MAX PLANCK INSTITUTE
FOR RESEARCH ON
COLLECTIVE GOODS

JOHANNES KRUSE

Discussion Paper
2025/12

THE ORDINARY MEANING BOT: SIMULATING HUMAN SURVEYS WITH LLMS

The ordinary meaning bot:
Simulating human surveys with LLMs

Comment
by
Johannes Kruse

Abstract

This Comment shows how large language models (LLMs) can help courts discern the "ordinary meaning" of statutory terms. Instead of relying on expert-heavy corpus-linguistic techniques (Gries 2025), the author simulates a human survey with GPT-4o. Demographically realistic AI agents replicate the 2,835 participants in Tobia's 2020 study on vehicle and yield response distributions with no statistically significant difference from the human data (Kolmogorov–Smirnov p = 0.915). The paper addresses concerns about hallucinations, reproducibility, data leakage, and explainability, and introduces the locked-prompt "Ordinary Meaning Bot," arguing that LLM-based survey simulation is a practical, accurate alternative to dictionaries, intuition, or complex corpus analysis.

Keywords: Ordinary meaning; large language models; prompt engineering; human survey simulation; alignment

JEL classification: K1, Z0

1. Introduction

To determine the ordinary meaning, key concept in U.S. law, judges can use various methods: They may rely on their intuition, consult dictionaries, employ corpus-linguistic techniques, or - more recently - use large language models (LLMs). The usefulness of these methods mainly depends on two factors: practicality and accuracy. Practicality considers the barriers that must be overcome before a technique can be used in court. Accuracy assesses how well a method captures the true ordinary meaning, which is the understanding of an average reader of a disputed term. With these factors in mind, this Comment first reviews the frequency-based corpus approach proposed by Stefan Gries (2025). It then introduces a new LLM-based technique that is not only easier to implement, but also produces results more closely aligned with Tobia (2020), which serves as the relevant gold standard.

2. Corpus-linguistics in practice

A key first step is to clarify what is meant by "corpus linguistics" when assessing the practicality of these methods for determining ordinary meaning. In this context, it is crucial to distinguish between two approaches. First, there is the do-it-yourself (DIY) approach, where judges with no specialized training rely on corpus linguistics tools themselves (Gries, 2025, p. 18). Second, there are more advanced methods, such as the one proposed by Stefan Gries. What makes Gries's approach particularly sophisticated is that it goes beyond simple frequency counts to include complex computational and statistical steps. It uses word-embedding models to represent concepts as multi-dimensional vectors and calculates a final score for a candidate term based on its weighted similarity to a set of semantic features derived from a large text corpus. This highly technical process requires significant expertise for proper execution. Naturally, this approach cannot be used directly by judges. However, this is not the author's intention. In his view, such an analysis should be left to the relevant experts, namely linguists. Relying on experts, however, presents the well-known practical challenges of dealing with expert testimony (Robertson, 2010). First, a judge must identify a suitable expert witness. This often triggers the infamous battle of experts, with each party presenting retained experts who offer contradictory opinions. Without the necessary technical expertise, the judge faces the difficult task of evaluating conflicting expert evidence. Additionally, this process involves considerable costs and delays. While this may be acceptable in proceedings before the Supreme Court, it is a different matter for lower-level cases with less at stake. These hurdles often discourage judges from calling an expert, especially when lower-threshold "alternatives" like ChatGPT are easily accessible.

3. Predicting human survey responses with LLMs

Now let's turn to a novel method for leveraging large language models to determine the ordinary meaning.

3.1 In a nutshell

The idea behind this method is straightforward: Human surveys are the best way to understand how an average person perceives something. So, we use LLMs to replicate these surveys. As Engel and McAdams (2024) argue, we don't ask the models directly if something is a "vehicle." Instead, we view them as tools for gathering evidence about how ordinary people interpret the term "vehicle".

To evaluate this approach, we first need a benchmark (ground truth). For the present purpose, we use a study published in 2020 by Kevin Tobia, which examines the ordinary meaning of 'vehicle' using 25 objects (such as a car, airplane, or skateboard) (Tobia, 2020). Based on responses from 2.835 participants, recruited via Amazon Mechanical Turk (MTurk), these results are regarded as "the best extant evidence of ordinary meaning of vehicles in the park" (Engel & McAdams, 2024, p. 259). Our process instructs GPT-4o to predict human responses and then compares the resulting distribution with Tobia's using a Kolmogorov–Smirnov (KS) test. A high p-value (typically> 0.05) indicates that the test found no statistically significant difference between the two distributions, meaning the hypothesis that they are indistinguishable cannot be rejected.

How should the model be prompted? Engel & McAdams showed that a belief prompt framed on a 7-point Likert scale yields reliable alignment ($p = 0.28$). Here is a simplified version: "We asked 2,835 people whether each object below is a vehicle (Yes/No). Please tell us what percentage you believe answered 'Yes'." Since LLMs are better at processing language than numbers, the prompt provides seven verbal ranges from "(almost) none" to "(almost) all." Each query is repeated 50 times to smooth out model randomness.

The present approach differs crucially from Engel and McAdams (2024): Rather than asking GPT for aggregate yes-rates, it seeks to approximate the individual response behavior that underlies those aggregates. The rationale is straightforward. Tobia's (2020) mean scores are merely the distilled essence of thousands of single judgements, each exhibiting non-trivial variance. To get closer to these results, one should start by considering the participants as individual influencing factors. The present approach, an LLM-supported study simulation with AI agents, aims to predict response behavior. The process has three steps:

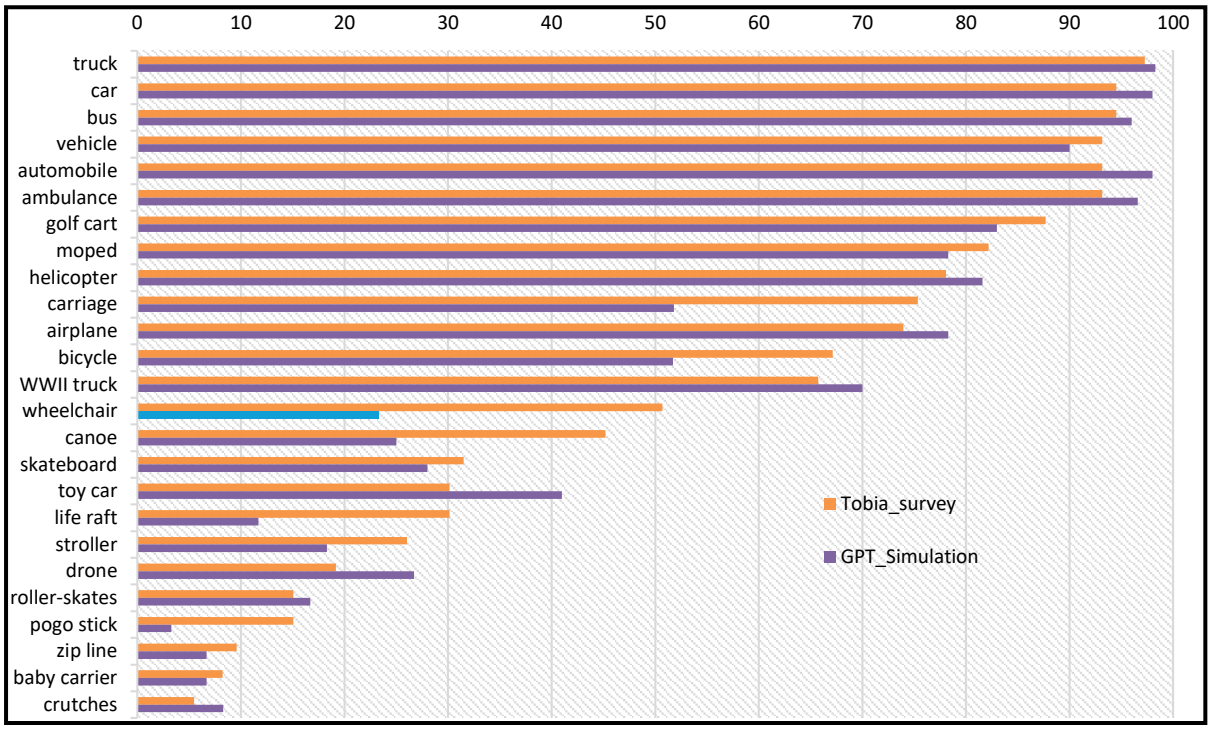
> (i) We identify 10 demographic factors relevant to ordinary-meaning judgments about "vehicle" (e.g., age, education, gender ratio, vehicle ownership) and estimate their distributions for the target population (MTurk workers).
>
> (ii) We generate 2,835 AI agents with 10-factor demographic profiles and cluster demographically similar individuals into subsamples of 25; each subsample is represented in the prompt by its modal traits. This aggregation into groups of 25 is

particularly useful given the relative demographic homogeneity of Amazon Mechanical Turk (MTurk) samples (Difallah et al., 2018).

(iii) For each subsample and each classification item ("Is X a vehicle?"), GPT-4o predicts the subsample's 7-point Likert response (one of seven percentage-range phrases). We repeat predictions across 50 runs (temperature = 0.9, top_p = 1.0) and include a short hint about typical online-survey variability in participant attention.

Now, let's review the results. The Kolmogorov-Smirnov test finds no statistically significant difference between the GPT responses and the gold standard ($p = 0.915$). As the figure below demonstrates, the GPT and human responses show close alignment across nearly all items.

Figure 1: Yes-rates ("Is X a vehicle?"): GPT-4o predictions vs. Tobia (2020) human data.



3.2 Two doubts/concerns regarding the proposed application

However, is this a desirable use case for LLMs? Gries (2025) raises doubts and highlights two key issues with LLMs: hallucinations and the replication problem. Hallucinations are a serious problem. They frequently occur when answering legal questions (Dahl et al., 2024; Deroy et al., 2023; Magesh et al., 2025). In the present context, however, the problem is limited. Just because LLMs frequently hallucinate when answering legal questions doesn't mean they will do so here. The opposite is more plausible. A major cause of hallucinations is a lack of knowledge (Huang et al., 2025, p. 1:7; Engel & Kruse, 2024a, p. 1001; 2024b, p. 8). The knowledge an LLM has in a specific field depends on the number of relevant documents it encountered during

pre-training (Kandpal et al., 2023, p. 1). This being so, it is not surprising that LLMs often hallucinate when answering open-ended legal questions. After all, only a small fraction of the training data is legal. However, this study is not concerned with applying legal knowledge, but rather with how the average reader understands a term. The training data contain rich information on this topic (Engel & McAdams, 2024, p. 248), and the human understanding of language is central to reinforcement learning. As the previous results clearly show, post-training refinements (Choi, 2025; Lee & Egbert, 2024) have not significantly impaired the ability of GPT to grasp the ordinary meaning of "vehicle".

Replication is challenging because LLM outputs aren't always consistent, so you might get different answers each time (Gries, 2025, p. 10). This problem is addressed by querying the model thousands of times for each item. When we average these results, we obtain a robust, statistically reliable score. Doing this helps us find the true signal amid the "noise" of individual LLM responses. Major causes of inconsistency, such as temperature and prompt differences, are removed with the proposed interface (see below).

Another potential concern is training-data contamination: the LLM may already be familiar with Tobia's publicly available study, making our "ground truth" partially embedded in its training data. Such leakage could happen not only through direct inclusion, but also via secondary diffusion. Although data contamination cannot be entirely dismissed, the current results cannot be solely attributed to it. Otherwise, Engel and McAdams would have shown a higher level of alignment, not a lower one. Compared to our method, their inquiry about the aggregated variables can be more clearly linked to and answered using the published study (since the average values are included). However, individual responses and specific demographic details remain unpublished and thus inaccessible.

Finally, the limitations of using proprietary LLMs should not be overlooked. Although we can identify an alignment, we cannot precisely explain how GPT arrived at this result - the proverbial "black box" issue. This lack of explainability is problematic, especially since interpreting statutes is an exercise of power, and it raises questions about legitimacy.

3.3 Practicality

But how practical is this approach? Unlike chatbots such as ChatGPT, which allow users to interact with LLMs without needing prior knowledge, the proposed pipeline requires more expertise. In its original form, it demands a deep understanding of LLMs and, for API use, at least basic coding skills. To make it more accessible, the pipeline was transformed into an interface: the "ordinary meaning bot".[1]

This interface offers guided access to LLMs and provides additional advantages in the given context. The sensitivity of LLMs to prompting and framing effects, along with the influence of temperature settings or model selection, challenges the consistency of methodology (Gries, 2025). This issue is particularly serious in litigation settings. Litigants could try to manipulate results by cherry-picking prompts (Choi, 2025, p. 25). The "ordinary meaning bot" helps reduce

1 You can view and try out the interface here: https://johkrus.github.io/The-ordinary-meaning-bot/.

this risk by preventing users from changing prompts or adjusting parameters like temperature. This ensures prompts and settings remain fixed, maintaining consistent outcomes and allowing results to be compared across different cases.

4. Outlook

This paper demonstrates that LLMs could serve as a functionally equivalent method to human surveys in testing ordinary meaning, and could offer a more methodologically sound option than traditional tools like dictionaries. At least for current, everyday language use.

However, sometimes we encounter meanings that are not well-covered in the training data because they are too specific or too historical (Gries, 2025). In those cases, we should supplement the LLM with additional knowledge using Retrieval-Augmented Generation (RAG). Besides using specialized corpora (like the Corpus of Historical American English), there is great potential in combining LLMs with corpus linguistics.

Further research is needed in this area, particularly because hallucinations may occur even in the context of RAG-based approaches (Gries, 2025, p. 10). Other possible applications of the current approach should also be explored, such as interpreting contract clauses.

Three key aspects are crucial for future discussions on this matter.

> (i) Inevitable adoption. Judges and lawyers already use LLMs in various ways, some more transparent than others, and they will keep doing so. The LLM genie is out of the bottle; a return to the pre-ChatGPT world is unrealistic. We must therefore face both the opportunities and the limitations of its use.
>
> (ii) Strengths and weaknesses. LLMs possess enormous potential and are constantly improving, yet they also have shortcomings. What is required, then, is a reflective use that takes these weaknesses into account—something guided interfaces can help to achieve.
>
> (iii) Relative yardstick. The value of LLMs in legal interpretation should be assessed only in light of the existing alternatives. Whether LLMs are perfect is not the point (they are not); the sole question is whether, in terms of accuracy and usability, they outperform judicial intuition, dictionaries, or legal corpus-linguistic methods.

References

Choi, J.H. (2025), “Off-the-Shelf Large Language Models Are Unreliable Judges.” https://papers.ssrn.com/abstract=5188865 (accessed July 14, 2025)

Dahl, M., Magesh, V., Suzgun, M., and Ho, D.E. (2024), “Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models,” *Journal of Legal Analysis*, 16(1), 64–93.

Deroy, A., Ghosh, K., and Ghosh, S. (2023), “How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?”

http://arxiv.org/abs/2306.01248 (accessed July 7, 2025)

Difallah, D., Filatova, E., and Ipeirotis, P. (2018), “Demographics and Dynamics of Mechanical Turk Workers,” in: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, New York, NY, USA, p. 135–143.

Engel, C., and Kruse, J. (2024a), “Kommentar ohne Autor,” *JuristenZeitung*, 79(22), 997.

Engel, C., and Kruse, J. (2024b), “Professor GPT: Having a Large Language Model Write a Commentary on Freedom of Assembly.” https://papers.ssrn.com/abstract=4994131 (accessed July 7, 2025)

Engel, C., and McAdams, R.H. (2024), “Asking GPT for the Ordinary Meaning of Statutory Terms,” *University of Illinois Journal of Law, Technology & Policy*, 2024, 235.

Gries, S. (2026), "Ordinary meaning in legal interpretation: a proposal from a corpus and a bit of an LLM) perspective," *Journal of Institutional and Theoretical Economics*, 181, forthcoming.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., et al. (2025), “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” *ACM Transactions on Information Systems*, 43(2), 1–55.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. (2023), “Large Language Models Struggle to Learn Long-Tail Knowledge,” in: *Proceedings of the 40th International Conference on Machine Learning*, PMLR, p. 15696–15707. (accessed July 7, 2025)

Lee, T.R., and Egbert, J. (2024), “Artificial Meaning?” https://papers.ssrn.com/abstract=4973483 (accessed July 22, 2025)

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C.D., and Ho, D.E. (2025), “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” *Journal of Empirical Legal Studies*, 22(2), 216–242.

Robertson, C. (2010), “Blind Expertise,” *New York University Law Review*, 85(1), 174.

Tobia, K. (2020), “Testing Ordinary Meaning,” *Harvard Law Review*, 134(2), 726–807.