (MLE) methods. Here we adopt an alternative Bayesian framework. While others have utilized Bayesian methods for estimating similar models the approach is not yet widespread (see Carpenter, 2018; Mattos and Ramos, 2022; Kaye and Firth, 2022; van Paridon et al., 2023). The Bayesian approach has several desirable properties relative to traditional MLE approaches. First, MLE methods can be fast, but the researcher's ability to use MLE rests on the assumption that they have pairwise comparisons for all possible items or responses (Mattos and Ramos, 2022; Kaye and Firth, 2022). Where the researcher does not have pairwise comparisons for every item, MLE methods will fail to converge. Sometimes this is within the researcher's control, but in many cases the researcher may not be able to generate a complete list of comparisons.

Further, in cases where the number of items to be ranked is very large, MLE methods may struggle with the computational complexity. Alternatively, while Markov Chain Monte Carlo sampling methods may sometimes run more slowly, they are able to generate estimates of the desired parameters even in cases where there are a large number of items to rank, and where not every item is paired with every other item. However, modern software for fitting Bayesian models, like Stan and its Hamiltonian Monte Carlo sampling procedures, have greatly reduced the time it takes to fit even more complex Bayesian models. Where speed might still be an issue, the choice to reduce the number of pairwise comparisons can still save the researcher time, though at the expense of larger errors in the posterior distributions

of the latent parameters.

Bayesian approaches also provide us with several options for dealing with the inherant uncertainty associated with the latent estimates derived from the Bradley-Terry models. Packages like `brms` include functions like `me()` that can be used to account for measurement error in particular predictors. Alternatively, researchers can simply sample from the posterior distributions of the Bradley-Terry estimates and run multiple iterations of downstream models to directly incorporate the uncertainty into their estimates.

Finally, while there are several `R` packages that can estimate Bradley-Terry Models, many rely on MLE for estimation, or they may depend upon package maintainers to keep functions updated and working.[5] The approach we outline here demonstrates how researchers can use Stan and `brms` to estimate Bradley-Terry models as multimembership mixed effects models. Depending on the structure of the pairwise comparison data, these models can be estimated using binomial or logistic regression and are generally easy to fit without relying on customized Bradley-Terry packages (for example see Firth, 2005; van Paridon et al., 2023).

Illustration: Scaling Economic Knowledge

To demonstrate the utility of the approach, we draw on data collected by DiGiuseppe et al. (2024) that measures knowledge about monetary policy. The data, collected on the Prolific platform, prompted respondents with the following question: "In a few sentences and without looking it up, can you explain how interest rates (i.e. the cost of borrowing money to buy a house or car) go up or down in the US economy?".[6] Respondents were asked to reply in 2-3 sentences. The aim of the study is to explore how knowledge about interest rates influences individual assessment of the Federal Reserve and support for its independence. This dataset

[5] Examples include the `BradleyTerry, BradleyTerry2, BradleyTerryScalable`, and `bpcs` packages. For more information see Firth (2005); Turner and Firth (2012); Kaye and Firth (2022); Mattos and Ramos (2022).

[6] Respondents were asked at two points later in the survey if they looked up the answers to knowledge questions. We removed those that confirmed they sought assistance from the data. We discuss below that we found several respondents used LLMs to answer the open-ended questions.

is useful, for our purposes, in that it collects both open and closed-ended questions relating to knowledge of the Federal Reserve and that it is a potentially difficult task for both human coders and LLMs.

We carry out this exercise with LLMs of various sizes and a mix of proprietary and open-source models (see Table 1). The use of multiple models is useful for testing the limits of the approach and to compare the consistency across models. These models include frontier LLMs (GPT-4o, GPT-4o Mini, Llama 3.1:405b) and two smaller large language models, Llama 3.1:7b and Google's Gemma2:2b and Gemma2:9b. We use API calls for the OpenAI models and the Large LLama 3.1 model. We run the smaller models locally using Ollama and the the R package 'rollama' to call on llama 3.2 locally (Gruber and Weber, 2024) on an Apple Macbook with an M3 Pro chip and 36GB of memory.

Model	**Parameters**	**Open Source**	**Access**	**Proportion Ties**	**Transitivity Score**
LLaMA 3.1:405B	405 Billion	Yes	API	0.024	99.1
GPT-4o	Not specified	No	API	0.009	96.1
GPT-4o mini	Not specified	No	API	0.040	96.7
LLaMA 3.1 8B	8 Billion	Yes	Local	0.062	96.0
Gemma:2B	2.6 Billion	Yes	Local	0.297	94.6
Gemma:9B	9 Billion	Yes	Local	0.065	98.2

Table 1: Properties of LLMs used in Illustration

In line with the workflow we described above, we paired each response with 20 other randomly selected responses, ensuring that $response_i \neq response_j$. This results in approximately 30–40 total comparisons for each response as any given respondent appears 20 times as $response_i$ and anywhere from 7 to 33 times as $response_j$. In total this yields a little over 30,000 total comparisons or lines in the data. Following the workflow, we prompt each LLM with the following prompt:

Pairwise Comparison LLM Prompt

"You are an expert in US economic policy. Your task is to determine which of two given statements contains a more knowledgeable response to the following question: "In a few sentences and without looking it up, can you explain how interest rates (i.e. the cost of borrowing money to buy a house or car) go up or down in the US economy? Respond with either '1' if the first statement contains more knowledge, '2' if the second statement contains more knowledge, or '0' if they are equal or incomparable. Compare these two statements and respond with 1, 2, or 0:", 1: [Statement 1], 2: [Statement 2]. Only reply with the integer 1, 2, or 0"

Table 1 shows that there are relatively few ties in most models. Although the smallest model appears to be less decisive, finding that 30% of all comparisons are tied. Beyond the ties, we see that the LLMs show consistency in their judgements. We examined the consistency of judgments across triplets where we had judgments for A, B, and C and examined how often there was a violation of transitivity. We see that no model is perfect. Yet, the transitivity scores ([total triplets - triplets with violations] / [total triplets]) demonstrate strong consistency.

Once we have the LLMs' judgments for these 30,000+ comparisons, we estimate a Bradley–Terry Model for the comparisons of each LLM. The first step is to resolve any ties in the data. Resolving these ties can be handled in a variety of ways, but here we simply randomly assign wins between the two options.

Once the ties are resolved we fit a Bradley-Terry model to estimate the following:

$$Pr(i \text{ beats } j) = \delta_i - \delta_j \tag{1}$$

The data are organized with two columns for $respondent_i$ and $respondent_j$ wherein the values in each vector correspond to an individual respondent ID number. A third column contains a binary indicator denoting whether or not $respondent_i$ won the comparison, with

a 1 indicating yes and a 0 otherwise.

Rather than using a customized package to estimate the Bradley-Terry Model, we fit a multi-membership mixed effects logit model using the `brms` package and Stan programming language (Bürkner, 2017, 2018; Stan Development Team, 2024; Gabry et al., 2024).

The model is simply a varying intercept model wherein we estimate separate intercepts for each individual respondent. Each pairwise comparison is simply subtracting the varying intercept for respondent$_j$ from respondent$_i$, as shown in Equation 1. Once each model is run we obtain varying intercept estimates for each individual respondent that allow us to rank order them according to their knowledge and ability to accurately answer the questions posed.[7]

Human-LLM Comparison of Comparisons

Our own prompting indicates that LLMs have a good understanding of the central process of setting interest rates. Beyond this, there is some indication that LLMs have strong domain knowledge in economics and finance but systematic evidence is limited (Yang et al., 2024; Hultberg et al., 2024). As such, we validate the LLMs against human coders in this specific task.

To create a human benchmark on a topic that most have limited understanding of, we recontacted 20 respondents from our original sample that offered an expert-level answer on our initial survey. Based on the results of the LLM ratings, we selected the top-20 responses in terms of knowledge. We then reviewed those responses to verify that they in fact provided an 'expert-level' response to the question on interest rates. We then recontacted these crowd-workers, via the Prolific Platform, and asked them to take part in a pairwise comparison exercise to rank several 20 pairs of responses randomly drawn from our dataset.[8] We ended up with 300 pairs of human rated pairwise comparisons. We then prompted the LLMs to compare these same pairs. Both Humans and LLMs were provided similar prompts.

[7] When estimating the models we use broadly regularizing priors to facilitate model convergence.
[8] As we note below, we screened out those that have used LLMs themselves.

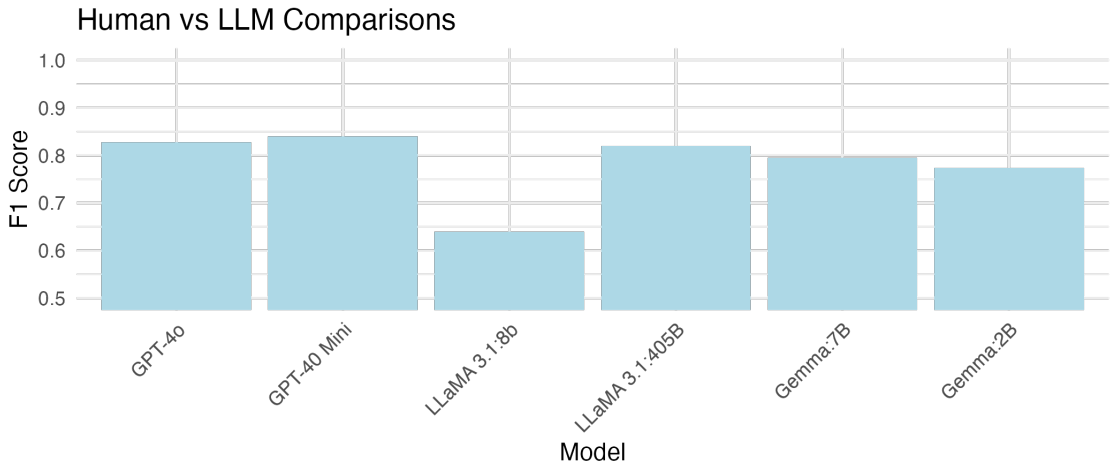

Figure 2: **Comparisons Human-LLM - F1 Score:** This figure reports the F1 score for Human-LLM comparisons of a subset of the original dataset of 300 pairwise comparisons.

Figure 2 reports the F1 score comparing the human raters with the LLMs for the small set of responses coded by the human coders.[9] The frontier models (GPT4o, GPT4o mini, and LLaMa 3.1:405B) achieve an F1 score above 0.8 compared to human coders, indicating a strong alignment with human assessments. The smaller models generate mixed results with Gemma: 7B and Gemma: 2b providing strong scores and Llama 3.1: 8b lagging behind. The F1 should be interpreted in context. In this exercise, we asked the respondents to select the profile with a "small preference" rather than indicate a tie to ease interpretation. Many of the human responses are going to have similar levels of knowledge. As such, there are likely to be a high level of ambiguous cases in the dataset. Compared to classification tasks which may have clearer borders between cases, the task here will naturally lead to more disagreement. Still, we find a strong F1 score. This gives us confidence that the underlying task generating the data is one that is well handled by LLMs.

[9]The F1 score reports a balance between the model's precision (correctness of its positive predictions) and recall (ability to identify all positive instances), representing a combined measure of the model's accuracy on the positive class. F1 Score $= 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$. Recall $= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$. Precision $= \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$.

Comparison with Closed Ended Responses

Thus far, we've seen that the pairwise comparison approach is well-suited for frontier models and the underlying task of comparisons is closely related to human decision making. We take an additional step to compare the estimates within respondent by comparing our BT estimates of individuals respondents to their own responses on closed-ended questions in the same survey. We compare these responses with the BT estimates derived from the pairwise comparisons of the largest open-source model, Llama 3.1 with 405 Billion parameters given we have a preference for open-source models.

In Figure 3, we plot the BT estimates in order but classify them based on how a respondent responded to the question "How familiar are you with the following US institution: The Federal Reserve". We see, in line with expectations, that the those that "have a fairly accurate idea of the duties of the institution" and "have an approximate idea of the duties of the institution" score higher on the scale. Those that "only know the institution by name" or "don't know" the institution rank consistently lower on the scale. Next, we turn to factual questions about the Federal Reserve. Given the large role of the Federal Reserve in setting interest rates, knowledge about the institutional structure of the Federal Reserve should be strongly related to knowledge about interest rates. Figure 4 presents the mean BT estimate by correctness of three factual questions in which respondents had to pick the correct answer from 4 choices. We see that those who could not identify which institution in the US government sets interest rates, who appoints the Fed. Chair, and those who could not identify the Fed Chair have significantly lower BT estimates. This gives us confidence that the BT estimates align with the the underlying construct - knowledge about the Federal Reserve and Interest Rates.

Comparing LLMs

We now proceed to compare the final BT estimates from the pairwise comparisons of each LLM. First, Figure 5 plots the 95% credible intervals around each respondent's latent es-

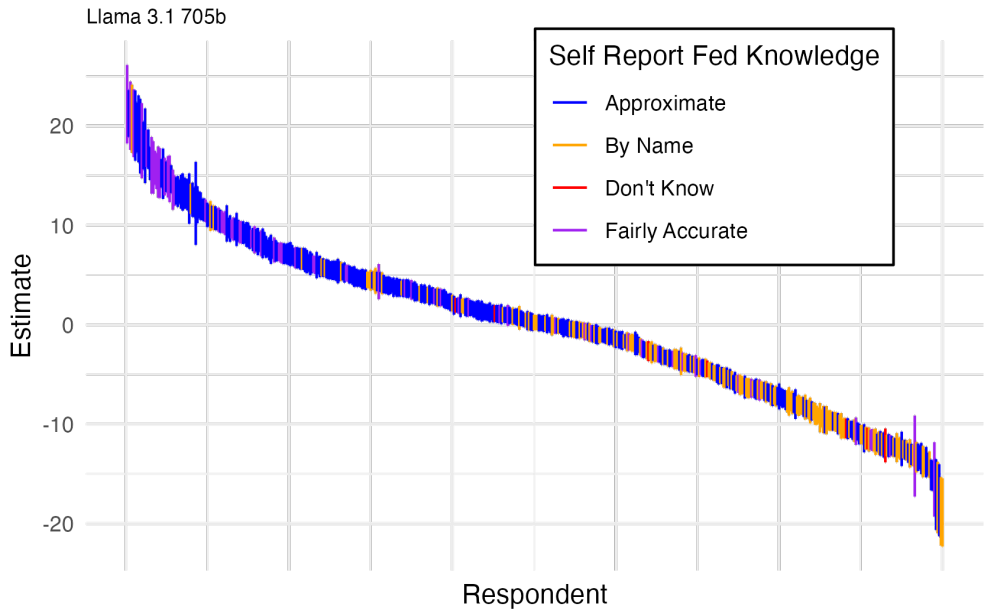

Figure 3: **BT Estimates by Self Reported Knowledge of the Fed:** Here we randomly selected 200 responses (for visibility) and plot the BT estimates in order of knowledge. The error bars of the estimates are colored based on responses to query about a respondents self reported knowledge of the Federal Reserve.

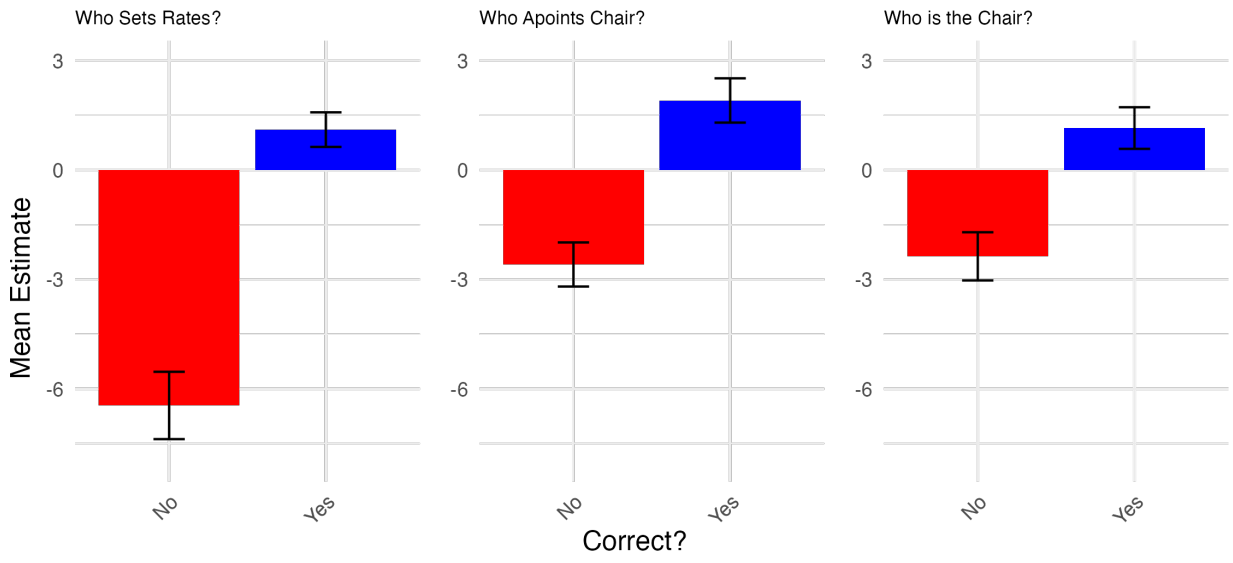

Figure 4: Mean BT Estimate by Correct and Incorrect Answers to Factual Questions about the Fed

timate from each of the 6 models. In each figures, we sorted latent estimates according to the median of their posterior distributions from the lowest to highest respondent knowledge ability within each model. The plots demonstrate a similar pattern across all but the smallest model. Further, we see that smallest model, Gemma 2:2B, has a smaller range of values, indicating that it has a more difficult time distinguishing "winners" from "losers". This corresponds with the greater number of ties Gemma2:2B identified.

The plots can show us the general distribution of the latent estimates but can't tell us if these distributions correlate. That is, do individual respondents who score high/low in the models from one LLM score similarly high/low in the models from a different LLM? Table 2 presents the correlations of the final BT scores for each model. Outside the smallest model, Gemma 2:2B, we see that the final correlation of all model's scores exceed 0.877. Correlation among the largest models (GPT4o and Llama 3.1:405B) exceeds 0.95. This suggests that, with this specific task, the LLMs are largely in a agreement about what constitutes a highly knowledgeable answer.

Table 2: **Correlation of BT Estimates of Pairwise Comparisons by LLM**

	Gemma 2:2B	Gemma 2:9B	GPT 4o mini	GPT 4o	Llama 3.1:405B	Llama 3.1:7B
Gemma 2:2B	1.000	0.702	0.647	0.619	0.625	0.702
Gemma 2:9B	0.702	1.000	0.925	0.914	0.927	0.894
GPT 4o Mini	0.647	0.925	1.000	0.928	0.951	0.907
GPT 4o	0.619	0.914	0.928	1.000	0.953	0.877
Llama 3.1:405B	0.625	0.927	0.951	0.953	1.000	0.902
Llama 3.1:7B	0.702	0.894	0.907	0.877	0.902	1.000

For comparison, we also prompted our LLMs to engage in zero-shot numerical ratings of individual statements. We asked each LLM to place each statement on scale reflecting the knowledge about interest rates from completely incorrect or irrelevant (0) to highly knowledgeable and accurate (10).[10] Figure 6 plots the distribution of these results and Table

[10]The prompt reads as follows: "You are an expert in US economic policy. Your task is to rate the given statement on a scale of 0-10 based on how knowledgeable it is in response to the following question: In a few sentences and without looking it up, can you explain how interest rates (i.e. the cost of borrowing money to buy a house or car) go up or down in the US economy? Rate the following statement on a scale of 0-10, where 0 is completely incorrect or irrelevant, and 10 is highly knowledgeable and accurate: [statement]. Respond ONLY with a single integer from 0 to 10, with no additional text."

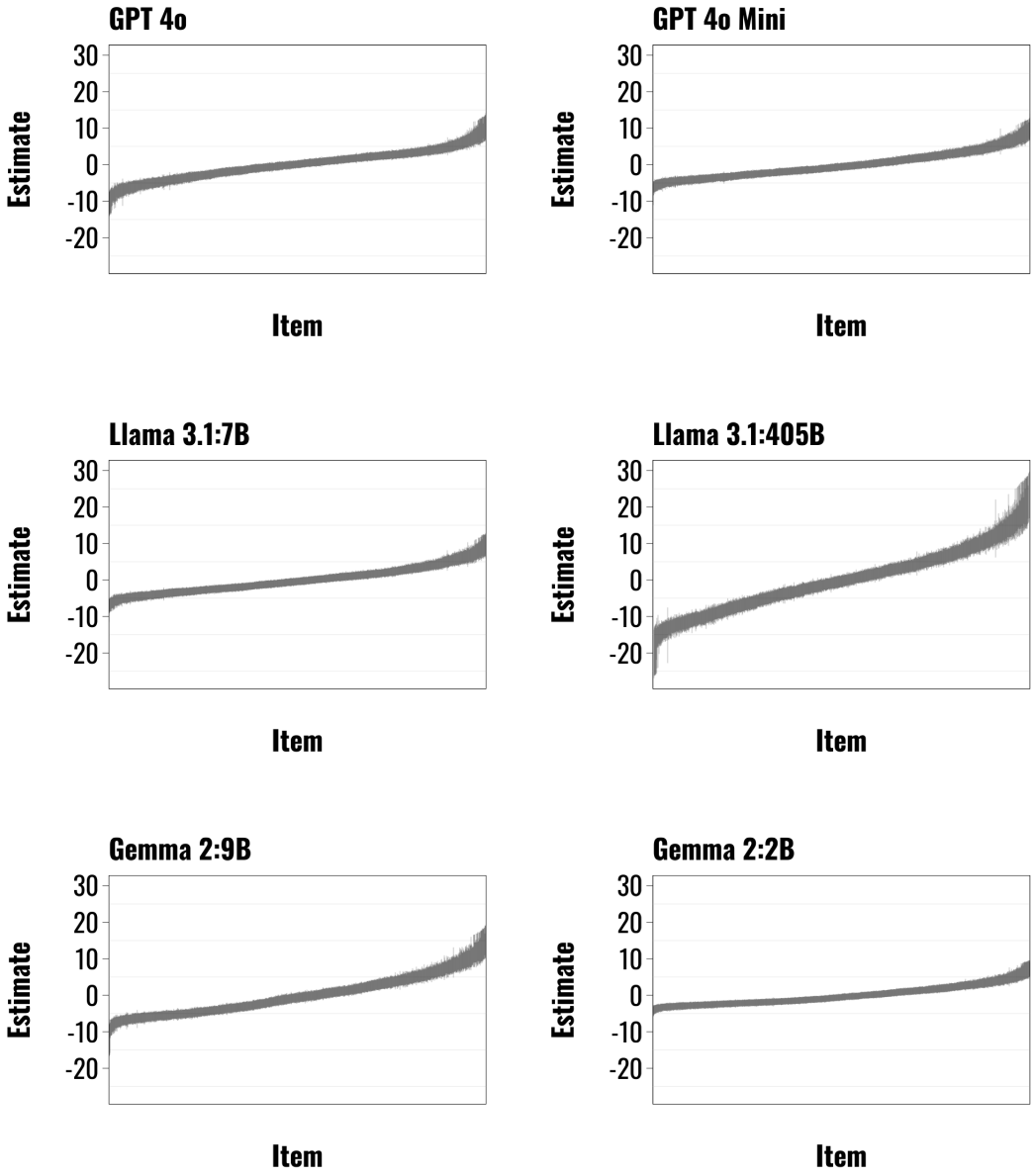

Figure 5: **Bayesian Bradley-Terry Estimates of Interest Rate Knowledge by LLM.** Each panel plots the 95% credible interval for the posterior distributions for each respondent in our dataset. Items are sorted highest to lowest within each panel.

Table 3: **Correlation of 0-10, One-shot, Ratings by LLM**

	Gemma 2:2B	Gemma 2:9B	GPT 4o mini	GPT 4o	Llama 3.1:405B	Llama 3.1:7B
Gemma 2:2B	1.000	0.877	0.819	0.769	0.838	0.737
Gemma 2:9B	0.877	1.000	0.903	0.847	0.920	0.784
GPT 4o mini	0.819	0.903	1.000	0.882	0.911	0.731
GPT 4o	0.769	0.847	0.882	1.000	0.846	0.711
Llama 3.1:405B	0.838	0.920	0.911	0.846	1.000	0.779
Llama 3.1:7B	0.737	0.784	0.731	0.711	0.779	1.000

Table 4: Note: The table shows correlation of the results of a prompt asking each LLM to rate the knowledge of the statement on a 0-10 scale. The temperature in each model was set to "0". N= 1402.

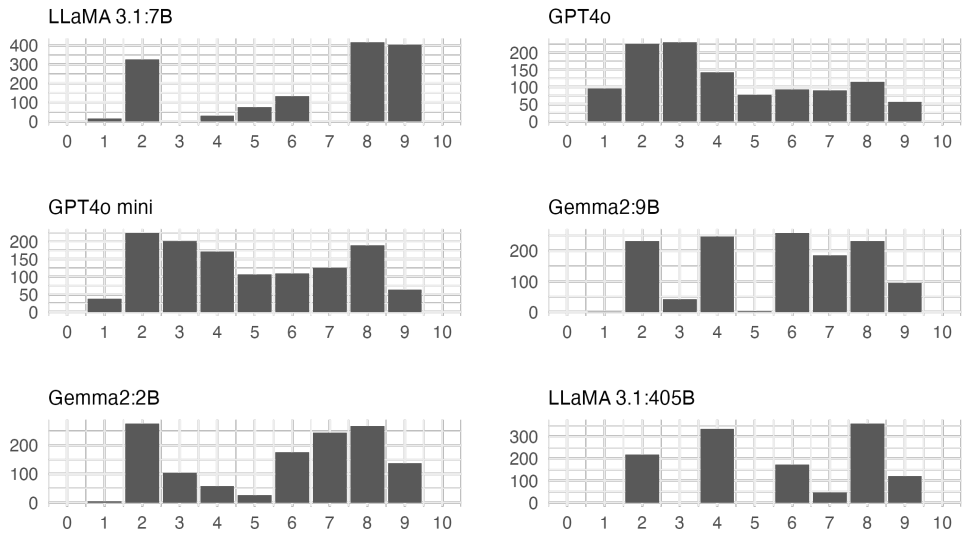

Figure 6: **Distribution of Ratings:** This figures presents the distributions of the 0-10 ratings of interest rate knowledge for each of the LLMs (N=1402).

3 presents the correlations of these ratings. Several things stand out. First, we see that while many of the models are given range of values to place a statement, they rely on a fraction of those. As such, zero-shot ratings may inherently limit the nuance of their output and thus miss key distinctions in between responses. This also gives further pause as it appears the patterns do not follow an apparent logic. It suggests that a inherent bias is selecting some numbers over others. Next, we see that among the largest 2 LLMs the correlation is high but not as high as the BT estimates. In theory, GPT4o and Llama 3.1:405B should have the most consistent responses. However, the correlation is only 0.846, compared with

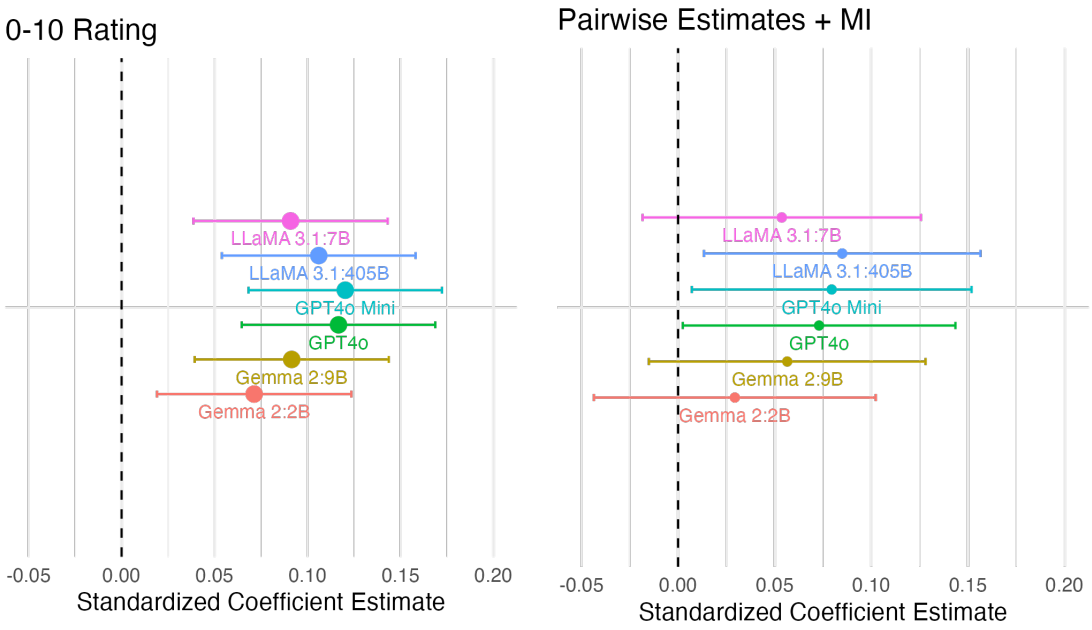

Figure 7: **Coefficient of Interest Rate Knowledge on Support for Central Bank Independence:** Both panels plot the standardized coefficients for of 'interest rate knowledge' in a linear model predicting support for central bank independence for each of the LLMs used in our analyses. Each model also includes controls for income and education. The left plot relies on a 0-10 one shot rating on knowledge

0.953 in the BT estimates. Among the smaller LLM ratings, we see a substantially higher correlation with both smaller and larger LLMs. At least when it comes to our task here, the pairwise comparison task appears well suited for use with larger LLMs but provides inconsistent results with models of lesser capabilities. In this domain, a simpler rating task produces more consistency.

The last step in our workflow is to draw from the distribution of Bradley Terry estimates and use a multiple (over)imputation framework to recover an aggregate estimate that incorporates the uncertainty in the latent variable. Here we use our latent variable, interest rate knowledge, as a predictor of support for central bank independence. The basic intuition is that when people have more knowledge about how and who sets the base interest rate they are more likely to favor independence. As such, we estimate linear models that include interest rate knowledge plus potential confounders: education and income. The right panel of Figure 7 presents the standardized coefficients of interest rate knowledge for each the

LLM derived latent variables. For comparison, we also plot the standardize coefficients of similar linear models for the zero-shot 0-10 ratings in the left panel. As we can see, the two approaches return point estimates of similar sizes. Yet, as expected, once we consider the uncertainty around the latent variable, the confidence intervals around the coefficient are considerably larger. Depending on the underlying LLM comparisons, this can result in the difference between a significant or insignificant result. The comparison nicely illustrates the potential consequences of relying solely on one-shot numerical ratings in downstream models.

Scope and Limitations

The framework presented here demonstrates that LLMs can be useful for scaling open-ended questions and that pairwise comparisons have several advantages over simply prompting an LLM to place a statement on a ordinal scale. We show that, at least when it comes to this task, LLMs' pairwise judgments correspond closely to expert or near-expert human coders, there is high agreement among different LLMs, and they outperform numerical ratings.

While the method is useful, it does have several limitations. First, researchers must be able to identify and phrase a question that will reveal the latent dimension of interest. Some concepts may still be better probed in the context of discrete factual questions. While others might benefit from a longer exposition. Given that the domain knowledge and capacities of LLMs have not been fully tested, it is still necessary to validate the use of LLMs with high quality benchmarks (Gilardi et al., 2023).

Second, some concepts may be easier to identify than others. We carried out an additional illustration (found in our Supplementary Appendix) in which we asked an LLM to identify 'uncertainty' in respondents expectations of the economic consequences of government action. In this exercise, we see less agreement across the different LLMs. However, the method still produces more consistent output than numerical ratings. Consequently, researchers should

do their due diligence to demonstrate that LLM output is consistent across and within models and that the findings are not dependent on the judgments of a single LLM.

Third, the ability of various LLMs to accurately and consistently evaluate the factual content of respondent answer depends on the integrity of the underlying LLM model and training data. If a given LLM cannot "retain" a piece of factual information then its ability to evaluate how factual a given response is will likely not be stable over time (Khatun and Brown, 2024). Other factors, like the framing of questions/prompts or the inclusion of extra or unnecessary language in prompts, can produce incorrect responses. Similarly, while multiple LLMs may rank particular options as the most likely correct responses, it is difficult to assess the degree of confidence or uncertainty across various LLMs, or how stable they are over time (Wang et al., 2024).

Finally, LLMs may potentially bring new life to the use of open-ended questions in survey research. However, they also bring risks beyond the structural and mechanical problems of the LLMs themselves. The increasingly widespread use of LLMs for completing various user tasks may interfere with efforts to collect user knowledge from surveys. We were suspicious that several of the open-ended responses from our Prolific respondents were themselves generated by LLMs. To explore this further, when we recontacted respondents to serve as our human benchmarks we again asked them to answer the question about interest rates. This time, we hid an additional request in the html in very and small transparent font. We asked that the response "mention Alan Greenspan". This served the purpose of providing a very specific instruction that is unlikely to be included without prompting but also one that would not arise too much suspicion if read by the human pasting in to the survey. We found that 5 of our top 20 respondents were using AI to answer the open-ended question. We removed these responses from the final dataset. While the use of LLMs by survey respondents were not detrimental to our analysis, it does show that respondents may rely on LLMs for cognitively demanding tasks like open-ended questions. Further, the prospect of AI agents completing surveys on their own raise the risk that entire survey forms will be completed by

AI. Consequently, survey researchers, whether using open or closed questions, should design strategies to identify these responses and drop them from the dataset.

References

Andre, Peter , Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart (2024). Narratives about the macroeconomy. Technical report, SAFE Working Paper.

Barrie, Christopher , Alexis Palmer, and Arthur Spirling (2024). Replication for language models: Problems, principles, and best practice for political science. https://github.com/ArthurSpirling/LargeLanguageReplication?tab=readme-ov-file. Working Paper.

Benoit, Kenneth , Kevin Munger, and Arthur Spirling (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science 63*(2), 491–508.

Bermejo, Vicente J , Nicolás Harari, Ramiro H Gálvez, and Andres Gago (2024). Llms outperform outsourced human coders on complex textual analysis. *Available at SSRN*.

Blackwell, Matthew , James Honaker, and Gary King (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research 46*(3), 303–341.

Bradley, Ralph Allan and Milton E Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Bürkner, Paul-Christian (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software 80*(1), 1–28.

Bürkner, Paul-Christian (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal 10*(1), 395–411.

Carlson, David and Jacob M Montgomery (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review 111*(4), 835–843.

Carpenter, Bob (2018). The bradley-terry model of ranking via paired comparisons. *RPubs*.

Converse, Jean M (1984). Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s. *Public Opinion Quarterly 48*(1B), 267–282.

Davidson, Roger R (1970). On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association 65*(329), 317–328.

Di Leo, Riccardo , Chen Zeng, Elias Dinas, and Reda Tamtam (2024). Mapping (a) ideology: A taxonomy of european parties using generative llms as zero-shot learners. *Available at SSRN 4907347*.

DiGiuseppe, Matthew , Carolina Garriga, and Andreas Kern (2024). Information, party politics, and public support for central bank independence. Working paper.

DiGiuseppe, Matthew and Patrick Shea (forthcoming). Information, uncertainty, & public support for brinkmanship during the 2023 debt limit negotiations. *British Journal of Political Science*. Forthcoming.

Firth, David (2005). Bradley-terry models in r. *Journal of Statistical Software 12*.

Gabry, Jonah , Rok Češnovar, Andrew Johnson, and Steve Bronder (2024). *cmdstanr: R Interface to 'CmdStan'*. R package version 0.8.1, https://discourse.mc-stan.org.

Geer, John G (1991). Do open-ended questions measure "salient" issues? *Public Opinion Quarterly 55*(3), 360–370.

Gilardi, F. , M. Alizadeh, and M. Kubli (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences 120*(30), e2305016120.

Gruber, Johannes B. and Maximilian Weber (2024, Apr). rollama: An r package for using generative large language models through ollama. *arXiv preprint*. A Preprint.

Haaland, Ingar K , Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart (2024). Measuring what is top of mind. Technical report, National Bureau of Economic Research.

Heseltine, Michael and Bernhard Clemm von Hohenberg (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics 11*(1), 20531680241236239.

Hultberg, Patrik T , David Santandreu Calonge, Firuz Kamalov, and Linda Smail (2024). Comparing and assessing four ai chatbots' competence in economics. *Plos one 19*(5), e0297804.

Kaye, Ella and David Firth (2022). Bradleyterryscalable.

Khatun, Aisha and Daniel G. Brown (2024). Trutheval: A dataset to evaluate llm truthfulness and reliability.

Kraft, Patrick W (2024). Women also know stuff: challenging the gender gap in political sophistication. *American Political Science Review 118*(2), 903–921.

Lazarsfeld, Paul F (1944). The controversy over detailed interviews—an offer for negotiation. *Public opinion quarterly 8*(1), 38–60.

Lincan Li, Jiaqi Li, Catherine Chen Fred Gui Hongjia Yang Chenxiao Yu Zhengguang Wang Jianing Cai Junlong Aaron Zhou Bolin Shen Alex Qian Weixin Chen Zhongkai Xue Lichao Sun Lifang He Hanjie Chen Kaize Ding Zijian Du Fangzhou Mum Jiaxin Pei Jieyu Zhao Swabha Swayamdipta Willie Neiswanger Hua Wei Xiyang Hu Shixiang Zhu Tianlong Chen Yingzhou Lu Yang Shi Lianhui Qin Tianfan Fu Zhengzhong Tu Yuzhe Yang Jaemin Yoo Jiaheng Zhang Ryan Rossi Liang Zhan Liang Zhao Emilio Ferrara Yan Liu Furong Huang Xiangliang Zhang Lawrence Rothenberg Shuiwang Ji Philip S. Yu Yue Zhao Yushun Dong (2024). Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.xxxxx*.

Loewen, Peter John , Daniel Rubenson, and Arthur Spirling (2012). Testing the power of arguments in referendums: A bradley–terry approach. *Electoral Studies 31*(1), 212–221.

Ludwig, Jens , Sendhil Mullainathan, and Ashesh Rambachan (2024). Large language models: An applied econometric framework. *arXiv preprint arXiv:2412.07031*.

Mattos, David Issa and Érika Martins Silva Ramos (2022). Bayesian paired comparison with the bcps package. *Behavior Research Methods 54*, 2025–2045.

Mellon, Jonathan , Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman (2024). Do ais know what the most important issue is? using language models to code open-text social survey responses at scale. *Research & Politics 11*(1), 20531680241231468.

Mens, Gaël Le and Aina Gallego (2023). Scaling political texts with chatgpt. *arXiv preprint arXiv:2311.16639*.

Narimanzadeh, Hasti , Arash Badie-Modiri, Iuliia G Smirnova, and Ted Hsuan Yun Chen (2023). Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction 7*(CSCW2), 1–29.

Ornstein, Joseph T. , Elise N. Blasingame, and Jake S. Truscott (2025). How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods*, 1–18.

Peysakhovich, Alexander and Adam Lerer (2023). Attention sorting combats recency bias in long context language models.

Poole, Keith T and Howard Rosenthal (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press, USA.

Rathje, Steve , Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences 121*(34), e2308950121.

Roberts, Margaret E , Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand (2014). Structural topic models for open-ended survey responses. *American journal of political science 58*(4), 1064–1082.

Stan Development Team (2024). *Stan Modeling Language Users Guide and Reference Manual, v2.36.0*.

Törnberg, Petter (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 08944393241286471.

Turner, Heather and David Firth (2012). Bradley-terry models in r: The bradleyterry2 package. *Journal of Statistical Software 48*.

van Paridon, JP , Ben Bolker, and Phillip Alday (2023). *lmerMultiMember: Multiple membership random effects*. R package version 0.11.8.

Wang, Weixuan , Barry Haddow, Alexandra Birch, and Wei Peng (2024, June). Assessing factual reliability of large language model knowledge. In K. Duh, H. Gomez, and S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico, pp. 805–819. Association for Computational Linguistics.

Wu, Patrick Y , Jonathan Nagler, Joshua A Tucker, and Solomon Messing (2023a). Concept-guided chain-of-thought prompting for pairwise comparison scaling of texts with large language models. *arXiv preprint arXiv:2310.12049*.

Wu, Patrick Y , Jonathan Nagler, Joshua A Tucker, and Solomon Messing (2023b). Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*.

Yang, Cehao , Chengjin Xu, and Yiyan Qi (2024). Financial knowledge large language model. *arXiv preprint arXiv:2407.00365*.

Zollinger, Delia (2024). Cleavage identities in voters' own words: Harnessing open-ended survey responses. *American Journal of Political Science 68*(1), 139–159.

Zucco Jr, Cesar , Mariana Batista, and Timothy J Power (2019). Measuring portfolio salience using the bradley–terry model: An illustration with data from brazil. *Research & Politics 6*(1), 2053168019832089.

A Appendix

A.1 Chain of Thought Prompting

In addition to asking an LLM to return just a final answers on which statement best aligned with the latent dimension of interest (knowledge), we also attempted a Chain of Thought (CoT) pair-wise comparison prompting approach following the recommendation of (Wu et al., 2023a). We used the following prompt with the llama 3.1:405B model:

"You are an expert in US economic policy. Your task is to determine which of two given statements contains a more knowledgeable response to the following question:", "In a few sentences and without looking it up, can you explain how interest rates (i.e., the cost of borrowing money to buy a house or car) go up or down in the US economy?", "Follow these steps to complete the task:", "Step 1: Write out your evaluation of Statement 1, discussing its strengths, weaknesses, and gaps in knowledge.", "Step 2: Write out your evaluation of Statement 2, discussing its strengths, weaknesses, and gaps in knowledge.", "Step 3: Compare your evaluations of the two statements and explain which one demonstrates greater knowledge, or why they are equal or incomparable.", "Step 4: Based on your reasoning, provide your final decision.", "Your response should include the full reasoning for each step, and the final decision must be presented as:", "Final Decision: [1] or Final Decision: [2] or Final Decision: [0]", "Here are the statements to evaluate:", "1:", [statement1], "2:", [statement2], "Write out your full evaluation and conclude with the final decision in the specified format."

We find, contrary to our expectations, that the CoT prompt performed worse than the direct prompt against the "close to expert" benchmark. As such, we present only the direct prompt in our central analysis.

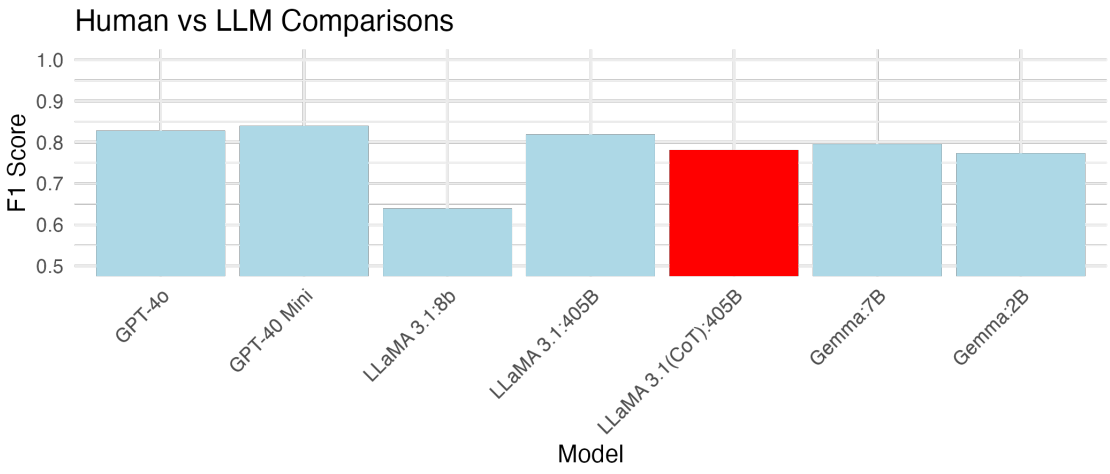

Figure A1: F1 - LLM-Human Comparison including Chain of Thought prompt

A.2 Additional Illustration: Uncertainty

In addition to the illustration above, we also applied our framework to an experimental setting where the variable of interest is a dependent variable. In a forthcoming paper, DiGiuseppe and Shea (ming attempt to manipulate respondents' uncertainty over the consequences of a Debt Ceiling breach in the run-up to the 2023 Debt Ceiling deadline in the United States. Figure **??** presents the images in the experiment. As a manipulation check, the authors asked respondents to report their expectations of what would happen if the US breached the debt ceiling. In their appendix, the authors used and LLM to rate the uncertainty of each response on a 0-10 scale and find that those in the treatment condition did, in fact, produce statements with greater uncertainty. Here, we apply a pairwise comparison approach to this analysis.

The figures and tables in this section indicate a few things. First, there is much less consistency in the model output in this task. The correlations are much lower among both the BT estimates and the 0-10 rankings. The BT estimates appear to be more consistent among the high-end models. Among the 3 frontier models (LLama 3.1:405B, GPT4o and GPT4o mini), the correlation of the final output ranges from 0.59 to 0.82. Still, this may be too low to have confidence in any particular model for this task.

Table A1: **Correlation of BT Estimates Uncertainty by LLM**

	Gemma_2:2B	Gemma_2:9B	GPT_4_0_Mini	GPT_4_0	Llama_3.1:405B	Llama_3.1:7B
Gemma_2:2B	1.00	0.55	0.23	0.06	-0.19	0.55
Gemma_2:9B	0.55	1.00	0.72	0.58	0.24	0.80
GPT_4_0_Mini	0.23	0.72	1.00	0.82	0.59	0.70
GPT_4_0	0.06	0.58	0.82	1.00	0.77	0.49
Llama_3.1:405B	-0.19	0.24	0.59	0.77	1.00	0.19
Llama_3.1:7B	0.55	0.80	0.70	0.49	0.19	1.00

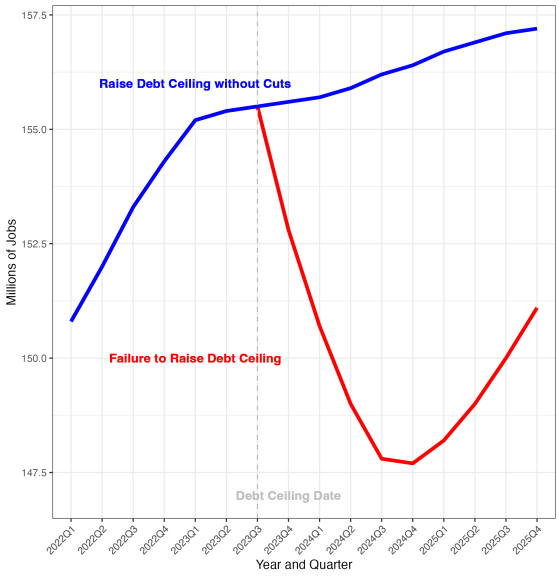

Figure A2: Control Figure

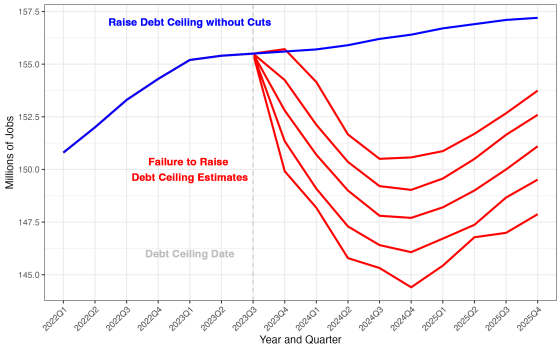

Figure A3: Treatment Figure

Figure A4: **DiGiuseppe and Shea's (2025) Treatment Conditions**

Table A2: Correlation Matrix of LLM 0-10 Ratings

	Gemma 2:2B	Gemma 2:9B	GPT 4o mini	GPT 4o	Llama 3.1:405B	Llama 3.1:7B
Gemma 2:2B	1.000	0.288	0.374	0.451	0.487	0.362
Gemma 2:9B	0.288	1.000	0.297	0.360	0.313	0.273
GPT 4o mini	0.374	0.297	1.000	0.566	0.632	0.548
GPT 4o	0.451	0.360	0.566	1.000	0.587	0.434
Llama 3.1:405B	0.487	0.313	0.632	0.587	1.000	0.620
Llama 3.1:7B	0.362	0.273	0.548	0.434	0.620	1.000

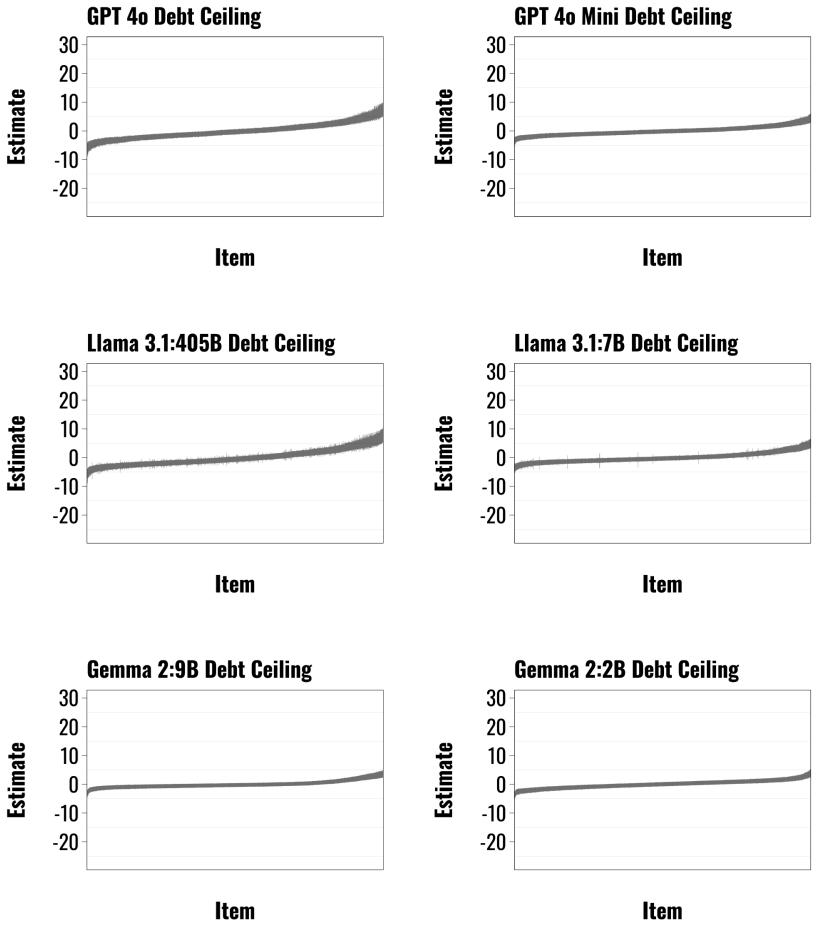

Figure A5: Bayesian BT Scores of Uncertainty

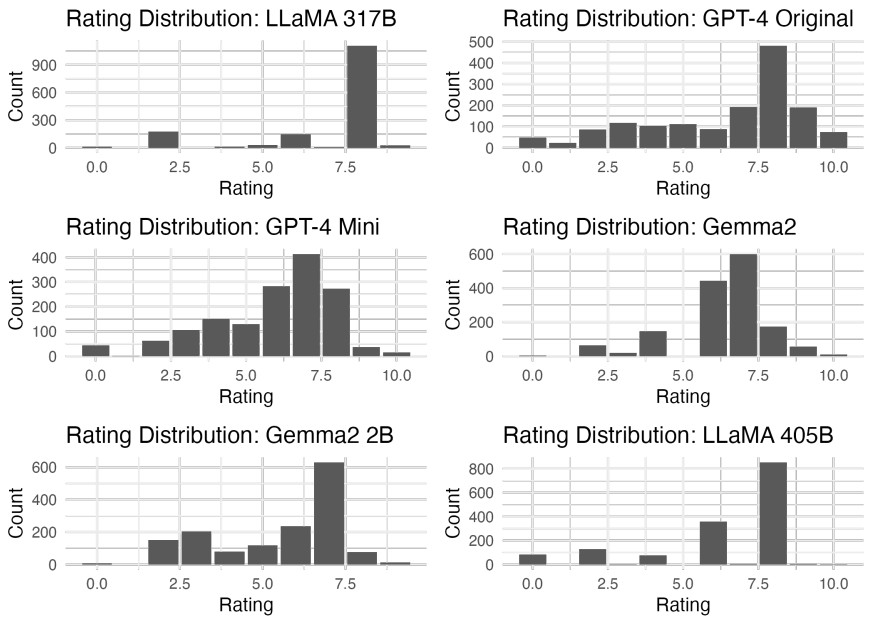

Figure A6: Distribution of Uncertainty 0-10 Ratings by LLM

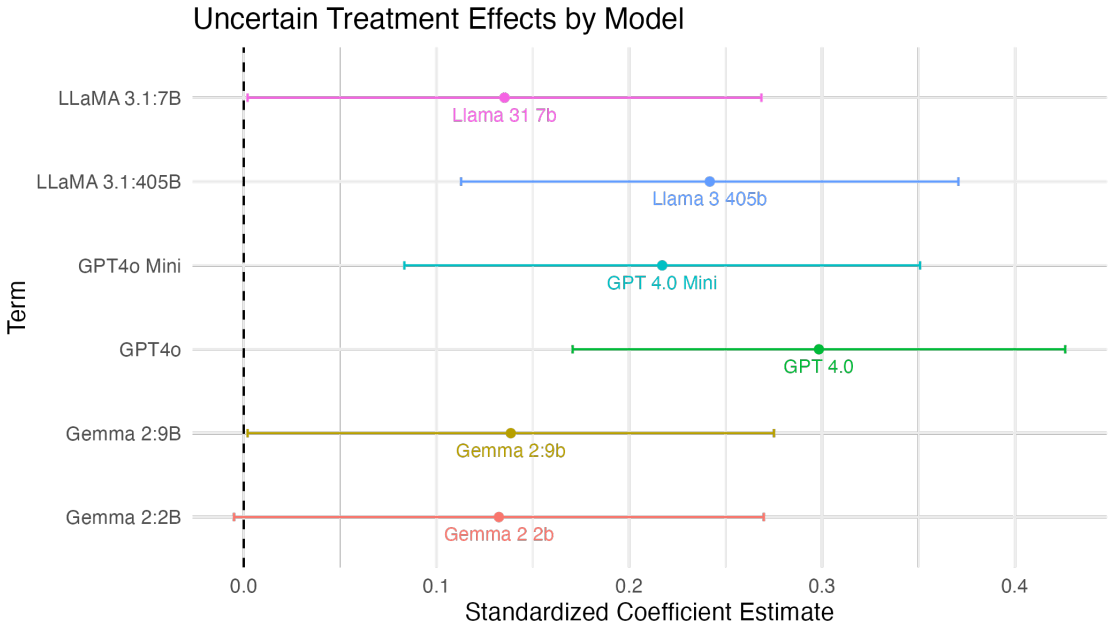

Figure A7: **ATE of treatments based on LLM codings of 'Uncertainty DV':** The point estimates indicate the average treatment effect and bars indicate the 95% confidence intervals.