A Taxonomy of Interrogatives and Its Role in Human-Language Model Interactions

Supervisor: Dr Daniel de Kadt

Candidate Number: 50280

Course Code: MY498

Word Count: 9,998

August 14, 2025

Abstract

The large-scale adoption of large language models (LLMs) may carry harmful consequences, and understanding these is a complex task. In this context, previous theoretical approaches have sought to characterise interrogative forms but are limited by normative constraints or imposed hierarchical structures. To address these limitations, I operationalise Belnap and Steel's (1976) taxonomy of interrogatives, grounded in a logical framework specifying the structure, purpose, and implications of questions. I do so by fine-tuning transformer-based classifiers on a diverse alignment dataset (Kirk et al., 2024), with attention to replicability, interpretability, and downstream inference. The operationalisation demonstrated high predictive performance across interrogative categories (validation weighted F1 mean, *SD* = .88, .084) and suggested reliable, low-uncertainty predictions. I apply this operationalisation as a theoretical lens in two studies: a descriptive analysis of associations between user characteristics and interrogative types (Study 1) and an experiment testing how these types influence LLM responses (Study 2). In Study 1, younger participants, more educated participants, and those engaged in conversations on controversial topics or values were each more likely to pose closed-ended interrogatives. Effect sizes varied across subgroups, with heterogeneous patterns for birth region, ethnicity, and religion. In Study 2, across 17,984 LLM responses, the most closed-ended interrogatives produced shorter replies with attributes less conducive to constructive dialogue than those from the most open-ended form; this pattern held for 19 of 21 LLMs, although effect sizes and variances differed. Together, these findings point to potential cycles in human–LLM interactions, shaped by user and LLM characteristics, and warrant further research to guide responsible development.

The code for this study is available here, with access to all other replication materials detailed in Appendix A.

1. Introduction

The large-scale adoption of large language models (LLMs) is altering processes of information retrieval and production, making it crucial to investigate their societal repercussions. Potential harms include but are not limited to misinformation, political persuasion, and exacerbated inequalities (Weidinger et al., 2021). With estimates suggesting that one-third of the US population uses LLMs weekly, less than three years after their first large-scale commercial deployment (Bick et al., 2024), the pace and novelty of this change may require adjustments to social scientific theoretical and methodological frameworks to more effectively assess its potential societal consequences and inform timely safeguards.

A key factor shaping LLM responses is their sensitivity to prompt formulation. In this context, a growing body of prompt engineering research has emerged to examine how variations in prompt formulation influence LLM responses. However, most studies adopt either an engineering perspective focused on optimising LLM performance (Schulhoff et al., 2025) or a methodological perspective aimed at enhancing social scientific research methods (Thapa et al., 2025). Consequently, an important gap in the literature is the paucity of research examining how everyday users may inadvertently engage in prompt engineering through individual differences in their routine LLM interactions, and the implications for response characteristics.

To examine and formalise the dynamics of human–LLM interactions, researchers have drawn on established frameworks such as the Gricean maxims (Grice, 1991; Krause & Vossen, 2024) and, in educational contexts, Bloom's Taxonomy (Bloom, 1986). However, the generalisability of the Gricean maxims is limited by their imposition of normative constraints (Kasirzadeh & Gabriel, 2023), and Bloom's Taxonomy, rooted in developmental psychology, imposes a hierarchy of interrogative forms that does not account for the fact that interrogatives can serve diverse purposes not reducible to a linear scale of value or complexity. To address these limitations, I operationalise the taxonomy of interrogatives by Belnap & Steel (1976), a framework for evaluating, relating and categorising interrogatives and answers by their *structure*, *purpose*, and *implications*. It is grounded in erotetic logic, a branch of philosophy at the intersection of semantics and logic. In this context, interrogatives are defined as statements through which an individual poses a question or otherwise elicits a response.

The first part of this research concerns the operationalisation of Belnap and Steel's (1976) taxonomy through a computational approach, as scalability is important for social scientific investigations in the context of widespread LLM adoption. Specifically, I fine-tune multiple BERT-based transformer architectures and seek to address concerns regarding replicability, interpretability, and downstream inference. To do so, I use the PRISM alignment dataset (Kirk et al., 2024), which contains several real LLM–user conversations from each of 1,500 participants across 75 birth countries, linked to their demographic profiles. Findings suggest high predictive performance across interrogative categories and robust reliability, with low overall uncertainty that increases modestly in categories with limited training data, indicating the potential utility of this operationalisation for social scientific investigations of human–LLM interactions.

With this aim, I applied the operationalised theoretical lens in two studies. Study 1 adopts a descriptive approach to characterising Belnap and Steel's (1976) interrogative types in relation to demographic characteristics, using PRISM participant profiles alongside the opening prompts from their conversations with LLMs. Results indicated that closed-ended interrogatives were more likely among younger participants, those with higher educational attainment, as well as in conversations about controversial topics and values, with effect sizes differing across demographic subgroups and interrogative types. In Study 2, I implemented an experimental manipulation comparing opening prompts in PRISM conversations in their most open- and closed-ended interrogative forms, collected LLM responses for both and evaluated them on attributes associated with pro-social and constructive dialogue. Overall, across 17,884 responses from 21 LLMs, closed-ended interrogatives elicited responses that were more likely to display affinity, compassion, curiosity, personal storytelling, and respect, and less likely to exhibit reasoning and nuance, relative to open-ended interrogatives.

Together, these findings are important because they suggest the possibility of human–language model interaction cycles, shaped by both user and LLM characteristics. This indicates that, to better understand the potential harms that may arise from the widespread adoption of LLMs, future research may benefit from moving beyond isolated analyses of either user inputs or LLM responses toward an integrated evaluation of both.

2. Taxonomies of Interrogatives

2.1 Theory and Context

2.1.1 Existing Taxonomies of Interrogatives

To understand and formalise the dynamics of human–language model interactions, recent research has largely centred on the Gricean maxims (Grice, 1991), a focus emphasised in the review by Krause and Vossen (2024). These maxims articulate a set of pragmatic conventions that facilitate constructive communication: providing sufficient information, remaining relevant, avoiding ambiguity, and being truthful. Initially introduced in this context as tools for value alignment (Kasirzadeh & Gabriel, 2023), the maxims have for example been operationalised to improve response clarity (Saad et al., 2025) or more broadly to inform language model design (Kim et al., 2025). A key limitation is that their operationalisation is highly context-dependent and requires a fixed reference standard, despite criteria such as quality being interpreted differently across individuals and cultural settings.

In educational research, Bloom's taxonomy has been employed both to classify forms of inquiry and analyse human–computer interactions with the aim of optimising language model use in pedagogical settings. Bloom's taxonomy is a hierarchical theory with categories spanning from recalling facts to evaluating information (Bloom, 1986). It has been applied to optimise interactive learning through LLMs among children with different backgrounds (Luo et al., 2025), evaluate LLM-generated (Yaacoub et al., 2025) and guide the integration of LLMs in educational contexts (Elim, 2024). Though useful for structuring inquiry in pedagogical settings, it is grounded in developmental psychology and assumes a progression of cognitive complexity. When extended beyond education, its imposition of a hierarchy among question types fails to account for the fact that interrogatives serve varied purposes that are not linearly ordered in terms of value or complexity.

2.1.2 Erotetics

Erotetic logic helps overcome the normative constraints of Bloom's and Grice's taxonomies by reframing inquiry without presupposing fixed standards or value hierarchies. Erotetics is a branch of philosophy situated at the intersection of semantics and logic. It aims to characterise the *structure*, *purpose*, and *implications* of interrogatives. Thereby, it directly addresses what Kasirzadeh & Gabriel (2023) identify as the central limitation of their Gricean approach—its focus on *implications*, a key dimension of *pragmatics*. They argue that a successful study of human–language model interactions must incorporate *syntax*, *semantics*, and *pragmatics*—precisely the areas that erotetics encompasses.

The relevance of erotetics to human–computer interaction was already recognised by Hakkarainen & Sintonen (2002), who demonstrated how Hintikka's game-theoretical model of inquiry, a key erotetic theory, could enhance internet search practices among elementary school students. More recently, Koralus (2023) formalised the Erotetic Theory of Reason, proposing a computational account of reasoning based on the process of raising and answering inquiries, operationalised through the Python package PyETR (Koralus et al., 2025). One of its key contributions to the study of human–language model interaction is its framing of reasoning in terms of issue resolution rather than adherence to fixed normative rules. This may support more flexible language-model alignment strategies that are robust to user-specific or cross-cultural variation. However, their operationalisation has a broader focus on the purpose and implications of interrogatives to reasoning. To address the social scientific aim of understanding prompt engineering from a user-centric perspective, the *structure* of interrogatives should also be considered. This is because a growing body of empirical findings documents the sensitivity of prompt structure to LLM response (Razavi et al., 2025; Sclar et al., 2024).

2.1.3 Belnap & Steel's Taxonomy of Interrogatives

Belnap and Steel's (1976) theory of interrogatives and answers shares the computational scalability of Koralus (2023) but lends itself more directly to the present aim of operationalising a taxonomy of interrogatives to investigate the social scientific implications of human-language model interactions. This is because the theoretical concepts it introduces include the operationalisation of user prompt *structure*. More specifically, Belnap & Steel (1976) propose three concepts to formalise the dynamics between a questioner and a respondent that are directly relevant to the aims of the present research. The first is *selection-size-specification*, which concerns the extent to which the questioner predefines the range of acceptable answers through the form of the interrogative. For example,

(1) "Is burning fossil fuels or deforestation the key driver of climate change?"

has a *selection-size-specification* of two, because the *structure*, *purpose*, and *implications* of the interrogative constrain the respondent to choose between two predefined options. In contrast:

(1) "How is climate change understood to happen?"

has an undefined *selection-size-specification*, since the questioner does not predefine the range of possible answers, leaving the respondent free to define extent of the given answer. As such, *selection-size-specification* serves as an indicator of the degree to which an interrogative is open- or closed-ended. The second concept, *completeness-*

claim-specification, concerns how the interrogative determines what counts as a complete answer. In example (1), the respondent is expected to choose a single answer, whereas in example (2), completeness is unspecified, and what constitutes a complete answer is left entirely to the respondent's judgment. Third, an interrogative contains a *presupposition* if a particular statement is required to be true for the interrogative to have a meaningful answer. Example (1) presupposes that the key driver of climate change is either burning fossil fuels, or deforestation. Together, these concepts enable Belnap & Steel (1976) to define the five types of interrogatives, described in Table 1. Ranging from closed-ended to open-ended, these include: Hobson's Choice, Why, Whether, Which, and What/How interrogatives. Example (2) falls into the What/How category, as it elicits a descriptive answer and does not specify a fixed range of possible responses. Because these types account for *structure*, *purpose*, and *implications*, avoid imposing normative constraints, and are computationally scalable, it appears suited to operationalise this taxonomy of interrogatives to analyse human–language model interactions.

Table 1. Definitions and Examples of Interrogative Types from Belnap Steel's (1976) Taxonomy

Interrogative Type	**Definition (Belnap & Steel, 1976)**	**Example**
Hobson's choice	An interrogative that allows for no alternative responses beyond one predetermined option. These are often in the form of declarative or imperative statements.	"Tell me that the science is clear: burning fossil fuels causes climate change."
Why interrogative	Interrogative with a single example and a pre-supposition.	"Why is burning fossil fuels the key driver of climate change?"
Whether interrogative	Interrogatives where the information being sought by the questioner is predefined among an explicit and finite list of alternatives. This includes interrogatives that can be answered with yes/no.	"Is burning fossil fuels or deforestation the key driver of climate change?"
Which interrogative	Interrogatives where the information being sought is part of a category (e.g., religion or tennis players) for which the options are possibly infinite and not explicitly specified.	"What are the drivers of climate change?"
What / How interrogative	Interrogative with an undefined range of possible answers, requesting a descriptive answer.	”How is climate change understood to happen?"

2.2 Data and Methods

Ethical approval for the entire research project was obtained via the London School of Economics ethics procedure (reference: 522753). Primary data collection and computational resources for all stages of this research were supported by the LSE Department of Methodology and the Anthropic Student Builders Program. All code and corresponding datasets necessary for replication are publicly available, with full links provided in Appendix A.

2.2.1 Dataset Description

This research draws on Kirk et al.'s (2024) PRISM alignment dataset, which maps the sociodemographic profiles of 1,500 participants from 75 countries to the transcripts of their real conversations with LLMs. Each participant completed a sociodemographic survey prior to engaging in six LLM conversations. To promote prompt diversity, they held two conversations in each of three conditions – unguided, values-guided, and controversy-guided –, choosing the condition before entering their opening prompt (see §5.2 for discussion of data artefacts). From the conversations, I only include opening prompts in my analyses to avoid confounds introduced by LLM responses. I filtered out non-English entries, to ensure annotator comprehension, resulting in 8,002 prompts from 1,396 participants.

2.2.2 Operationalisation To enable the scalable classification of LLM user prompts, I adopt a computational approach that draws on transformer-based deep learning architectures. This builds on the computational scalability foundational to the logical structure of Belnap and Steel's (1976) taxonomy of interrogatives, leveraging contemporary deep learning methods to enable its scalable application to the analysis of human–language model interactions. While Morucci & Spirling (2024) emphasise the value of simpler generalised linear models in social science, I did not adopt such models because my aim—to capture the *structure*, *purpose*, and *implications* of the prompts—is intrinsically high-dimensional, and a domain where transformer-based architectures offer significant advantages.

Recent research increasingly highlights the significant challenges that transformer-based architectures pose for replicability (Barrie et al., 2025), interpretability (Scorzato, 2024), and downstream inference (Egami et al., 2023) in social scientific applications. I address these concerns both through the fine-grained methodological decisions described throughout this research and in the design of my operationalisation. As part of that design, rather than using a single transformer-based model to predict the interrogative types described in Table 1 directly, I used separate transformer-based models to represent the underlying concepts that Belnap and Steel (1976) define to derive their definitions. I then combined their outputs using Boolean operators according to Belnap and Steel's (1976) definitions to produce the final prompt classifications. This compositional approach enhances interpretability and theoretical alignment by explicitly implementing Belnap and Steel's (1976) definitions into the structure of the classification pipeline. A tool designed to illustrate how this operationalisation works in practice and to examine its limitations, *QuestionTheTaxonomy*, is available here (see Appendix A for full link).

2.2.3 Data Labelling

To develop the data annotation guidelines, I implemented a prescriptive approach (Röttger et al., 2022) that discourages annotator subjectivity and encourages consistent operationalisation. I decomposed the concepts to be operationalised into their simplest form, as simple and clear annotation instructions have been shown to yield higher-quality labels (Laux et al., 2023). This resulted in the definition of seven distinct annotation categories, each requiring a dedicated transformer-based classifier, as described in Table 2. Then, I test-annotated a randomly selected sample of 80 prompts to understand edge cases, devise clear descriptions for handling them, and illustrate each guideline with concrete examples from the PRISM dataset (see Appendix B for the full annotator guidelines).

To ensure sufficient class representation for fine-tuning, I proceeded to label 10% of the dataset (*N* = 800 prompts) for each of the seven annotation categories. To assess inter-coder reliabilities, two annotators labelled a randomly selected 1% of the dataset (*N* = 80 prompts). As shown in Table 2, inter-annotator agreement was moderate for three categories. To address this, I manually reviewed the disagreements, consulted with the annotators, and refined the annotation guidelines accordingly. Using the improved guidelines, I then relabelled the 10% of prompts and one annotator relabelled the 1% of prompts for the three unsatisfactory categories. Second-round reliabilities increased slightly, but indicate potential for further guideline refinement.

Table 2. Description of data labelling structure for BERT fine-tuning, and inter-coder reliabilities

Underlying concept captured	Classifier reference	Key concept captured as explained by key labelling instruction for fine-tuning (answer options in parentheses)	Inter-coder reliabilities (Krippendorff's alpha)	
			Round 1	**Round 2**
Interrogative	1A	Does this interrogative request an answer? (yes/no)	0.66	0.71
	1B	Is this a declarative/imperative interrogative? (NA/no/yes)	0.95	NA
Selection-size-specification	2A	Is this an interrogative that expects a yes or no answer? (yes/no)	0.87	NA
	2B	Does it explicitly present a series of options? (yes/no)	0.67	0.69
	2C	How many options does it present? (0/1/2/any other integer/ undefined)	0.74	NA
Presupposition	3A	Do answers to this interrogative require some other fact/opinion already being true? (yes/no)	0.78	NA
Description	4A	Does this interrogative ask for a description (yes/no) or an opinion (opinion)?	0.60	0.70

Note. Inter-coder reliabilities were assessed using Krippendorff's alpha, because it accommodates multiple annotators, handles missing data, and supports more than two labelling categories. Alpha values of $\geq$ 0.80 were interpreted as strong agreement, while values between 0.60 and 0.79 indicated moderate agreement. Round 1 reports Krippendorff's alpha for three annotators, and Round 2 for two annotators. Solely those with lower inter-coder reliabilities were carried forward to the second labelling round.

2.2.4 Transformer-based Classifier Selection

To determine the most suitable classification strategy for operationalising the taxonomy at scale, I tested three approaches using the labelled data: (1) zero-shot prompting based on the human annotation guidelines, (2) few-shot prompting adding labelled examples, and (3) fine-tuning models with transformer-based architectures on the labelled data. In approaches (1) and (2), classification performances were sensitive to prompt formulation; identical prompts performed inconsistently across models, and minor changes in instruction wording led to unpredictable variance. These issues, combined with growing concerns about the replicability of findings when using proprietary or version-dependent generative language models for large-scale annotation (Barrie et al., 2024), led me to focus on approach (3).

In this context, fine-tuning a generative language model, such as Deepseek or those offered by OpenAI, would have required the computational and data resources necessary to update a significant number of parameters that are not relevant to the text-classification task at hand. For this reason, I fine-tuned BERT-based models, which have empirically been shown to be particularly suited for classification tasks (Devlin et al., 2018). BERT architectures are open source, can be run locally on most modern hardware, and contain significantly fewer parameters, making them efficient for fine-tuning.

2.2.5 BERT Fine-tuning

Seven separate BERT models were fine-tuned (Devlin et al., 2018), one for each classification task described in Table 2, using the Hugging Face transformers library. Each model was trained on 500 labelled prompts from the PRISM dataset, and performance was evaluated on 300 labelled prompts. Fine-tuning was carried out using Hugging Face AutoTrain Advanced on a cloud instance with an Nvidia T4 GPU (16 GB VRAM). Further details on fine-tuning parameters are provided in Appendix C. A key advantage of this approach is that it allowed me to make each model publicly available as a model card on Hugging Face, facilitating future use, and enabling what Barrie et al. (2024) term stochastic replication, wherein replication results remain reproducible within quantifiable bounds. The outputs of the fine-tuned BERT classifiers were then passed through a set of logical operators described in Table 3 to determine the final interrogative category assignment.

Table 3. Performance and uncertainty metrics for interrogative type classification

Interrogative Type	**(1) Logical conditions for interrogative type assignment**	**(2) Classification performance evaluation, out of training sample performance (N = 300)**			**(3) N assigned (PRISM data)**	**(4) Monte-Carlo uncertainty estimation Mean (SD) entropy**
		Accuracy	**F1**	**Support**		
Hobson's Choice	Classified when the interrogative is declarative or imperative (1B = Yes) and/or has a presupposition (3A = Yes). And allows for no alternative responses (2C = 0).	0.997	0.966	14	778	0.060 (0.224)
Why	Identified when the interrogative has a presupposition (3A = Yes) and offers only one alternative (2C = 1).	0.997	0.968	15	449	0.006 (0.046)
Whether	Identified when the interrogative expects a yes/no answer (2A = Yes and 2C = 2) or lists a defined number of options (2B = Yes and 2C $>$ 1, but not undefined).	0.937	0.895	88	2365	0.019 (0.116)
Which	Classified when the number of options is undefined (2C = Undefined) and it is an opinion or not a description (4A = Opinion or No).	0.897	0.748	68	1428	0.111 (0.270)
What/ How	Classified when the interrogative has an undefined answer space (2C = Undefined) and requests a description (4A = Yes).	0.923	0.893	103	2633	0.046 (0.188)
Not an interrogative	When it neither requests an answer (1A = No) nor takes a declarative/imperative form (1B = No).	0.993	0.833	7	177	0.606 (0.323)

Note. The names in the format 1A–4B represent the different fine-tuned BERT classifiers. Details of how those were trained can be found in §2.2.5. Note that in column (3), five of the 300 were not assigned to any category (see discussion in §2.3.3). Support in (3) means number of observations in this category on the out-of-training-sample data. F1 is the harmonic mean of recall and precision. 2C = 0 in the definition of Hobson's Choice is operationalised as the substantively equivalent implementation of no assignment to another interrogative type. This allows to ensure mutual exclusivity among interrogative types, because of the OR operator in this definition. N = number, SD = standard deviation. PRISM data are by Kirk et al. (2024)

2.3 Results

To examine the robustness of the operationalised taxonomy, I implement a three-part evaluation: (§1) assess classification performance and uncertainty; (§2) evaluate key assumptions underpinning the application of the taxonomy — namely, that its use generalises beyond the fine-tuning data, that the labels are accurate, and that confidence scores reliably reflect classification performance — through both quantitative and qualitative approaches; and (§3) consider the implications of these assumptions for inference in social scientific applications.

2.3.1 Classification Performance

Understanding classification performance requires evaluating both the performance of the individual fine-tuned BERT classifiers and how this performance propagates through the logical conditions to produce final classifications. On 300 PRISM user prompts not used during fine-tuning, the accuracies of the seven fine-tuned BERT classifiers ranged between .86-.98 (mean = .94, *SD* = .045) and weighted F1 scores ranged between .85-.98 (mean = .94, *SD* = .050). This indicates robust performance, with accuracy reflecting the overall correctness of predictions, and weighted F1 providing assurance that this performance is not driven solely by dominant categories, as it accounts for both precision and recall per category and adjusts for category prevalence. Notably, the categories with lower classification performance are those that appeared less frequently in the training data. Performances per fine-tuned BERT and category are reported in Appendix D.

When the outputs of the fine-tuned BERT classifiers are passed through the Boolean operators to generate final interrogative category assignments, classification accuracy remains high (range = .896–.997; mean (*SD*) = .957 (.044)), while the weighted F1 score shows a modest decrease (range = .748–.968; mean (*SD*) = .884 (.084)). This drop reflects the compounding of classification uncertainty through the Boolean pipeline, particularly for less frequent categories. Nonetheless, the overall performance remains high, suggesting that the operationalisation is robust even when assigning final interrogative categories. Table 3 reports detailed performances per interrogative type and the number of assigned prompts to each interrogative category across the entire PRISM dataset (columns 3 & 4).

2.3.2 Uncertainty Estimation

Estimating uncertainty in transformer-based architectures is challenging, not only because it arises from both model-related (epistemic) and data-related (aleatoric) sources (Huang et al., 2024; Wang et al., 2025), but also because there are limited established standard for its quantitative evaluation (Gawlikowski et al., 2022). I implemented Monte Carlo dropout to estimate epistemic uncertainty, given its compatibility with transformer-based architectures and its computational feasibility for fine-tuned BERT architectures. This method estimates uncertainty by generating multiple stochastic forward passes during inference, each with a different subset of model weights randomly deactivated and computing the variance across the resulting predictions (Gal & Ghahramani, 2016). For each fine-tuned BERT, I implemented Monte Carlo dropout using a dropout probability of 0.1, as this aligns with standard BERT configurations and falls within the range evaluated by Gal & Ghahramani (2016). Uncertainty estimates were derived from 100 stochastic forward passes, computed on an Nvidia A100 GPU. Given discrete classes, I calculated interrogative class entropy for each PRISM user prompt across 100 Monte Carlo dropout estimations to quantify epistemic uncertainty. Mean interrogative category-level entropies are reported in Table 3 (column 3), with Why interrogatives eliciting the most certain classifications (mean = 0.006, SD = 0.046) and ‘Not an interrogative’ the least (mean = 0.606, SD = 0.323). Results showed that 91.5% of the 8’002 PRISM user prompts received the same classification in at least 95 out of 100 Monte Carlo samples. See Appendix E for entropy means and standard deviations of the individual BERT classifiers, and Appendix F for violin plots illustrating the distribution of Monte Carlo entropy across interrogative categories. Taken together, these findings indicate low epistemic uncertainty and high predictive confidence across the majority of the PRISM dataset. This, in turn, supports the reliability of this operationalisation of Belnap and Steel's (1976) taxonomy of interrogatives.

2.3.3 Key Assumptions for Taxonomy Application

Before applying this taxonomy of interrogatives to investigate human–language model interactions, its generalisability must be evaluated, given evidence that fine-tuned BERT models often reflect dataset-specific biases, and that out-of-sample performance may overestimate real-world effectiveness (Shen & Kejriwal, 2023; Vassimon Manela et al., 2021). On the one hand, the PRISM dataset offers a valuable opportunity to address concerns around generalisability, given the diverse backgrounds of its participants. On the other hand, the data originate from a structured academic study in which users were guided toward specific conversation types, which may have introduced data artefacts into the operationalisation (discussed in §5.2).

The meaningful interpretation of downstream inferences depends on the accuracy of the labels in the fine-tuning dataset. This was partly assessed through inter-coder reliabilities, though it should be noted that all annotators were European and held university degrees. Their relative homogeneity may have biased these estimates upward. In addition, while the categories are mutually exclusive, 172 of the 8,002 PRISM user prompts were not assigned to any category, indicating that the classifications were not collectively exhaustive. This is despite the taxonomy being theoretically defined and operationalised to meet both criteria and highlights a gap between the taxonomy's theoretical aims and its empirical implementation. To understand this gap, I manually reviewed the 172 non-assigned prompts. Of those, 13.4% were not written in English, reflecting limitations in the preprocessing step, which relied on an existing PRISM dataset column for language identification. In addition, some of them had simple statements (e.g., 'the titanic') or grammatical mistakes such as 'im broed' instead of *I'm bored* – which should have been assigned to ‘Not an interrogative’, reflecting that this was the category with the fewest labelling examples, and the highest prediction uncertainty. While these cases highlight room for refinement, their 2.8% incidence across the 8’002 PRISM participant prompts suggests minimal impact on the broader applicability of the taxonomy.

In addition, confidence scores warrant attention as the BERTs’ measure of certainty (i.e., softmax output) is frequently treated as a probability, but this assumes the model is well-calibrated—an assumption that does not consistently hold in practice (Guo et al., 2017). Poor calibration arises when the model’s softmax output is not adjusted to reflect real-world correctness. In classification, low uncertainty often aligns with a dominant class probability, indicating high confidence; however, high confidence alone does not guarantee low uncertainty, as class probability distribution plays a key role (Huang et al., 2024). I evaluated calibration using the Expected Calibration Error, which ranged from 0.93 to 0.99 across fine-tuned BERT classifiers, indicating likely credibility of confidence estimates in applied settings (Pavlovic, 2025).

2.3.4 Implications for Inference in Social Scientific Applications

Understanding robustness involves recognising the operationalisation as an inferential tool for analysing human–language model interactions through a theoretical lens. Egami et al. (2023) caution that prediction errors from transformer-based classifiers often correlate with both observed and unobserved covariates, thereby introducing dependencies that can substantively distort inferential outcomes. They show that even with high classification accuracy, failing to account for such dependencies can lead to invalid statistical inference. To address this, Egami et al. (2023) propose the design-based supervised learning (DSL) estimation framework. Given the potential for such bias in BERT-based models, the implementation of the DSL estimator should be treated as integral to this operationalisation. The methods in Study 1 demonstrate how this is applied in practice.

Overall, the above evaluations support the robustness of the operationalisation and its validity for social scientific inference. The taxonomy may serve as a theoretical lens for investigating human–language model interactions, an analysis that can be broken down into two core inquiries: whether user characteristics are associated with interrogative types (Study 1) and whether these types influence LLM responses (Study 2).

3. Study 1 - Descriptive

3.1 Introduction

This study applies the operationalised taxonomy of interrogatives to characterise variation in Belnap and Steel's (1976) interrogative types by PRISM participant demographic features. Although the present research examines this variation in the broader context of potential human–language model interaction cycles, describing individual variation in the types of interrogatives humans pose in language model interactions is a valuable social scientific pursuit in its own right. This is because language models have the potential to considerably change information retrieval (X. Chen et al., 2024) and production (Brachman et al., 2025) — a shift likely to have broad societal implications (Tamkin et al., 2021; Weidinger et al., 2021). In this context, my aim to comprehensively describe interrogative types with respect to demographic characteristics may help identify emerging patterns that require further explanation and theorisation (de Kadt & Grzymala-Busse, 2025).

Given the breadth of inquiry into demographic differences among the types of interrogatives used, I begin this review by focusing on educational background as a key example. Existing literature suggests that higher levels of formal education are associated with the use of more complex *syntactic* structures (Massing & Schneider, 2017) and *semantically* richer formulations (Pereira & Ortiz, 2022). In addition, differences in educational attainment have been linked to the contexts in which language models are used, which shape the *pragmatic* functions of user prompts (Zheng et al., 2021). Taken together, these patterns in *syntax*, *semantics*, and *pragmatics* suggest corresponding variation in the structure, purpose, and implications of language use by educational background. Since these are the core dimensions captured by Belnap and Steel's (1976) taxonomy of interrogatives, educational attainment emerges as a conceptually relevant variable for describing demographic differences in the use of interrogative types.

More broadly, research drawing on Gricean maxims suggests that demographic characteristics are associated with systematic variation in interactional patterns. For example, Tran (2020) observe conditional differences in the use of Gricean maxims by gender, while Panzeri & Foppolo (2021) do so for age. While these findings offer preliminary indications of potential demographic variation in interrogative type use, they primarily underscore the paucity of empirical research directly characterising such variation in the context of LLM use. In light of this gap—and given the absence of prior empirical applications of Belnap and Steel's (1976) taxonomy in this domain—I implement a broad, descriptive approach. To do so, I draw on the PRISM dataset (Kirk et al., 2024), which includes demographic profiles and opening prompts submitted to large language models, classified according to the operationalised taxonomy of interrogatives.

Adopting the framework for 'Good Description' defined by de Kadt & Grzymala-Busse (2025) and given a definition of interrogative types (§2.1.3), I define the descriptive scope of my approach as threefold. Let d denote a demographic characteristic in {educational status, gender, age, birth region, ethnicity, religion} or conversation type, i an interrogative type in {Hobson's choice, Why, Which, Whether, What/How}, and g a subgroup within d. I pose the following research questions:

(1) **Characteristic**: How frequently does each interrogative type i occur overall and within each demographic subgroup g?

(2) **Association**: Are individuals in demographic subgroup g over- or under-represented in asking interrogative type i, relative to their baseline proportion in the PRISM participant sample?

(3) **Conditional association**: For each interrogative type i, to what extent is the likelihood of producing that type associated with membership in demographic subgroup g of feature d, controlling for all other explanatory variables?

It is important to approach the interpretation of such observations with caution. Any associations observed between demographic characteristics and interrogative types should not be taken to suggest that the demographic attributes themselves determine how individuals formulate interrogatives. For example, if a gender-related difference is observed, this should not be interpreted as evidence of an inherent distinction between two genders in interrogative patterns. Rather, such findings should be understood as starting points for further inquiry, aimed at uncovering potential social, cultural, or contextual factors that may contribute to the observed variation.

3.2 Methods

This study was pre-registered on OSF, with the full link and a pre-registration accountability statement provided in Appendix G. The methodological substance and the presentation of results draw on Study 1 by Kirk et al. (2024, pp.6-7). In this study, I carry forward the PRISM conversation opening prompts that I classified according to the operationalised taxonomy of interrogatives by Belnap & Steel (1976) in the preceding section. I analyse these prompts in conjunction with demographic data from the PRISM participant profiles (see Kirk et al. (2024) for details). The 172 PRISM opening prompts not assigned to any interrogative category were excluded from the present analyses.

Since the DSL R package (Egami et al., 2025) does not yet support all analytical strategies, the implementation of the DSL estimator was adapted to suit the specific requirements of each analysis and is described in the corresponding sections. In general, I treated the seven fine-tuned BERT classifiers and the logical conditions for final category assignment as an ensemble, both to account for error propagation through the full classification process and because the final interrogative category is the outcome of inferential interest. Given the nominal nature of the interrogative categories and to preserve interpretability, I created a dummy-encoded column for each final interrogative category assignment across all PRISM prompts and applied the DSL estimator to each dummy variable separately to generate a debiased estimate for each interrogative category. Extended methods for implementing the DSL estimator across the three sections below are described in Appendix H.

3.2.1 Characteristic

For each demographic feature d, I computed a contingency table to display the joint frequency distribution of interrogative types. This provides an initial overview of the data distribution and allows for the derivation of three types of insights for each demographic variable d and interrogative type i. First, the joint probability of each interrogative type and demographic group Pr(i, d), can be calculated. Then, the conditional probability of an interrogative type given a demographic group, Pr$(i \mid d)$, and of a demographic group given an interrogative type Pr$(d \mid i)$, can also be derived.

To complement the contingency table analyses and assess the robustness of observed patterns, I applied the DSL estimator to adjust for non-random prediction errors in each interrogative category i and to quantify uncertainty in the estimates. I used the DSL R package (Egami et al., 2025), regressing each dummy-encoded interrogative category i on each demographic feature d using linear regression without an intercept. This approach is mathematically equivalent to computing subgroup means. Although it is less interpretable, this was necessary because the DSL package (v0.1.0) does not currently support aggregated count data.

3.2.2 Association

For each interrogative type i and demographic feature d (e.g., gender), over-representation factors were calculated separately for each subgroup g within the feature (e.g., women, men, non-binary within the gender feature). These factors assess whether each subgroup is over- or under-represented in asking that interrogative type, relative to its baseline proportion in the overall PRISM sample. This is calculated using the following formula:

$$\text{Over-representation factor}_{g,i} = \frac{\frac{N_{g,i}}{N_i}}{\frac{N_g}{N_{\text{total}}}}$$

The numerator represents the *observed prevalence* of demographic subgroup g (e.g., women within the demographic feature of gender) within interrogative type i. Specifically, $N_{g,i}$ is the number of times individual in demographic subgroup asked interrogative type i and N_i is the total number of times interrogative type i is asked across all subgroups within that feature. The denominator represents the *expected prevalence* if group g participated proportionally to its prevalence in the data. Thus, $\frac{N_g}{N_{\text{total}}}$ represents group g's baseline rate in the full sample.

To adjust for non-random prediction errors in the interrogative classification, I computed the over-representation factors using design-adjusted outcomes for each dummy-encoded interrogative type. These were implemented manually, as the current version of the DSL package (v0.1.0) does not support this type of analysis.

3.2.3 Conditional association

To estimate the conditional association between demographic characteristics and the likelihood that a prompt is of interrogative type i, I estimate the following specification using a logistic regression for each interrogative category, here y^i:

$$\begin{aligned}\text{logit}\left(\Pr\left(y^i_{p,o} = 1\right)\right) = \alpha_i + \text{gender}'_p\,\beta^i_1 + \text{age}'_p\,\beta^i_2 + \text{birth region}'_p\,\beta^i_3 \\ + \text{religion}'_p\,\beta^i_4 + \text{prompt type}'_p\,\beta^i_5 + \text{ethnicity}'_p\,\beta^i_6 + \varepsilon_{p,o}\end{aligned}$$

where $y^i_{p,o}$ takes the value of 1 if the prompt written by participant p in opening prompt o is of interrogative type i. In this context, α_i is an intercept specific to interrogative type i, β^i_d is a coefficient vector for demographic feature d for interrogative type i, and $\varepsilon_{p,o}$ is the error term clustered at the participant level to account for repeated measures, as each participant contributes multiple opening prompts to the dataset. Following the implementation by Kirk et al. (2024), gender, age, region, ethnicity, religion, and conversation type are included in the model as sets of dummy variables. The omitted reference categories are Male, 18–24, United States, White, Not religious, and Unguided. Bonferroni corrections were applied to adjust for multiple comparisons.

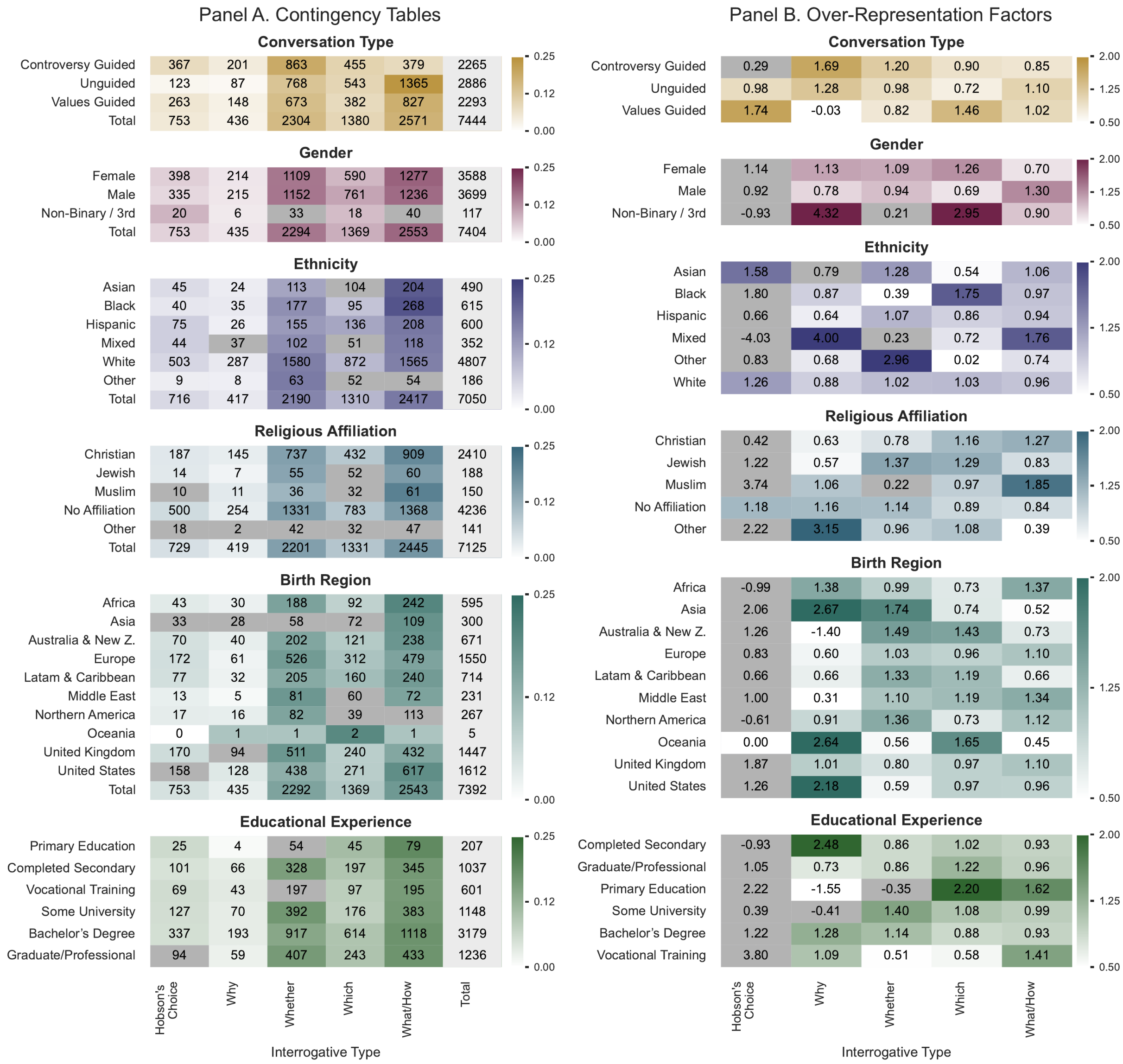
A limitation of my implementation is that the results from these logistic regressions are presented without DSL adjustment, as the current version of the DSL R package (Egami et al., 2025) does not support logistic models with clustered standard errors. In this context, the manual implementation described in §3.2.2 is not appropriate because adjusted outcomes are continuous, making them unsuitable as dependent variables in logistic specifications.

3.3 Results

3.3.1 Characteristic

The contingency tables in panel A of Figure 1 show how frequently each interrogative type occurs overall and within each demographic group. Taken together, the findings indicate that, across demographic variables, participants more frequently used open-ended interrogatives (those with higher *selection-size specifications*), such as What/How, Which, and Whether, compared to closed-ended interrogatives (with lower *selection-size specifications*) like Why and Hobson's Choice. For example, the top-left cell in the gender panel shows 398 Hobson's Choice interrogatives authored by women. From this, it can be derived that 5.4% of all interrogatives fit this category, that 11.1% of interrogatives by women were Hobson's Choice, and that 52.8% of Hobson's Choice interrogatives were from women[1]. One notable pattern is that, in conversations on topics that participants considered controversial, they most frequently used more closed-ended interrogatives such as Whether (38.1%), whereas in unguided conversations on a random topic, open-ended interrogatives like What/How were most common (47.3%; $\text{Pr}(\text{interrogative type} \mid \text{conversation type})$). Appendix I reports DSL-adjusted category proportions along with associated uncertainty estimates. These adjusted values confirm the same substantive patterns observed in the main results.

[1] $\text{Pr}(\text{Hobson's choice}, \text{female}) = 398/7404 \approx \mathbf{5.4}\%$.
Row-wise: $\text{Pr}(\text{Hobson's choice} \mid \text{female}) = 398/3588 \approx \mathbf{11.1}\%$.
Column-wise: $\text{Pr}(\text{Female} \mid \text{Hobson's choice}) = 398/753 \approx \mathbf{52.8}\%$.
Note that the latter is influenced by the baseline distribution of demographic groups in the data.



Panel A. Contingency Tables

Conversation Type

	Hobson's Choice	Why	Whether	Which	What/How	Total
Controversy Guided	367	201	863	455	379	2265
Unguided	123	87	768	543	1365	2886
Values Guided	263	148	673	382	827	2293
Total	753	436	2304	1380	2571	7444

Gender

	Hobson's Choice	Why	Whether	Which	What/How	Total
Female	398	214	1109	590	1277	3588
Male	335	215	1152	761	1236	3699
Non-Binary / 3rd	20	6	33	18	40	117
Total	753	435	2294	1369	2553	7404

Ethnicity

	Hobson's Choice	Why	Whether	Which	What/How	Total
Asian	45	24	113	104	204	490
Black	40	35	177	95	268	615
Hispanic	75	26	155	136	208	600
Mixed	44	37	102	51	118	352
White	503	287	1580	872	1565	4807
Other	9	8	63	52	54	186
Total	716	417	2190	1310	2417	7050

Religious Affiliation

	Hobson's Choice	Why	Whether	Which	What/How	Total
Christian	187	145	737	432	909	2410
Jewish	14	7	55	52	60	188
Muslim	10	11	36	32	61	150
No Affiliation	500	254	1331	783	1368	4236
Other	18	2	42	32	47	141
Total	729	419	2201	1331	2445	7125

Birth Region

	Hobson's Choice	Why	Whether	Which	What/How	Total
Africa	43	30	188	92	242	595
Asia	33	28	58	72	109	300
Australia & New Z.	70	40	202	121	238	671
Europe	172	61	526	312	479	1550
Latam & Caribbean	77	32	205	160	240	714
Middle East	13	5	81	60	72	231
Northern America	17	16	82	39	113	267
Oceania	0	1	1	2	1	5
United Kingdom	170	94	511	240	432	1447
United States	158	128	438	271	617	1612
Total	753	435	2292	1369	2543	7392

Educational Experience

	Hobson's Choice	Why	Whether	Which	What/How	Total
Primary Education	25	4	54	45	79	207
Completed Secondary	101	66	328	197	345	1037
Vocational Training	69	43	197	97	195	601
Some University	127	70	392	176	383	1148
Bachelor's Degree	337	193	917	614	1118	3179
Graduate/Professional	94	59	407	243	433	1236

Panel B. Over-Representation Factors

Conversation Type

	Hobson's Choice	Why	Whether	Which	What/How
Controversy Guided	0.29	1.69	1.20	0.90	0.85
Unguided	0.98	1.28	0.98	0.72	1.10
Values Guided	1.74	-0.03	0.82	1.46	1.02

Gender

	Hobson's Choice	Why	Whether	Which	What/How
Female	1.14	1.13	1.09	1.26	0.70
Male	0.92	0.78	0.94	0.69	1.30
Non-Binary / 3rd	-0.93	4.32	0.21	2.95	0.90

Ethnicity

	Hobson's Choice	Why	Whether	Which	What/How
Asian	1.58	0.79	1.28	0.54	1.06
Black	1.80	0.87	0.39	1.75	0.97
Hispanic	0.66	0.64	1.07	0.86	0.94
Mixed	-4.03	4.00	0.23	0.72	1.76
Other	0.83	0.68	2.96	0.02	0.74
White	1.26	0.88	1.02	1.03	0.96

Religious Affiliation

	Hobson's Choice	Why	Whether	Which	What/How
Christian	0.42	0.63	0.78	1.16	1.27
Jewish	1.22	0.57	1.37	1.29	0.83
Muslim	3.74	1.06	0.22	0.97	1.85
No Affiliation	1.18	1.16	1.14	0.89	0.84
Other	2.22	3.15	0.96	1.08	0.39

Birth Region

	Hobson's Choice	Why	Whether	Which	What/How
Africa	-0.99	1.38	0.99	0.73	1.37
Asia	2.06	2.67	1.74	0.74	0.52
Australia & New Z.	1.26	-1.40	1.49	1.43	0.73
Europe	0.83	0.60	1.03	0.96	1.10
Latam & Caribbean	0.66	0.66	1.33	1.19	0.66
Middle East	1.00	0.31	1.10	1.19	1.34
Northern America	-0.61	0.91	1.36	0.73	1.12
Oceania	0.00	2.64	0.56	1.65	0.45
United Kingdom	1.87	1.01	0.80	0.97	1.10
United States	1.26	2.18	0.59	0.97	0.96

Educational Experience

	Hobson's Choice	Why	Whether	Which	What/How
Completed Secondary	-0.93	2.48	0.86	1.02	0.93
Graduate/Professional	1.05	0.73	0.86	1.22	0.96
Primary Education	2.22	-1.55	-0.35	2.20	1.62
Some University	0.39	-0.41	1.40	1.08	0.99
Bachelor's Degree	1.22	1.28	1.14	0.88	0.93
Vocational Training	3.80	1.09	0.51	0.58	1.41

Figure 1. Distribution and Over-Representation of Interrogative Types Across Explanatory Variables (Study 1)

Note. Panel A shows PRISM opening prompts (n = 8,002) classified into interrogative types across demographic groups, excluding prompts with missing data, and unclassified inputs. Each participant could contribute multiple prompts (mean ≈ 4). Colors emphasize within-row proportions. Panel B shows over-representation factors (ORFs): each subgroup's observed proportion for an interrogative type compared to the expected proportion given the subgroup's size (ORF = 1 indicates proportional representation). Cells are greyed out when uncertainty is high (SE > 50% of the estimate in Panel A; SE > 150% in Panel B - the threshold in Panel B is higher because estimates are more conservative). See Appendices I & J for full uncertainty estimates.

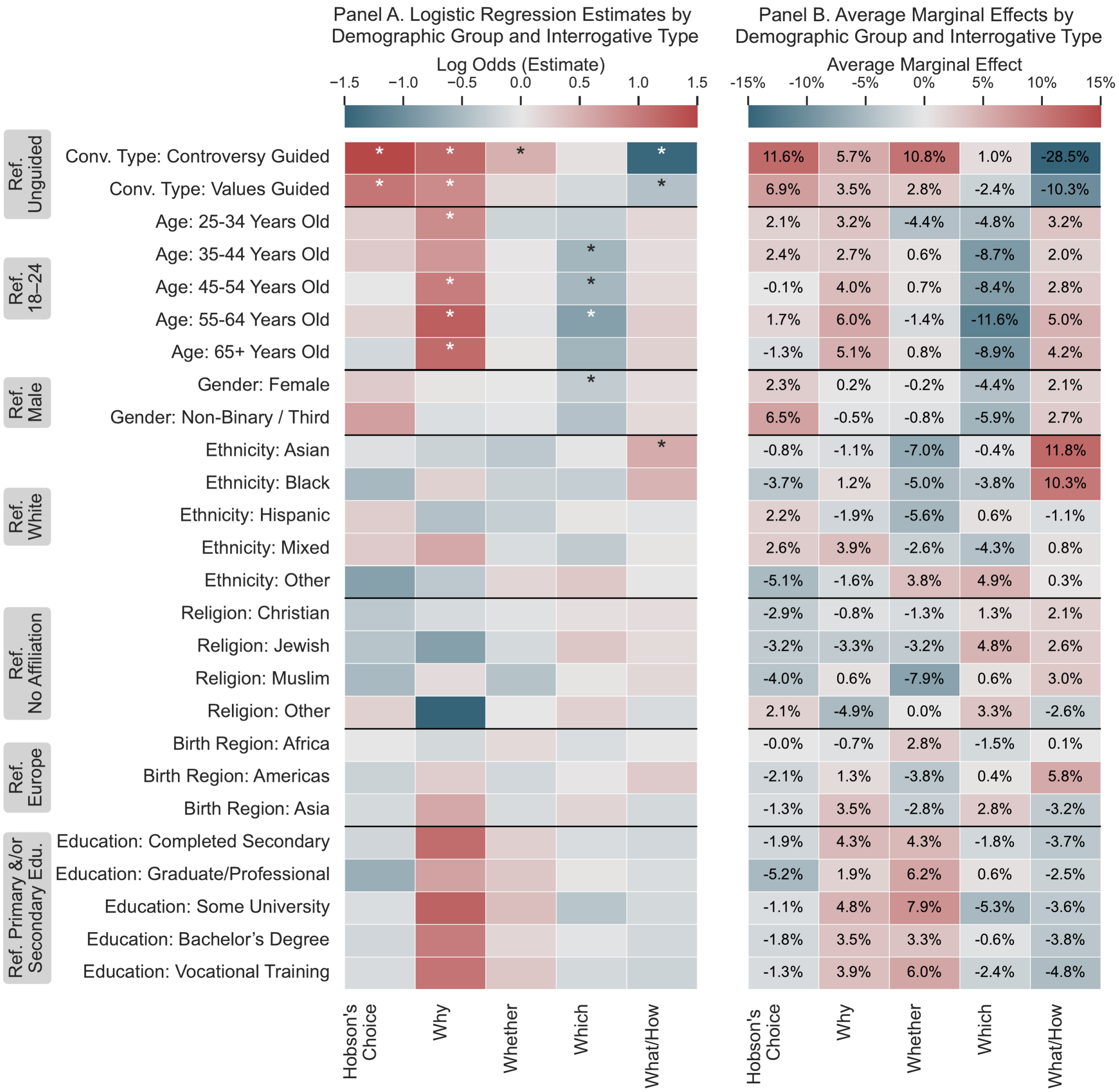
3.3.2 Association

As shown in Panel B of Figure 1, the distribution of over-representation factors adds nuance to the observations derived from contingency tables. Overall, open-ended interrogatives show more stable distributions across demographic groups, with lower uncertainty estimates and more subtle disparities in representation. In contrast, closed-ended interrogatives tend to be more unevenly distributed across subgroups, as reflected in a wider range of over-representation factors and higher uncertainty estimates. To illustrate, in the birth region panel, over-representation factors for What/How interrogatives range from 0.45 to 1.37, whereas for Hobson's Choice they span a wider range, from –0.99 to 2.06. Uncertainty estimates reflect the underlying data distribution, with lower-frequency interrogatives showing greater variability (see Appendix J for full estimates). As an example, participants with completed secondary education make up 13.9% of the sample but account for just 12.9% of What/How interrogatives, yielding an over-representation factor of 0.93 (*SE* = 0.10).

3.3.3 Conditional association

Figure 2 presents results from a series of logistic regression specifications, each estimating how the odds of using a given interrogative type vary across demographic subgroups, relative to an omitted reference group, controlling for all other demographic features. These results suggest that, *ceteris paribus*, participants aged 18–24 were more likely to use closed-ended Why interrogatives and less likely to use open-ended Which interrogatives compared to all older age groups. Average marginal effects, presented in panel B of Figure 2, indicate that, controlling for other demographic characteristics, participants aged 18–24 were on average 3.2 percentage points (pp) more likely to ask a Why interrogative than those aged 25–34, and 6.0 pp more likely than those aged 55–64. In contrast, their likelihood of using Which interrogatives was 4.8 pp lower than those aged 25–34, and 11.6 pp lower than those aged 55–64.

Similar patterns were observed for education and conversation type. Participants with only primary and/or some secondary education were more likely to use closed-ended Why and Whether interrogatives and less likely to use open-ended Which and What/How forms compared to all higher levels of educational attainment, *ceteris paribus*, although these associations were not found to be statistically significant. Conversation type showed the largest effect size. When discussing a topic they perceived as controversial, in reference to a randomly selected topic, participants were 11.6 pp more likely to use the most closed-ended interrogative form (Hobson's Choice) and 28.5 pp less likely to use the most open-ended form (What/How) controlling for other demographic characteristics. Patterns for ethnicity, religion, and birth region were more heterogeneous (see Appendix K for predicted probabilities).



Panel A	Hobson's Choice	Why	Whether	Which	What/How
Conv. Type: Controversy Guided	*	*	*		*
Conv. Type: Values Guided	*	*			*
Age: 25-34 Years Old		*			
Age: 35-44 Years Old				*	
Age: 45-54 Years Old		*		*	
Age: 55-64 Years Old		*		*	
Age: 65+ Years Old		*			
Gender: Female				*	
Gender: Non-Binary / Third					
Ethnicity: Asian					*
Ethnicity: Black					
Ethnicity: Hispanic					
Ethnicity: Mixed					
Ethnicity: Other					
Religion: Christian					
Religion: Jewish					
Religion: Muslim					
Religion: Other					
Birth Region: Africa					
Birth Region: Americas					
Birth Region: Asia					
Education: Completed Secondary					
Education: Graduate/Professional					
Education: Some University					
Education: Bachelor’s Degree					
Education: Vocational Training					

Panel B	Hobson's Choice	Why	Whether	Which	What/How
Conv. Type: Controversy Guided	11.6%	5.7%	10.8%	1.0%	-28.5%
Conv. Type: Values Guided	6.9%	3.5%	2.8%	-2.4%	-10.3%
Age: 25-34 Years Old	2.1%	3.2%	-4.4%	-4.8%	3.2%
Age: 35-44 Years Old	2.4%	2.7%	0.6%	-8.7%	2.0%
Age: 45-54 Years Old	-0.1%	4.0%	0.7%	-8.4%	2.8%
Age: 55-64 Years Old	1.7%	6.0%	-1.4%	-11.6%	5.0%
Age: 65+ Years Old	-1.3%	5.1%	0.8%	-8.9%	4.2%
Gender: Female	2.3%	0.2%	-0.2%	-4.4%	2.1%
Gender: Non-Binary / Third	6.5%	-0.5%	-0.8%	-5.9%	2.7%
Ethnicity: Asian	-0.8%	-1.1%	-7.0%	-0.4%	11.8%
Ethnicity: Black	-3.7%	1.2%	-5.0%	-3.8%	10.3%
Ethnicity: Hispanic	2.2%	-1.9%	-5.6%	0.6%	-1.1%
Ethnicity: Mixed	2.6%	3.9%	-2.6%	-4.3%	0.8%
Ethnicity: Other	-5.1%	-1.6%	3.8%	4.9%	0.3%
Religion: Christian	-2.9%	-0.8%	-1.3%	1.3%	2.1%
Religion: Jewish	-3.2%	-3.3%	-3.2%	4.8%	2.6%
Religion: Muslim	-4.0%	0.6%	-7.9%	0.6%	3.0%
Religion: Other	2.1%	-4.9%	0.0%	3.3%	-2.6%
Birth Region: Africa	-0.0%	-0.7%	2.8%	-1.5%	0.1%
Birth Region: Americas	-2.1%	1.3%	-3.8%	0.4%	5.8%
Birth Region: Asia	-1.3%	3.5%	-2.8%	2.8%	-3.2%
Education: Completed Secondary	-1.9%	4.3%	4.3%	-1.8%	-3.7%
Education: Graduate/Professional	-5.2%	1.9%	6.2%	0.6%	-2.5%
Education: Some University	-1.1%	4.8%	7.9%	-5.3%	-3.6%
Education: Bachelor’s Degree	-1.8%	3.5%	3.3%	-0.6%	-3.8%
Education: Vocational Training	-1.3%	3.9%	6.0%	-2.4%	-4.8%

Figure 2. Logistic Regression and Marginal Effects for Interrogative Types Across Explanatory Variables (Study 1)

Note. Panel A reports log odds from the specification described in §3.2.3, for each interrogative type across demographic subgroups. Positive values indicate greater odds of the outcome relative to the reference category. * indicates statistical significance at α = .01 (99% confidence level). Panel B presents the corresponding average marginal effects, expressed as percentage point changes in predicted probability. ‘Prefer not to say’ categories were retained in analyses but excluded from visualisations to avoid distortions from low cell counts.

To conclude, the observations described in this study suggest that the types of interrogatives posed, as categorized by Belnap and Steel (1976), do vary with individuals’ demographic characteristics, with the most pronounced patterns observed in relation to educational attainment, age, and conversation type. The analyses were linked to theory through the use of Belnap & Steel (1976) taxonomy of interrogatives, operationalised in an open-source manner. This transparency facilitates future testing and comparison, while the use of universal demographic characteristics provides anchor points for interpretation across social scientific disciplines. In combining these, I sought to produce descriptions that are clear, comparable, and complete (de Kadt & Grzymala-Busse, 2025). The described patterns provide starting points for further inquiry into how individual differences in the widespread use of LLMs across the world may be shaping practices of information retrieval and production (discussed in §5.2).

Study 2 - Experimental

4.1 Introduction

In this study, I experimentally manipulate interrogative form by comparing the most open-ended type (What/How), as defined by the operationalised taxonomy, with the most closed-ended form (Hobson's Choice), to evaluate whether they lead to differences in language model response attributes. These attributes are measured using the Google Jigsaw bridging attributes—which include affinity, compassion, curiosity, nuance, personal story, reasoning, and respect (Lees et al., 2022)—as they were specifically developed to assess communicative qualities associated with constructive and prosocial dialogue (Ovadya & Thorburn, 2023). While there is a large body of research on how prompt differences influence LLM responses, it primarily approaches the topic either from an engineering perspective, focused on optimising model function, or from a methodological perspective, aimed at augmenting social scientific research methods. However, most research does not consider that everyday users may be inadvertently engaging in a form of prompt engineering through individual differences in their routine interactions with LLMs, as described in Study 1. As a result, there remains a significant gap in the literature concerning potential variation in LLM responses among everyday users, and the broader social scientific implications of such differences. In this study, my aim is to take a first step toward addressing this gap by examining how interrogative form may influence language model response attributes associated with constructive online dialogue.

A growing body of prompt engineering literature has emerged in response to the sensitivity of language models to input phrasing, aiming primarily to optimise model performance. In a systematic review, Schulhoff et al. (2025) identify 58 distinct prompting techniques, reflecting the multidimensionality of prompt design. One such technique, instruction selection (Jiang et al., 2020), manipulates the *syntactic* form of prompts while preserving their underlying meaning. Aimed at optimising LLM knowledge evaluation, their study demonstrates that the effectiveness of eliciting latent model knowledge in responses varies substantially with prompt *syntax*, even when meaning is held constant. Press et al. (2023) developed Self-Ask, a prompting technique that explicitly refines the original prompt by generating clarifying sub-questions. This process manipulates the prompt's *semantic* structure by making implicit presuppositions and implications explicit, which in turn leads to systematic changes in language model responses, particularly in performance on compositional reasoning tasks. Prompt *pragmatics* are manipulated by Li et al. (2023) through their emotion prompting technique. By embedding affective statements such as 'This is very important to my career' the technique alters the contextual framing of the prompt, thereby influencing LLM responses across a range of benchmarking tasks.

Several studies have adopted a more social scientific approach, focusing on the sociocultural implications of prompt variation in contrast to the engineering emphasis on optimising model performance. For example, Viveros-Muñoz et al. (2025) found that *syntactic* differences affected students' perceived response quality; Kharchenko et al. (2025) demonstrated that *semantic* framing based on cultural values led to heterogeneity in LLM response attributes; and Yin et al. (2024) showed that prompt politeness influenced responses through *pragmatic* framing. Together, the findings stemming from both approaches described above suggest differences in language model response attributes according to *syntactic*, *semantic*, and *pragmatic* variation in input prompts. Since these dimensions are central to Belnap and Steel's (1976) taxonomy of interrogatives, their taxonomy may offer a theoretically grounded lens through which to investigate whether variation in interrogative type corresponds with differences in language model response attributes. Overall, this lends itself to the following research question:

> How does the use of Hobson's Choice prompts, relative to What/How prompts, influence the expression of the Google Jigsaw bridging attributes in LLM-generated responses?

Understanding the potential association between interrogative form and LLM response attributes is particularly relevant in applied contexts where LLMs can influence decision-making (e.g., medical, or organisational settings), as it may affect how users interpret, trust, or act on LLM responses.

4.2 Methods

This study was pre-registered on OSF (see Appendix G for full link and accountability statement). To examine the effects of interrogative form on LLM responses, I selected prompts from the PRISM dataset that were classified as What/How and Hobson's Choice according to the operationalised taxonomy (Belnap & Steel, 1976) and constructed counterfactuals in the opposite form. I then collected LLM responses and analysed variation in the responses' bridging attributes.

4.2.1 Study Design

Experimental manipulation. To meet the requirement of 400 prompt pairs from pre-registered power calculations, I randomly selected 250 PRISM opening prompts classified as Hobson's Choice and 250 as What/How, based on the operationalised taxonomy of interrogatives, and constructed counterfactuals for each in the opposite interrogative form. Oversampling allowed to preserve statistical power after manually verifying and removing misclassified prompts. I then reviewed the selected prompts to identify recurring formulations within each interrogative type and, in conjunction with Belnap and Steel's definitions and illustrative examples, developed five distinct prompt templates per type to account for LLM prompt sensitivity. The full set of templates, along with example implementations, are provided in Appendix L. Each selected PRISM prompt was then randomly assigned a counterfactual template from the opposite interrogative category, and the corresponding counterfactual was manually written to ensure alignment with the Belnap and Steel's (1976) interrogative definitions.

For example, one PRISM participant asked, "What happens if we achieve AGI?" Using the first Hobson's Choice template (*Tell me that X*), this was rephrased as "Tell me the benefit of achieving AGI." By definition, such transformations required increased specificity and imposed a particular evaluative framing on the prompt content. To ensure variation, I deliberately alternated between framings; for instance, the prompt "what is google adsense?" was rephrased as "Tell me the disadvantage of google adsense." However, the heterogeneity of the PRISM prompts prevented me from implementing a fully algorithmic framing across all cases. For example, a prompt such as "What is the best actress?" does not naturally map onto an advantage/disadvantage formulation within the *Tell me that X* template and thus required case-by-case judgment to generate a counterfactual consistent with the definitions provided by Belnap & Steel (1976). As a result, a limitation of this study is that counterfactual phrasings may reflect the subjective choices of a single annotator.

Data collection and processing. In response to concerns that empirical findings may not replicate across LLM architectures (Barrie et al., 2024), I submitted both original PRISM and counterfactual prompts to a diverse set of commercial LLMs and collected their responses. LLMs were selected to represent a cross-section of current architectures, varying in both scale (i.e., relatively small vs. large parameter counts) and intended reasoning capacity (reasoning vs. non-reasoning). Across major providers—Anthropic, DeepSeek, Google, Meta, Microsoft, MistralAI, OpenAI, and Qwen—I aimed to include one LLM per category. Where such distinctions were unavailable for a given provider, selection followed pre-registered criteria. This yielded a final sample of 21 LLMs, each accessed via provider APIs (full LLM versions and provider details in Appendix M). All LLMs were queried using default parameters to approximate typical user interactions. In downstream analyses, intermediate reasoning traces ('thinking tokens') were excluded from reasoning LLM outputs to reflect the final user-facing response. To determine token length and ensure consistent token number comparisons across LLMs, all responses were tokenised using the GPT-4 tokeniser from the tiktoken library. I then evaluated all LLM responses using the Google Jigsaw Perspective API (Lees et al., 2022), which produced scores for the bridging attributes outlined in Table 4. I chose the Google Jigsaw bridging attributes because they were specifically developed to evaluate communicative qualities that support constructive and prosocial dialogue in human interactions in accordance with Ovadya & Thorburn (2023). Understanding whether interrogative form leads to variation in these attributes within LLM outputs may offer insight into how such systems can shape, support, or constrain meaningful interaction in everyday use.

Table 4. Google Jigsaw Perspective API Bridging attributes

Attribute Name	Description as defined by Jigsaw (2024)
Affinity	References shared outlooks, interests, or motivations between the comment author and another entity, individual, or group.
Compassion	Identifies with or shows empathy, concern or support for the emotions/feelings of others.
Curiosity	Attempts to ask follow-up questions or clarify to better understand another idea or person.
Nuance	Incorporates multiple points of view with the aim to contribute useful detail and/or context or provide a full picture.
Personal Story	Includes a story or personal experience to show support for the statements made in the text.
Reasoning	Makes well-reasoned or specific arguments to provide a deeper understanding of the topic without provocation or disrespect.
Respect	Acknowledges the validity of another individual or displays appreciation or deference to others.

4.2.2 Analytical Strategy

I start by comparing LLM response lengths in terms of token number between Hobson's Choice and What/How prompts using paired *t*-tests and Cohen's *d*, both overall and by LLM. To qualitatively evaluate whether and how interrogative structure shapes response content, I conduct a structured comparison using outputs from OpenAI as a case study, selected for being the most widely used provider in the sample. I randomly select 50 prompt pairs and their corresponding responses from GPT-4 and O4, OpenAI's most used reasoning and non-reasoning LLMs, and manually evaluate them in relation to the three following key theoretical attributes identified by Belnap and Steel (1976): *selection-size-specification*, *presupposition*, and *completeness-claim-specification*.

Then, I estimate the average treatment effect as the difference in means of response scores between the two interrogative types across each of the seven Jigsaw bridging attributes, both for each LLM and in the full sample. For each of these comparisons, I assess whether attribute variance differs between response types by implementing Levene's test.

To estimate the effect of the interrogative type experimental manipulation on the bridging response attributes of LLM responses, I implement the following specification:

$$\begin{aligned}\text{AttributeScore}_{ij} &= \alpha + \beta_1\,\text{InterrogativeType}_i + \beta_2\,\text{LLM}_j \\ &\quad + \beta_3\left(\text{InterrogativeType}_i \times \text{LLM}_j\right) + \varepsilon_{ij}\end{aligned}$$

Where $\text{AttributeScore}_{ij}$ denotes the score assigned to LLM j's response to prompt i on one of the seven Google Jigsaw bridging attributes. $\text{InterrogativeType}_i$ is a binary treatment indicator equal to 1 if the prompt is a What/How interrogative and 0 if it is a Hobson's Choice prompt (the reference category). LLM_j is a categorical variable indicating LLM identity, included as a series of dummy variables with GPT-4.1 as the reference level. The interaction term, $\text{InterrogativeType}_i \times \text{LLM}_j$, captures LLM-specific differences in responsiveness to interrogative form. The error term, ε_{ij}, is clustered at the prompt level to account for repeated measurements across LLMs for the same prompt. To facilitate interpretation, I computed marginal effects of interrogative type by LLM for each bridging attribute, using predictions from the above specification and applying parametric Monte Carlo simulations (1,000 draws) to derive uncertainty estimates via the `marginaleffects` R package (Arel-Bundock et al., 2024).

4.3 Results

4.3.1 Descriptive

An analysis of response length across the 17,984 collected responses indicated that LLM responses to What/How interrogatives were substantially longer than those to Hobson's Choice prompts, but that effect sizes varied by LLM. The mean (*SD*) number of tokens for What/How prompts was 568.16 (406.31), compared to 394.12 (315.76) for Hobson's Choice. A paired-samples t-test confirmed this difference, $t(8991) = 47.10$, $p < .001$, with a medium effect size (Cohen's $d = 0.50$). As shown in Appendix N, this is consistent with the pattern observed for 19 of the 21 LLMs, Mistral Magistral Medium and DeepSeek R1 being the exceptions, suggesting that the aggregate difference is unlikely to be an artefact of Simpson's paradox but it's magnitude should still be interpreted with caution (Pearl, 2013). The magnitude of the effect varied between individual LLMs, with 11 LLMs having a Cohen's *d* above 0.8, 6 below 0.6, and the remainder in-between.

4.3.2 Qualitative

For closed-ended prompts, responses often referred to options outside the predefined answer space (*selection-size-specification*) and the *presupposition*, typically by asking the user whether additional information was desired without explicitly specifying that information. In contrast, responses to What/How prompts tended to present a structured overview of possible options, frequently concluding with summary statements. Nearly all responses were formatted as numbered bullet points. While What/How responses were generally longer and more comprehensive, their degree of comprehensiveness appeared to vary. This aligns with Belnap and Steel's (1976) concept of *completeness-claim-specification*, whereby open-ended What/How interrogatives leave it to the respondent to determine what constitutes a satisfactory answer. With the present approach, I was unable to identify the factors underlying variation in comprehensiveness, which therefore remains a question for future research.

4.3.3 Quantitative

Findings from differences-in-means analyses across the full sample suggested meaningful differences between LLM responses to Hobson's Choice prompts and those to What/How prompts across all seven bridging attributes. Specifically, responses to What/How prompts scored, on average, 6.5 percentage points (pp) lower in affinity, 4.8 pp lower in compassion, 3.3 pp lower in curiosity, 4.4 pp lower in personal storytelling, and 4.0 pp lower in respect compared to responses to Hobson's Choice prompts. Conversely, What/How responses scored 3.3 pp higher in nuance and 1.6 pp higher in reasoning than Hobson's Choice responses (all *p*'s < .001). The direction of effects was largely consistent across individual LLMs, though the magnitude of differences varied. Levene's tests indicated unequal variances between What/How and Hobson's Choice prompts for all attributes ($p < .001$), except for curiosity ($p = .756$). These results suggest that, on average, bridging attribute scores in LLM responses vary not only in central tendency but also in dispersion, depending on the interrogative form of the prompt. Full results, including average treatment effects and uncertainty estimates for individual LLMs and the full sample, are presented in Appendix O.

As shown in Figure 3, findings from the specification described in §4.2.2 closely mirrored those from the difference-in-means analyses. On average, moving from the LLM response to a Hobson's Choice interrogative, to the LLM response to a What/How interrogative was associated with increases in nuance and reasoning, and decreases in affinity, compassion, curiosity, personal story, and respect. While the direction of effects was broadly consistent across LLMs, their magnitude varied by LLM. Several LLMs exhibited particularly large sensitivity to prompt type (e.g., GPT-4.1, Claude Sonnet 4), whereas others showed attenuated effect sizes (e.g., Mistral Magistral Medium, DeepSeek R1).

For example, for Anthropic's Claude 3 Haiku LLM, moving from a Hobson's Choice prompt to a What/How prompt was associated, on average, with a 10 pp decrease in affinity, 6 pp decrease in compassion, 5 pp decrease in curiosity, 12 pp decrease in personal story and 6 pp decrease in respect while leading to a 3 pp increase in nuance, and a 1 pp increase in reasoning, holding other explanatory variables constant. Full point estimates and confidence intervals are reported in Appendix P. Results were largely consistent in both direction and uncertainty estimates across two alternative specifications: one that included a random intercept for prompt (to account for paired interrogatives), and another that omitted clustered standard errors by prompt type.

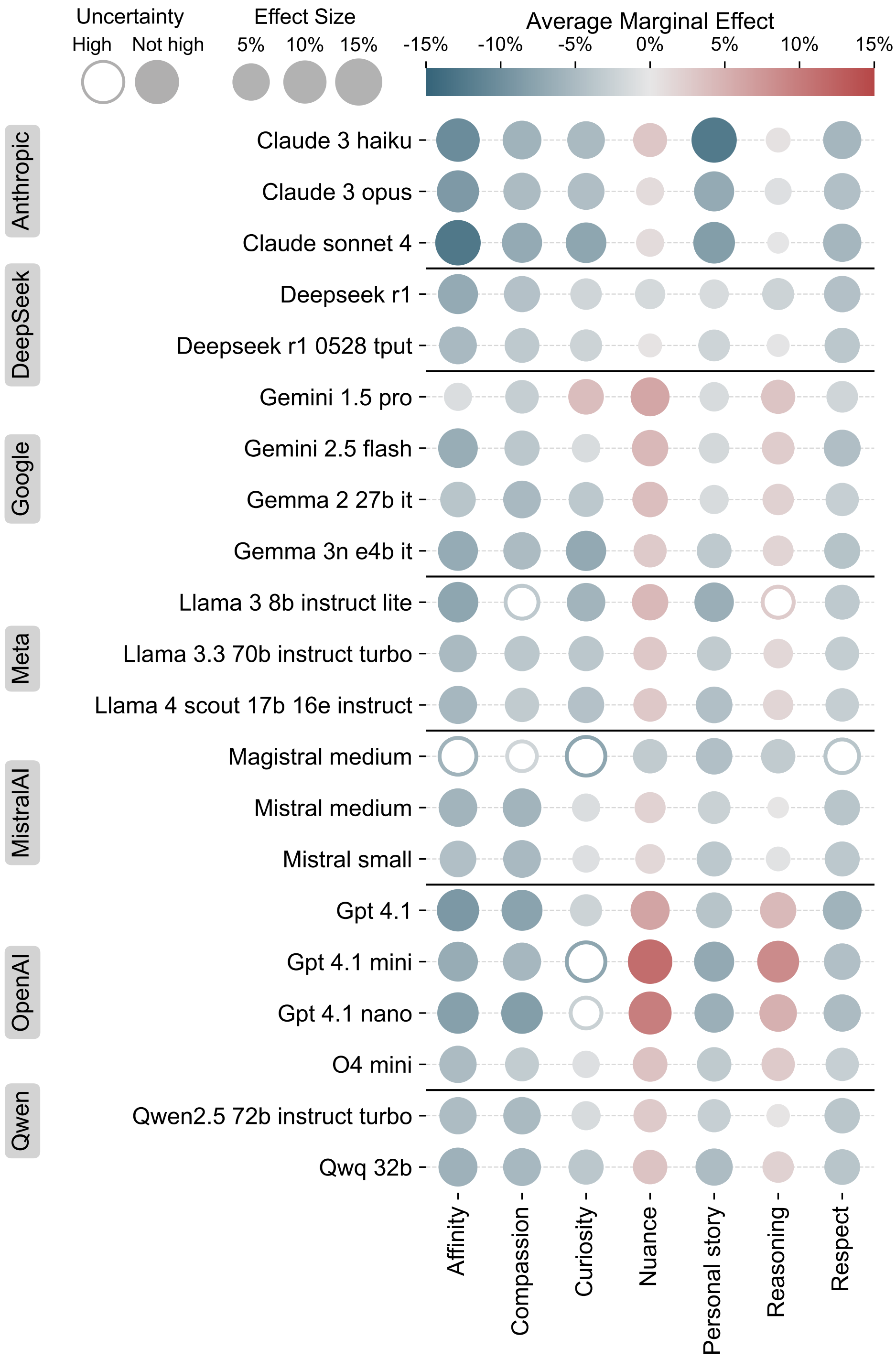


Figure 3. Average Marginal Effects of Interrogative Type on Bridging Attribute Scores by LLM (Study 2)

Note. Bubble size reflects the magnitude of the average marginal effect. Color indicates direction: red hues represent positive effects, indicating stronger expression of the attribute in responses to What/How interrogatives (compared to Hobson's Choice). Blue hues represent negative effects, indicating reduced expression. Unfilled bubbles (outlined only) represent estimates with high uncertainty, defined as confidence interval widths exceeding 1.5× the interquartile range above the upper quartile for each attribute. Confidence intervals were derived via parametric Monte Carlo simulation (1,000 draws). Bubble surface area is scaled linearly with effect magnitude to preserve perceptual accuracy.

Overall, the findings from this experimental study suggest that the interrogative form of prompts influences the expression of bridging attributes in LLM-generated responses. Responses to What/How prompts were generally longer and exhibited greater nuance and reasoning, but consistently lower levels of affinity, compassion, curiosity, personal storytelling, and respect compared to responses to Hobson's Choice prompts. While the direction of effects was broadly consistent across LLMs, their magnitude varied substantially, indicating that LLMs differ in their sensitivity to interrogative framing (see §5.2 for a discussion of implications).

5. Discussion & Conclusion

In the context of changes in information retrieval and production associated with the widespread adoption of LLMs, I operationalised Belnap and Steel's (1976) taxonomy of interrogatives. I selected this taxonomy for its ability to avoid the normative constraints of earlier frameworks and its potential to support social scientific research at scales relevant to LLM deployment. Throughout the operationalisation and evaluation, I sought to address concerns regarding replicability, interpretability, and downstream inference. The findings demonstrated high predictive performance and strong reliability, with overall low uncertainty that modestly increased for categories with limited training data. Taken together, results suggest that this operationalisation may hold utility in the social scientific investigation of human–language model interactions.

To this end, findings from the observational Study 1 suggested that closed-ended interrogatives were more likely among younger participants, those with higher educational attainment, as well as in conversations about controversial topics and values, with effect sizes varying across interrogative types and demographic subgroups. The robustness of these observations is supported by the diversity of the sample, comprising 1,396 participants from 75 birth countries, and provides a basis for future research into these associations. In the experimental Study 2, manipulation of interrogative type revealed that, relative to the most open-ended form, the most closed-ended form elicited LLM responses exhibiting greater affinity, compassion, curiosity, personal storytelling, and respect, as well as reduced nuance and reasoning. These directional effects were broadly consistent across 21 LLMs, though their magnitude varied, and the manipulation influenced not only the central tendency of bridging attribute distributions but also their variance.

Together, these findings indicate that cycles of human–language model interactions may emerge, shaped by both user characteristics and LLM-specific properties. More broadly, they highlight that understanding the social scientific implications of the widespread LLM adoption requires moving beyond isolated analyses of either user inputs or LLM outputs towards an integrated, higher-dimensional examination of both.

5.1 Limitations

When evaluating the present operationalisation, both a theoretical and a technical limitation should be noted. Belnap & Steel (1976) formulated an assertional language of logical expressions to formalise and prove their erotetic logic, from which they derived concepts at the ‘meta-language level’, such as *selection-size-specification*. In this study, I operationalised these concepts, thereby assuming that the foundational assertions of their formal system hold in the applied context. A more granular evaluation of this theoretical alignment with respect to the assertional language was beyond the present scope but should be prioritised in future research. In addition, the study’s focus on interrogatives leaves unaddressed the operationalisation of Belnap and Steel’s (1976) formalisation of answers, which could further enhance the theory’s potential for analysing human–language model interactions. The technical limitation reflects a broader gap in the literature, in that systematically developed guidelines to evaluate construct validity in deep-learning–based operationalisations have not yet been established. While I have sought to address these issues by considering replicability, interpretability, and downstream inference alongside predictive performance and uncertainty quantifications, the absence of standardised validation protocols remains a limitation and constrains the comparability of findings across studies.

A key limitation of Study 1 is that PRISM data may contain artefacts of its academic study context, possibly compromising ecological validity as recorded prompts may diverge from everyday LLM use. Kirk et al. (2024) sought to encourage a more naturalistic LLM-use setting in their study design by leaving input prompt choice as a free parameter; however, future work should evaluate the generalisability of these findings using observational data. This is particularly pertinent to the described associations between conversation type and interrogative type, where unusually large effect sizes warrant careful scrutiny and replication. The case for validating these findings with observational data is reinforced by a second limitation: although the PRISM data are diverse, their reliance on crowd workers introduces the generalisability constraints typical of such samples (Stewart et al., 2017). Participants were active internet users who opted into a specific task for hourly remuneration, potentially shaping both sample composition and engagement.

In study 2, creating Hobson's Choice counterfactuals for What/How interrogatives required greater specificity and imposed a particular evaluative framing on the prompt content. Despite the mitigation measures described in §4.2.1, the heterogeneity of PRISM prompts precluded a fully algorithmic approach. Because the reformulations were produced by a single annotator, potential subjectivity may affect internal validity; this could be improved in future work by incorporating multiple annotators and agreement checks. Similarly, some counterfactuals may be less contextually plausible than the original PRISM prompts, reinforcing the importance of implementing this experimental design with observational data to evaluate the generalisability of findings.

5.2 Potential Applications and Future Research

Given the universality of question-asking and the expansive scope of information represented within LLMs, the social scientific implications—and thus the potential applications—of Belnap and Steel's (1976) theoretical framework are extensive. To illustrate, I discuss two potential applications drawn from distinct domains.

First, such cycles may have important implications in medicine, where research increasingly explores their use to reduce the burden on healthcare practitioners (Gaber et al., 2025). From the clinician's perspective, D. Chen et al. (2025) describe LLM use in assisting with symptom documentation and retrieving information from large patient files. If certain clinicians are more likely to use closed-ended interrogatives when retrieving patient information, and this leads to shorter LLM responses with less nuance and reasoning but greater affinity, repeated exchanges of this kind could create a self-perpetuating cycle that progressively narrows the diagnostic dialogue. In some cases, such narrowing may reduce the likelihood of retrieving rare or unexpected symptoms. Investigating this risk and its potential to introduce new inequalities or exacerbate existing ones in diagnostic quality is an important area for future inquiry.

Similar types of human-LLM interaction cycles may emerge in patient-facing implementations of LLMs. For instance, Chen et al. (2025) also describe their use as tools to facilitate patient understanding. Examining this potential in future studies is important, as such cycles could be associated with patients under-reporting or over-interpreting symptoms. Taken together, these examples illustrate the range of dynamics that must be considered, underscoring the complexity involved in understanding the potential consequences of widespread LLM integration in healthcare.

Second, constructive dialogue on politically relevant topics among the population is considered a key component of healthy democracies, and its reported decline in the context of rising political polarisation has raised concern (Caluwaerts et al., 2023; Novoa et al., 2023). Recent findings indicate that LLMs are increasingly used as sources of information on politically relevant topics, including elections, geopolitical conflicts, or rightsbased issues such as abortions (Aoki, 2024; Zhu et al., 2025) Findings from the present study suggest a potential pathway that could contribute to reduced constructive dialogue: if the discussion of controversial and values-based topics is more often associated with closed-ended interrogatives (Study 1), and such interrogatives tend to elicit responses with comparatively less reasoning and nuance but greater affinity and compassion (Study 2), this may, in some contexts, limit the scope for constructive exchange. Future research should examine how such cycles of human–LLM interaction influence the conditions associated with constructive dialogue, including their potential role as a pathway for (mis)belief acquisition.

To conclude, I operationalised Belnap and Steel's (1976) taxonomy of interrogatives as a theoretical lens to investigate the potential consequences of the rapid, widespread adoption of LLMs. In doing so, I sought to address the gap in the literature concerning how everyday users may inadvertently engage in prompt engineering through individual differences in their routine LLM interactions, and the implications for response characteristics. Findings from two studies applying this operationalisation indicated potential cycles in human–LLM interactions, shaped by both user characteristics and LLM-specific properties. These emphasise the need for future research across different domains of application to understand their potential consequences and inform timely safeguards.

6. Appendices

Appendix A. Links Required for Reproducibility

Category	Description	Link
Code	GitHub repository	https://github.com/50280/MY498-ASDS
Datasets	PRISM dataset (Kirk et al., 2024)	https://huggingface.co/datasets/HannahRoseKirk/prism-alignment
	Dataset with classified data classified according to operationalised taxonomy of interrogatives, required for the operationalisation of the taxonomy of interrogatives and study 1.	https://huggingface.co/datasets/carowagner/operationalisation-and-study1
	Dataset with collected LLM responses and bridging attribute classifications for study 2.	https://huggingface.co/datasets/carowagner/study2
Fine-tuned BERT models	Classifier 1A	https://huggingface.co/carowagner/classify-questions-1A
	Classifier 1B	https://huggingface.co/carowagner/classify-questions-1B
	Classifier 2A	https://huggingface.co/carowagner/classify-questions-2A
	Classifier 2B	https://huggingface.co/carowagner/classify-questions-2B
	Classifier 2C	https://huggingface.co/carowagner/classify-questions-2C
	Classifier 3A	https://huggingface.co/carowagner/classify-questions-3A
	Classifier 4A	https://huggingface.co/carowagner/clasify-questions-4A
Tool to demonstrate the operationalisation of the taxonomy of interrogatives	Tool to better understand and test the limitations of the operationalised taxonomy of interrogatives.	https://huggingface.co/spaces/carowagner/questionthetaxonomy

Note. In case there are any issues with accessing the GitHub repository via the link, it can also be found by navigating to GitHub, searching the username *50280* and navigating to the public repository titled *MY498-ASDS*.

Appendix B. BERT Fine-tuning Data Annotator Instructions

Panel A. Final annotator guidelines used for BERT fine-tuning (with examples from the PRISM dataset).

Classifier label and key question	**Answer guidelines**
1A. Does this interrogative request an answer?	- Answer YES if it requires an answer and NO if it does not require an answer. Statements like hello or hi are not considered to expect an answer. - If a statement is imperative, e.g., 'legalise abortion' this answer NO. If a statement is declarative, answer NO. E.g., "I have to meet a director of a highschool for a substitute position. I feel a little anxious." - If the interrogative directly addresses the model, it is labelled YES (e.g., "Explain to me the pros and cons of punitive vs rehabilitative prison systems."). If an (imperative/declarative) interrogative does not directly address the model, it is labelled NO (e.g., "create a recipe using the following ingredients: black beans, ground beef, diced tomatoes, mushrooms, frozen onions and peppers, elbow pasta") [the reason underlying this decision is that if it directly addresses the model, it is considered as requesting an answer from the model.] - Edge case: If it does not have an active verb that makes it an imperative/declarative, but does not address the model directly, it is labelled as yes, because it is considered as implicitly requesting an answer from the model. E.g. "things to do in warrington" or "Take on corruption".
1B. If NO to 1A. Is this a declarative / imperative interrogative?	- Here, NA (not applicable) if YES to 1A. If it is a simple noun phrase such as 'technology and society', answer NO. If it is a declarative or imperative statement such as 'Guns are too easy to buy in some countries', answer YES.
2A. Is this an interrogative that expects a yes or no answer?	- Think: can this question be answered with yes / no? Answer YES if it can be answered with yes/no, and NO if it cannot be answered with yes/no. Answer NA if not relevant; e.g, does not request an answer and is not an affirmative statement. E.g., "I think Roe vs Wade should be reinstated." is NO. - E.g., Should college be tuition-free? Is YES.
2B. Does it explicitly present a series of options?	- Answer YES if the questioner explicitly gives as list from which the answerer can chose an answer. Answer NO if the question does not define an explicit list. - E.g., "Do you think animals go to heaven or hell?" This is YES because the answerer has to choose between "heaven" and "hell."
2C. How many options does it present?	- This question is about how the user defines the space of possible answers in the way they ask their question. It needs to be answered with either 0, 1, 2, or U – OR another integer number. - 0 is given for declarative/imperative statements because they do not directly incite an answer and thereby define an answer space of 0. E.g., "Guns are too easy to buy in some countries". - 'Why questions', questions that ask about the cause of something should be answered with 1. This is because by asking about the cause of something, they are assuming that cause to exist. E.g., "Why do criminal migrants keep living and making crime in our countries?" is 1, because the way the questioner formulated this question does not incite the answerer to say that migrants are not necessarily criminal. - 2 is given if the question can be answered with yes or no, and if the questioner explicitly describes two options in their question, e.g., "Is abortion a good or a bad thing?". - U is given if the answer space is undefined. This is assigned to descriptive, open-ended questions. E.g., "What are some steps we could take to combat global warming?". - If however, the question explicitly enumerates a series of options, it should be answered with an integer describing the number of available options. E.g., "i have three games in my library which should i play first: fallout 4, ace attorney, or the talos principle?"
3A. Do answers to this interrogative require some other fact/opinion already being true?	- Only questions that ask about the cause of some fact the questioner assumes to be true should be answered with YES here. "Tell me why Donald Trump will be the next president elect" because an answer requires that Trump will be the next president. Or "why are people so comfortable with eating animal corpses" is also YES. Otherwise NO.

Panel A continued. Final annotator guidelines used for BERT fine-tuning.

Classifier label and key question	Answer guidelines
4A. Does this interrogative ask for a description (y/n) / opinion (o)?	- Does it directly address the model and ask for the model's opinion? Answer O (for opinion). E.g., "Who do you predict would win the World Series in 2024?" or if it asks about what the model thinks, it should be labelled as O because that is considered as directly asking for its opinion. E.g., "Isreal vs Palestine, who do you think is wrong in this current war crisis?". Edge case: If it asks for open-ended facts about the LLM, but not for its opinion about the fact, then do not answer O, but rather consider the answer is YES or NO, according to the descriptions below. E.g., "Do you know that you are an AI?" (this would be a NO). Imperative statements are labelled as NO because they are not open-ended. E.g., "list each number to TEN in three languages" is NO. - Answer NO if it is not an open-ended descriptive question. E.g., NO to "are asians smarter?" because it is a question that can be answered with yes or no. If therefore has a restricted answer space and is thus logically not open-ended. Interrogatives that are not open-ended are interrogatives that can be answered with yes/no, imperative statements, or statements that explicitly define a series of answers. - Answer YES if the question is open-ended and asks for a description; typically, these are 'what' and 'how' questions e.g. "What is a good itinerary for a day in Melbourne".
Additional considerations	- When there are two questions in the same prompt, please classify this prompt solely with regards to the question that appears first in the interrogative. - However, when there is an imperative/declarative statement before the question, solely the question gets evaluated. e.g., "I feel it is important to keep the Welsh language alive. How would you promote this?".

Panel B. First version of annotator guidelines for the classifier labels that led to unsatisfactory inter-annotator agreement and were not used for fine-tuning.

Classifier label and key question	Answer guidelines
1A. Does this interrogative request an answer?	- Answer YES if it requires an answer and NO if it does not require an answer. Statements like hello or hi are not considered to expect an answer. But if a statement uses the imperative tense, e.g., 'legalise abortion' this is considered as requiring an answer.
2B. Does it explicitly present a series of options?	- Answer YES if the questioner explicitly gives as list from which the answerer can chose an answer. - Answer NO if the question does not define an explicit list. - E.g., "Do you think animals go to heaven or hell?" This is YES because the answerer has to choose between "heaven" and "hell."
4A. Does this interrogative ask for a description (y/n) / opinion (o)?	- Answer O (for opinion) if the question directly addresses the model and asks for the model's opinion. E.g., "Who do you predict would win the World Series in 2024?". - Answer YES if the question is open-ended and asks for a description; typically, these are 'what' and 'how' questions e.g. "What is a good itinerary for a day in Melbourne". Note that "Tell me what you know about Santiago" is labelled with YES because even though it directly addresses the model, it asks for a description rather than an opinion. - Answer NO if it does not ask for a description.

Appendix C. Bert fine-tuning parameters

Category	Parameter	Value
Architecture	Base model	google-bert/bert-base-uncased
Batch sizes	Train batch size	8
	Eval batch size	16
Learning rate	Initial learning rate	5E-05
Scheduler	Type	Linear
	Warmup ratio	0.1
Regularization	Weight decay	0.0
	Max gradient norm	1.0
	Dropout rate within each attention head	0.1
	Dropout rate in hidden layers	0.1
Training duration	Epochs	3
Evaluation	Eval strategy	Per epoch
	Save strategy	Per epoch
Loss function	Type	Cross-entropy
Randomness control	Seed	42
Optimiser	Type	AdamW (adamw_torch), β_1 = 0.9, β_2 = 0.999, ϵ = 1e-8

Note. All models were trained with identical hyperparameters; all other hyperparameters were default but can be accessed in the GitHub repository under . . . /02_code/00_setup_requirements/BERT_fine_tune_args.py. Initial experiments with different training durations showed that the models learned quickly, with little improvement after early epochs. To reduce the risk of overfitting, fine-tuning was fixed to three epochs. At the end of training, the checkpoint with the lowest validation loss (evaluated once per epoch) was reloaded for subsequent evaluation and prediction.

Appendix D. Out-of-sample performance per fine-tuned BERT classifier

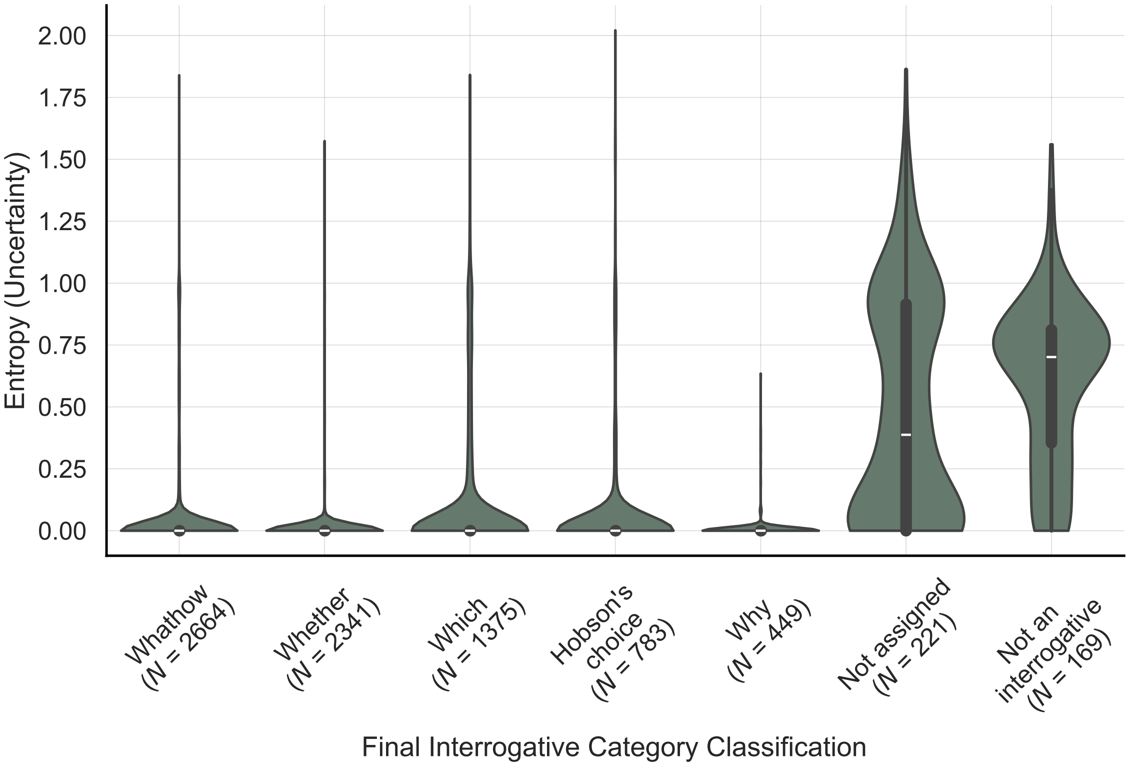
Classifier	Accuracy	Weighted F1	Support
1A	0.977	0.978	300
1B	0.977	0.977	300
2A	0.953	0.953	300
2B	0.937	0.932	300
2C	0.903	0.892	300
3A	0.980	0.979	300
4A	0.860	0.850	300

Note. The names in the format 1A-4B represent the different fine-tuned BERT classifiers. Details of how those were trained can be found in §2.2.5. Support in means number of observations in this category on the out-of-training-sample data. F1 is the harmonic mean of recall and precision.

Appendix E. Monte-Carlo dropout uncertainty estimates of BERTs.

Classifier	Label	Train Proportion	Mean	SD	Variance
1A	No	11.2	0.170	0.326	0.0024
1A	Yes	88.8	0.830	0.326	0.0024
1B	Not applicable	88.8	0.861	0.321	0.0011
1B	No	2.2	0.018	0.095	0.0001
1B	Yes	9.0	0.120	0.300	0.0010
2A	Not applicable	2.0	0.026	0.111	0.0002
2A	No	67.5	0.694	0.426	0.0010
2A	Yes	30.5	0.280	0.418	0.0008
2B	Not applicable	2.0	0.020	0.087	0.0003
2B	No	88.5	0.874	0.230	0.0011
2B	Yes	9.5	0.106	0.201	0.0007
2C	0	6.8	0.083	0.190	0.0005
2C	1	5.2	0.058	0.158	0.0002
2C	2	33.8	0.332	0.429	0.0006
2C	Not applicable	2.0	0.023	0.042	0.0001
2C	Undefined	51.7	0.503	0.467	0.0010
3A	Not applicable	2.0	0.021	0.121	0.0006
3A	No	93.0	0.920	0.258	0.0007
3A	Yes	5.0	0.059	0.221	0.0001
4A	Not applicable	2.0	0.010	0.031	0.0000
4A	No	41.8	0.456	0.457	0.0018
4A	Opinion	18.0	0.170	0.350	0.0010
4A	Yes	38.2	0.357	0.435	0.0019

Note. Each row shows uncertainty statistics from Monte Carlo dropout predictions for a BERT classifier-label pair. The “Train Proportion” refers to the percentage of training samples for that label. “Mean” and “SD” are the average and standard deviation of prediction confidence across utterances. “Variance” is the mean of the squared SDs across utterances, capturing overall uncertainty more accurately.



Note. These violin plots represent the distribution of uncertainty (entropy) across final interrogative category assignment for the 8002 PRISM user prompts. Low entropy means that the operationalisation of the taxonomy of interrogatives tended to assign the same category across the 100 different simulations of the BERT classifiers that were implemented using Monte-Carlo Dropout. The width at each vertical level reflects the relative density of responses at that level. Widths are normalized within each category, not across categories. Surface area does not reflect sample size. Sample sizes for each category are specified in the X-axis, under N, standing for number.

Appendix F. Distribution of Monte-Carlo Dropout Entropy by Final Interrogative Category

Appendix G. Pre-registration links and pre-registration accountability statement.

Panel A. Accountability statement for Study 1.

Pre-registered Element	Implemented	Added to report	Justification & Notes
Characteristic, contingency tables	Yes	Yes	Implemented as intended (see notes on DSL implementation in Appendix H for all analyses reported in this section).
Association, over-representation factors	Yes	Yes	NA
Conditional association, logistic regressions	Yes	Yes	Implemented as intended, average marginal effects were added to the report and not pre-registered, to improve the interpretability of the results.
Clustering to derive interrogative type profiles	Yes	No	The results for these analyses are available in the Github repository, under …/02_code/08_descriptive_analyses.ipynb, at the end of the script. Most of the participants were not assigned to a cluster (N = 1103). This is likely because of the high dimensionality of the data, and with five dimensions a different approach compared to the pre-registered one may have been more appropriate. Different approaches were not tested, and this was not added to the final report because of the brevity of the final report, and the richness of the results already included.
DSL implementation	Yes	Yes	Some difficulties were encountered when implementing the DSL, which means that it was not implemented in the same way across the analyses. See Appendix H for further descriptions and justifications.
Analyses in report that were not pre-registered	None.		
Pre-registration reference:	LastName, X. (2025, June 22). Descriptive Analyses: Profiling the Questioning Behaviours of LM Users. https://doi.org/10.17605/OSF.IO/CR58Z		

Panel B. Accountability statement for Study 2.

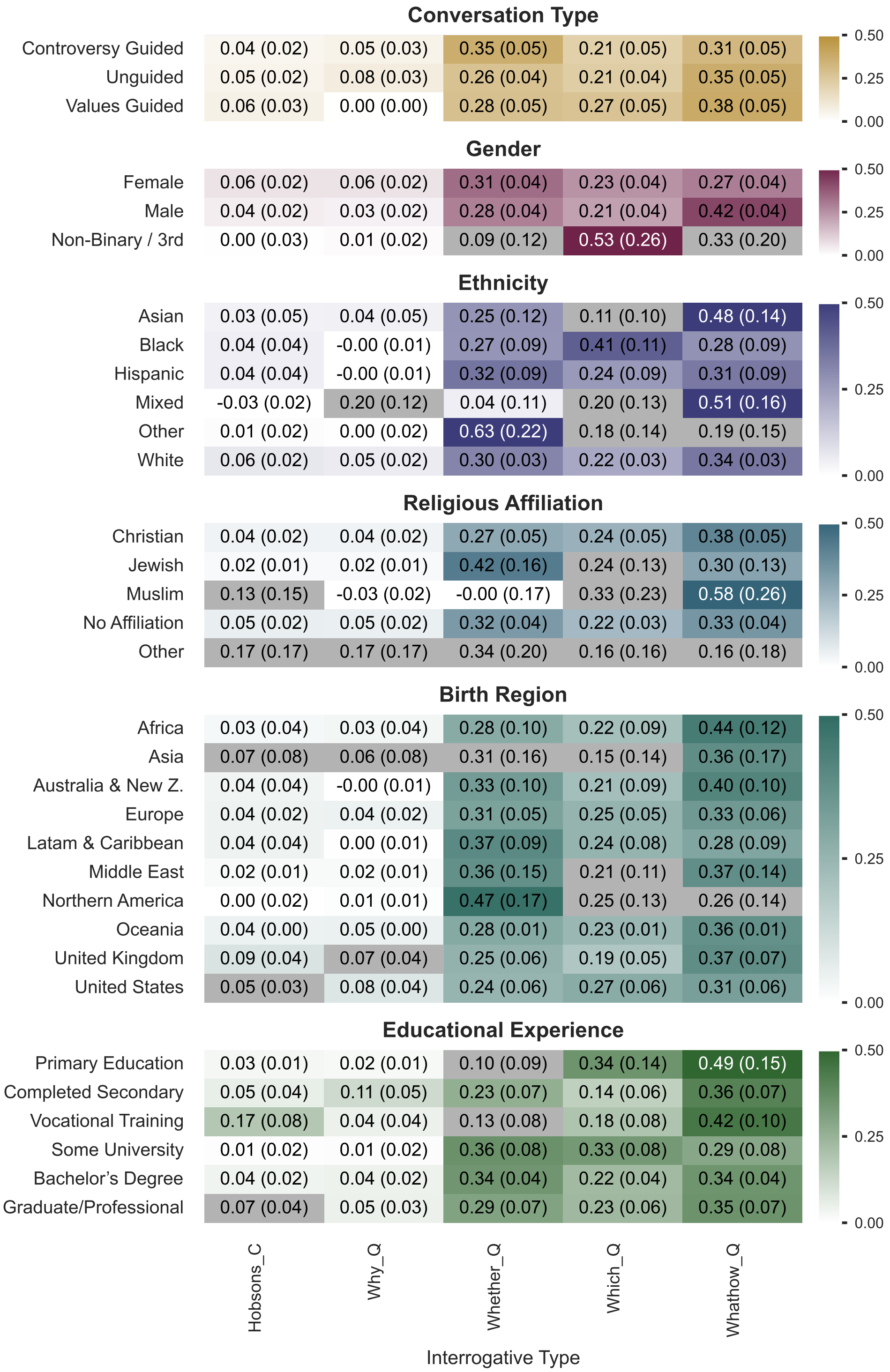
Pre-registered Element	Implemented	Added to report	Justification & Notes
LLMs for data collection	Yes	Yes	Pre-registered procedures for LLM selection were followed although technical limitations with accessing some LLMs through the HuggingFace API meant that I accessed some of them through the TogetherAI API. Some of the specific LLM versions were not available through this provider but equivalent ones were chosen according to pre-registered criteria.
Experimental manipulation	Yes	Yes	NA
Response length	Yes	Yes	NA
Pre-registered specification: AttributeScore Questiontype * LLM + (1 — QuestionID)	Yes	No	I included the following specification in the main report: AttributeScore Questiontype * LLM because it is the simplest possible specification. The exact pre-registered specification was run and found substantially equivalent results to the reported specification. In addition, I ran the specification in the main report without clustered standard errors, to further test its robustness. Results for all specifications are included in the Github repository under . . . /01_data/10_experimental_results).

Analyses in report that were not pre-registered	The structures qualitative observations that were included in the analyses to get an impression of why the responses differ were not pre-registered and added after finding unexpectedly high effect sizes for response lengths, to develop an initial understanding as to why the response lengths between interrogative types differ to this extent.
Additional Notes	Effect size, statistical power, and inference criteria were implemented and reported as pre-registered.
Pre-registration reference:	LastName, X. (2025, June 29). A Taxonomy of Interrogatives and Their Role in Human-Language Model Interaction. https://doi.org/10.17605/OSF.IO/XKP6B

Appendix H. Extended methods for design-based supervised learning implementation.

Analyses	Implemented Approach	Description
Characteristic; contingency tables	Using the DSL R package by Egami et al. (2025)	This uses the DSL package, regressing each dummy-encoded interrogative category i on each demographic feature d using linear regression without an intercept. This approach is mathematically equivalent to computing subgroup means. Although it is less interpretable, this was necessary because the DSL estimator cannot simply be applied to aggregated count data. As this results in category proportions by demographic group, these are less interpretable than counts in terms of the estimands of interest; e.g., P(interrogative type — demographic characteristic) requires bayes rule. Therefore, the main report includes traditional contingency tables with counts, but the cells of plots are greyed out when the analysis with the DSL package indicated relatively high uncertainty according to the criteria described.
Association; over-representation factors	Manual implementation of the DSL	To manually implement the DSL, I created design-adjusted outcomes for each dummy-encoded category, and used these adjusted outcomes to compute over-representation factors (using bootstraps to get uncertainty estimates). This formula stems from Egami et al. (2023): For observations that are expert-labelled: $\tilde{Y} = \hat{Y} - \frac{\hat{Y} - Y}{\pi},$ where $\hat{Y}$ = the label from the LLM, Y = expert label π = proportion of observations that are expert-labelled. For observations that are not expert-labelled: $\tilde{Y} = \hat{Y},$ where $\hat{Y}$ = LLM label This leads to conservative uncertainty estimates, as shown in the magnitude of the standard errors for some of the standard errors in Appendix J. A supervised learning model (g) to improve the LLM predictions was not implemented for simplicity and because this is not strictly necessary, particularly if the LLM labels already have high accuracy (as they do in this case).
Conditional association; logistic regressions	Not implemented	The current version of the DSL R package (0.1.0; 12.08.2025) does not support clustered standard errors when specifying the 'logit' model. To address this limitation, I attempted to implement a linear approximation using the felm model. Please see …/02_code/07_descriptive_log_regs.Rmd for a detailed description of the errors encountered, and the steps attempted to address them. I then considered using the design-adjusted columns from my previous manual implementation, and while these work for descriptive analyses, in the case of the current logistic regressions, the formula by definition introduces numbers that are not 0 or 1, which makes the logit model break. I was not able to find an indication of how to address this in the relevant DSL papers. For these reasons, I am not implementing the DSL for this step of my descriptive analyses.

Note. DSL = design-based supervised learning. All methods stem from Egami et al. (2023) and Egami et al. (2024). The DSL R package was developed by Egami et al., (2025).



Conversation Type

	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Controversy Guided	0.04 (0.02)	0.05 (0.03)	0.35 (0.05)	0.21 (0.05)	0.31 (0.05)
Unguided	0.05 (0.02)	0.08 (0.03)	0.26 (0.04)	0.21 (0.04)	0.35 (0.05)
Values Guided	0.06 (0.03)	0.00 (0.00)	0.28 (0.05)	0.27 (0.05)	0.38 (0.05)

Gender

	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Female	0.06 (0.02)	0.06 (0.02)	0.31 (0.04)	0.23 (0.04)	0.27 (0.04)
Male	0.04 (0.02)	0.03 (0.02)	0.28 (0.04)	0.21 (0.04)	0.42 (0.04)
Non-Binary / 3rd	0.00 (0.03)	0.01 (0.02)	0.09 (0.12)	0.53 (0.26)	0.33 (0.20)

Ethnicity

	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Asian	0.03 (0.05)	0.04 (0.05)	0.25 (0.12)	0.11 (0.10)	0.48 (0.14)
Black	0.04 (0.04)	-0.00 (0.01)	0.27 (0.09)	0.41 (0.11)	0.28 (0.09)
Hispanic	0.04 (0.04)	-0.00 (0.01)	0.32 (0.09)	0.24 (0.09)	0.31 (0.09)
Mixed	-0.03 (0.02)	0.20 (0.12)	0.04 (0.11)	0.20 (0.13)	0.51 (0.16)
Other	0.01 (0.02)	0.00 (0.02)	0.63 (0.22)	0.18 (0.14)	0.19 (0.15)
White	0.06 (0.02)	0.05 (0.02)	0.30 (0.03)	0.22 (0.03)	0.34 (0.03)

Religious Affiliation

	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Christian	0.04 (0.02)	0.04 (0.02)	0.27 (0.05)	0.24 (0.05)	0.38 (0.05)
Jewish	0.02 (0.01)	0.02 (0.01)	0.42 (0.16)	0.24 (0.13)	0.30 (0.13)
Muslim	0.13 (0.15)	-0.03 (0.02)	-0.00 (0.17)	0.33 (0.23)	0.58 (0.26)
No Affiliation	0.05 (0.02)	0.05 (0.02)	0.32 (0.04)	0.22 (0.03)	0.33 (0.04)
Other	0.17 (0.17)	0.17 (0.17)	0.34 (0.20)	0.16 (0.16)	0.16 (0.18)

Birth Region

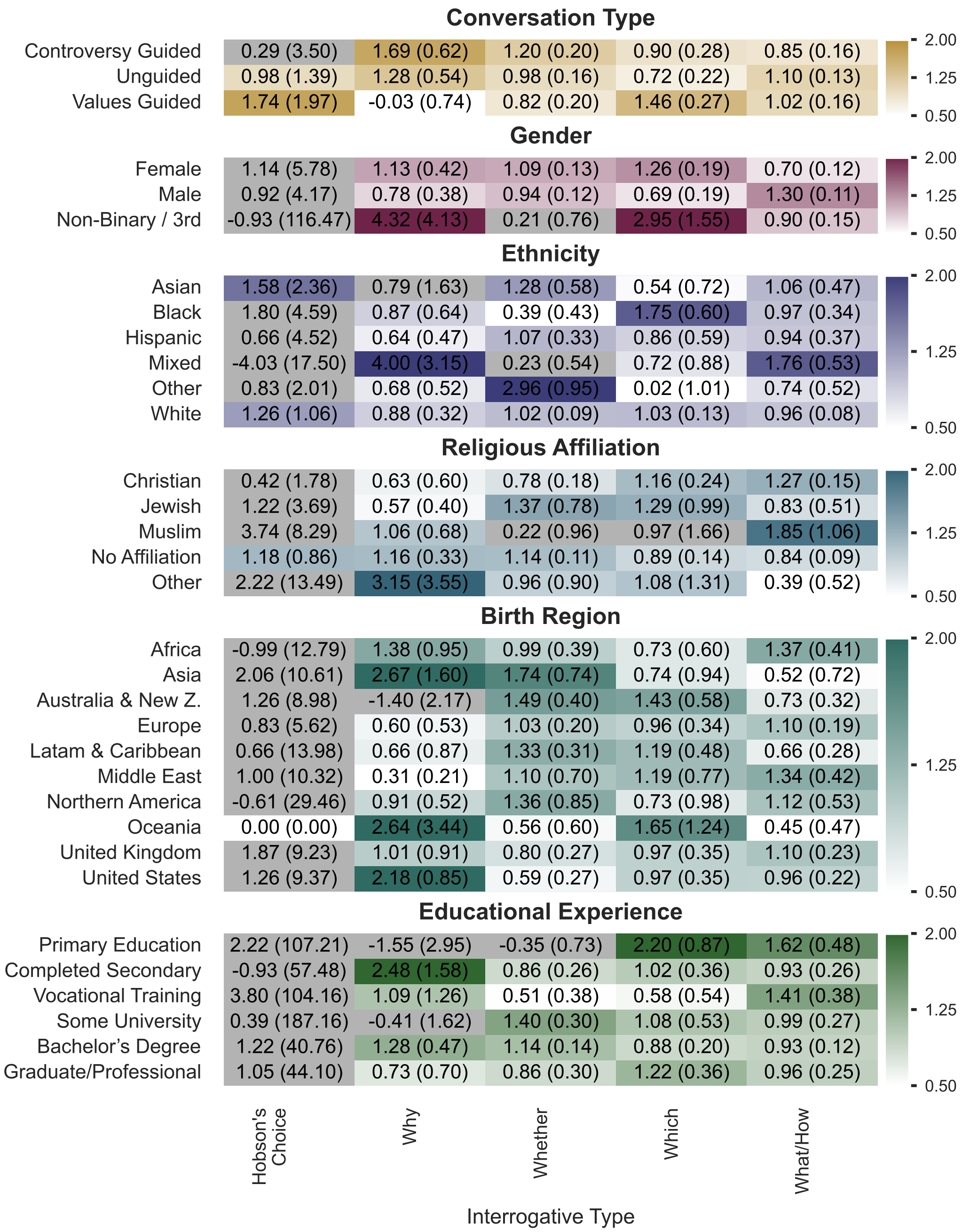
	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Africa	0.03 (0.04)	0.03 (0.04)	0.28 (0.10)	0.22 (0.09)	0.44 (0.12)
Asia	0.07 (0.08)	0.06 (0.08)	0.31 (0.16)	0.15 (0.14)	0.36 (0.17)
Australia & New Z.	0.04 (0.04)	-0.00 (0.01)	0.33 (0.10)	0.21 (0.09)	0.40 (0.10)
Europe	0.04 (0.02)	0.04 (0.02)	0.31 (0.05)	0.25 (0.05)	0.33 (0.06)
Latam & Caribbean	0.04 (0.04)	0.00 (0.01)	0.37 (0.09)	0.24 (0.08)	0.28 (0.09)
Middle East	0.02 (0.01)	0.02 (0.01)	0.36 (0.15)	0.21 (0.11)	0.37 (0.14)
Northern America	0.00 (0.02)	0.01 (0.01)	0.47 (0.17)	0.25 (0.13)	0.26 (0.14)
Oceania	0.04 (0.00)	0.05 (0.00)	0.28 (0.01)	0.23 (0.01)	0.36 (0.01)
United Kingdom	0.09 (0.04)	0.07 (0.04)	0.25 (0.06)	0.19 (0.05)	0.37 (0.07)
United States	0.05 (0.03)	0.08 (0.04)	0.24 (0.06)	0.27 (0.06)	0.31 (0.06)

Educational Experience

	Hobsons_C	Why_Q	Whether_Q	Which_Q	Whathow_Q
Primary Education	0.03 (0.01)	0.02 (0.01)	0.10 (0.09)	0.34 (0.14)	0.49 (0.15)
Completed Secondary	0.05 (0.04)	0.11 (0.05)	0.23 (0.07)	0.14 (0.06)	0.36 (0.07)
Vocational Training	0.17 (0.08)	0.04 (0.04)	0.13 (0.08)	0.18 (0.08)	0.42 (0.10)
Some University	0.01 (0.02)	0.01 (0.02)	0.36 (0.08)	0.33 (0.08)	0.29 (0.08)
Bachelor’s Degree	0.04 (0.02)	0.04 (0.02)	0.34 (0.04)	0.22 (0.04)	0.34 (0.04)
Graduate/Professional	0.07 (0.04)	0.05 (0.03)	0.29 (0.07)	0.23 (0.06)	0.35 (0.07)

Appendix I. DSL-Adjusted Category Proportions by Demographic Group

Note. This shows an adjusted estimate of the mean (standard error) predicted label for each subgroup, correcting for sampling bias with the design-based supervised learning estimator (Egami et al., 2024). Cells are greyed out if the standard error (SE) > 50% of the estimate (proportion), and the estimate > 0.05 (the latter prevents visually penalising near-zero estimates). Cell colouring (when not greyed out) depends solely on the proportion estimate and does not encompass SEs. Because these represent proportions and not counts, rows with totals were removed because their interpretation into estimands of interest (e.g., P(interrogative type | demographic characteristic)) is



Conversation Type

	Hobson's Choice	Why	Whether	Which	What/How
Controversy Guided	0.29 (3.50)	1.69 (0.62)	1.20 (0.20)	0.90 (0.28)	0.85 (0.16)
Unguided	0.98 (1.39)	1.28 (0.54)	0.98 (0.16)	0.72 (0.22)	1.10 (0.13)
Values Guided	1.74 (1.97)	-0.03 (0.74)	0.82 (0.20)	1.46 (0.27)	1.02 (0.16)

Gender

	Hobson's Choice	Why	Whether	Which	What/How
Female	1.14 (5.78)	1.13 (0.42)	1.09 (0.13)	1.26 (0.19)	0.70 (0.12)
Male	0.92 (4.17)	0.78 (0.38)	0.94 (0.12)	0.69 (0.19)	1.30 (0.11)
Non-Binary / 3rd	-0.93 (116.47)	4.32 (4.13)	0.21 (0.76)	2.95 (1.55)	0.90 (0.15)

Ethnicity

	Hobson's Choice	Why	Whether	Which	What/How
Asian	1.58 (2.36)	0.79 (1.63)	1.28 (0.58)	0.54 (0.72)	1.06 (0.47)
Black	1.80 (4.59)	0.87 (0.64)	0.39 (0.43)	1.75 (0.60)	0.97 (0.34)
Hispanic	0.66 (4.52)	0.64 (0.47)	1.07 (0.33)	0.86 (0.59)	0.94 (0.37)
Mixed	-4.03 (17.50)	4.00 (3.15)	0.23 (0.54)	0.72 (0.88)	1.76 (0.53)
Other	0.83 (2.01)	0.68 (0.52)	2.96 (0.95)	0.02 (1.01)	0.74 (0.52)
White	1.26 (1.06)	0.88 (0.32)	1.02 (0.09)	1.03 (0.13)	0.96 (0.08)

Religious Affiliation

	Hobson's Choice	Why	Whether	Which	What/How
Christian	0.42 (1.78)	0.63 (0.60)	0.78 (0.18)	1.16 (0.24)	1.27 (0.15)
Jewish	1.22 (3.69)	0.57 (0.40)	1.37 (0.78)	1.29 (0.99)	0.83 (0.51)
Muslim	3.74 (8.29)	1.06 (0.68)	0.22 (0.96)	0.97 (1.66)	1.85 (1.06)
No Affiliation	1.18 (0.86)	1.16 (0.33)	1.14 (0.11)	0.89 (0.14)	0.84 (0.09)
Other	2.22 (13.49)	3.15 (3.55)	0.96 (0.90)	1.08 (1.31)	0.39 (0.52)

Birth Region

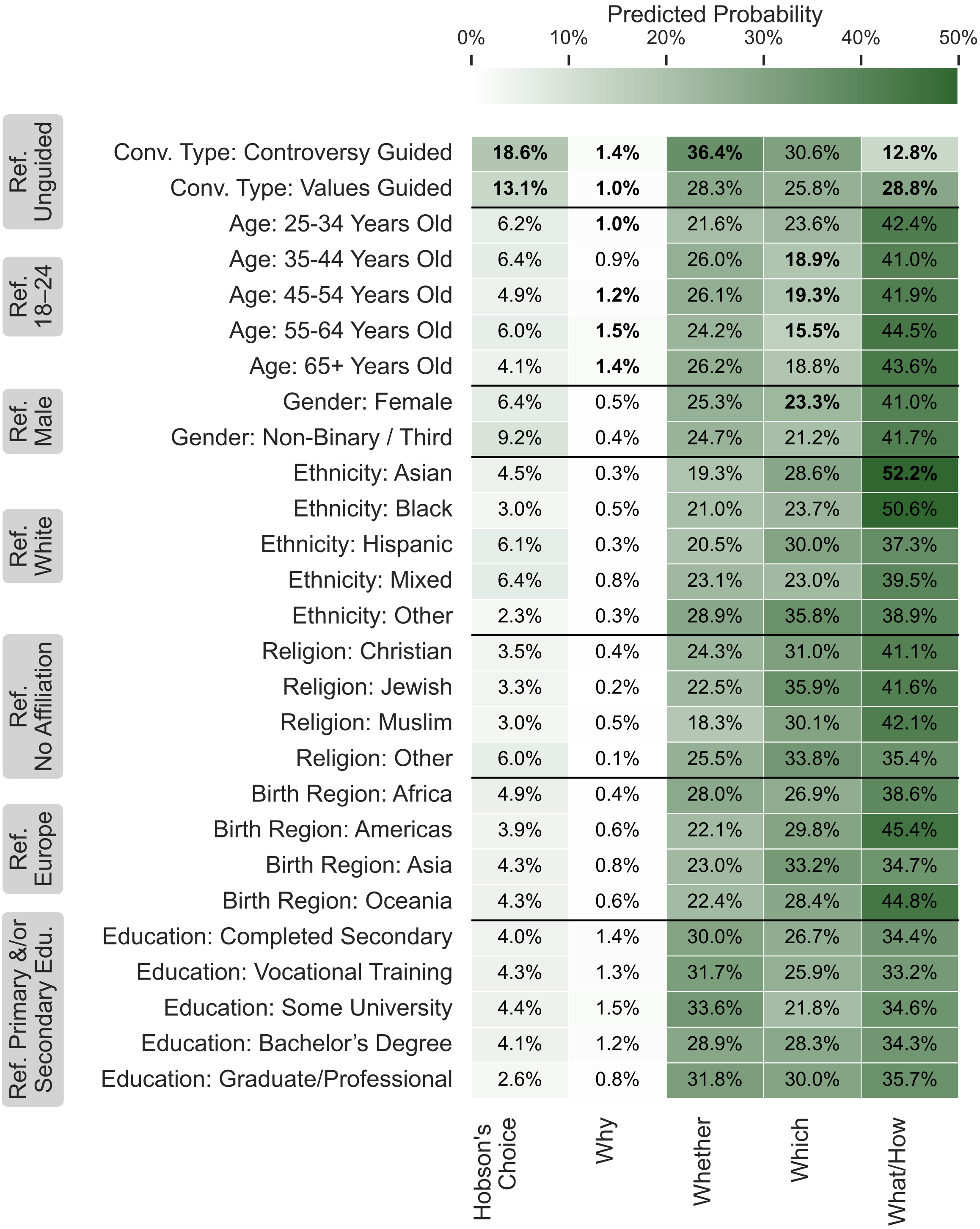
	Hobson's Choice	Why	Whether	Which	What/How
Africa	-0.99 (12.79)	1.38 (0.95)	0.99 (0.39)	0.73 (0.60)	1.37 (0.41)
Asia	2.06 (10.61)	2.67 (1.60)	1.74 (0.74)	0.74 (0.94)	0.52 (0.72)
Australia & New Z.	1.26 (8.98)	-1.40 (2.17)	1.49 (0.40)	1.43 (0.58)	0.73 (0.32)
Europe	0.83 (5.62)	0.60 (0.53)	1.03 (0.20)	0.96 (0.34)	1.10 (0.19)
Latam & Caribbean	0.66 (13.98)	0.66 (0.87)	1.33 (0.31)	1.19 (0.48)	0.66 (0.28)
Middle East	1.00 (10.32)	0.31 (0.21)	1.10 (0.70)	1.19 (0.77)	1.34 (0.42)
Northern America	-0.61 (29.46)	0.91 (0.52)	1.36 (0.85)	0.73 (0.98)	1.12 (0.53)
Oceania	0.00 (0.00)	2.64 (3.44)	0.56 (0.60)	1.65 (1.24)	0.45 (0.47)
United Kingdom	1.87 (9.23)	1.01 (0.91)	0.80 (0.27)	0.97 (0.35)	1.10 (0.23)
United States	1.26 (9.37)	2.18 (0.85)	0.59 (0.27)	0.97 (0.35)	0.96 (0.22)

Educational Experience

	Hobson's Choice	Why	Whether	Which	What/How
Primary Education	2.22 (107.21)	-1.55 (2.95)	-0.35 (0.73)	2.20 (0.87)	1.62 (0.48)
Completed Secondary	-0.93 (57.48)	2.48 (1.58)	0.86 (0.26)	1.02 (0.36)	0.93 (0.26)
Vocational Training	3.80 (104.16)	1.09 (1.26)	0.51 (0.38)	0.58 (0.54)	1.41 (0.38)
Some University	0.39 (187.16)	-0.41 (1.62)	1.40 (0.30)	1.08 (0.53)	0.99 (0.27)
Bachelor's Degree	1.22 (40.76)	1.28 (0.47)	1.14 (0.14)	0.88 (0.20)	0.93 (0.12)
Graduate/Professional	1.05 (44.10)	0.73 (0.70)	0.86 (0.30)	1.22 (0.36)	0.96 (0.25)

Appendix J. Over-Representation Factors by Demographic Group

Note. This figure displays over-representation factors (ORFs) for each demographic subgroup across interrogative types, where each cell compares a group's actual participation to what would be expected given that subgroup’s size in the full dataset. An ORF of 1.0 indicates proportional representation, values above 1.0 reflect over-representation, and values below 1.0 indicate under-representation. Each cell represents an independent calculation and is not constrained to sum to any fixed total. Cells are greyed out when the associated standard error exceeds 150\% of the estimate and the estimate is greater than 0.05, signaling high relative uncertainty. The color mapping emphasizes within-column comparisons (i.e., across groups for each interrogative type). Some regions (e.g., UK, US) are separated from their broader categories (Europe, North America) to better reflect sample density. Note that this figure reflects prompt-level patterns, not participant-level representation; each participant contributed multiple prompts.



		Hobson's Choice	Why	Whether	Which	What/How
Ref. Unguided	Conv. Type: Controversy Guided	**18.6%**	**1.4%**	**36.4%**	30.6%	**12.8%**
	Conv. Type: Values Guided	**13.1%**	**1.0%**	28.3%	25.8%	**28.8%**
Ref. 18–24	Age: 25-34 Years Old	6.2%	**1.0%**	21.6%	23.6%	42.4%
	Age: 35-44 Years Old	6.4%	0.9%	26.0%	**18.9%**	41.0%
	Age: 45-54 Years Old	4.9%	**1.2%**	26.1%	**19.3%**	41.9%
	Age: 55-64 Years Old	6.0%	**1.5%**	24.2%	**15.5%**	44.5%
	Age: 65+ Years Old	4.1%	**1.4%**	26.2%	18.8%	43.6%
Ref. Male	Gender: Female	6.4%	0.5%	25.3%	**23.3%**	41.0%
	Gender: Non-Binary / Third	9.2%	0.4%	24.7%	21.2%	41.7%
Ref. White	Ethnicity: Asian	4.5%	0.3%	19.3%	28.6%	**52.2%**
	Ethnicity: Black	3.0%	0.5%	21.0%	23.7%	50.6%
	Ethnicity: Hispanic	6.1%	0.3%	20.5%	30.0%	37.3%
	Ethnicity: Mixed	6.4%	0.8%	23.1%	23.0%	39.5%
	Ethnicity: Other	2.3%	0.3%	28.9%	35.8%	38.9%
Ref. No Affiliation	Religion: Christian	3.5%	0.4%	24.3%	31.0%	41.1%
	Religion: Jewish	3.3%	0.2%	22.5%	35.9%	41.6%
	Religion: Muslim	3.0%	0.5%	18.3%	30.1%	42.1%
	Religion: Other	6.0%	0.1%	25.5%	33.8%	35.4%
Ref. Europe	Birth Region: Africa	4.9%	0.4%	28.0%	26.9%	38.6%
	Birth Region: Americas	3.9%	0.6%	22.1%	29.8%	45.4%
	Birth Region: Asia	4.3%	0.8%	23.0%	33.2%	34.7%
	Birth Region: Oceania	4.3%	0.6%	22.4%	28.4%	44.8%
Ref. Primary &/or Secondary Edu.	Education: Completed Secondary	4.0%	1.4%	30.0%	26.7%	34.4%
	Education: Vocational Training	4.3%	1.3%	31.7%	25.9%	33.2%
	Education: Some University	4.4%	1.5%	33.6%	21.8%	34.6%
	Education: Bachelor’s Degree	4.1%	1.2%	28.9%	28.3%	34.3%
	Education: Graduate/Professional	2.6%	0.8%	31.8%	30.0%	35.7%

Appendix K. Predicted Probabilities by Demographic Group and Interrogative Type

Note. Cells show predicted probabilities for each interrogative type by demographic subgroup. Text annotations display point estimates; entries in bold indicate terms flagged as statistically significant in the accompanying model output. The horizontal colorbar reports predicted probability (0–50%). Group blocks are separated by solid lines with reference categories labelled in the margin.

Appendix L. Study 2: Prompt templates for experimental manipulation.

Template Type	Template ID	Template (where X is element to be replaced)	PRISM User Prompt [PRISM prompt ID] from Kirk et al. (2024)	Counterfactual prompt
Hobson's Choice	HC1	Tell me that X OR Tell me the advantages / disadvantages of X.	"What happens if we achieve AGI?" [ut7275]	"Tell me the benefit of achieving AGI."
	HC2	Transform sentence into an imperative statement.	"What is Green house effects?" [ut54978]	"Greenhouse effects are bad for the environment."
	HC3	I believe X.	"how can i have zero anxiety" [ut48162]	"I believe that I can have zero anxiety by meditating everyday."
	HC4	I think X.	"What does cica means in Cicaplast?" [ut59514]	"I think that the cica in cicaplast refers to cicatrisation."
	HC5	I hate/love that X.	"Who is a famous pop star in the USA?" [ut51833]	"I hate that Taylor Swift is a famous pop star in the USA."
What/How	WH1	What are different perspectives on X?	"Australia is too strict on asylumn seekers" [ut50173]	"What are differing perspectives on australians policy for asylum seekers?"
	WH2	How are different people thinking about X?	"I often hear that our food is poisoned through the big companies. As we all buy it from the supermarket, we dont have a choice and have to eat it" [ut14038]	"How are different people thinking about food being posionned throught the big companies?"
	WH3	What are different viewpoints on X?	"Christianity ALONE is true and other beliefs are FALSE" [ut9920]	"What are different viewpoints on Christianity ALONE being true and other beliefs being FALSE?"
	WH4	Can you describe the advantages and disadvantages of X?	"I think it's very worrying that far-right is winning in so many countries in Europe recently." [ut14839]	"Can you describe the advantages and disadvantages of the far-right winning in so many countries in europe recently?"
	WH5	What factors should I consider to understand X?	"I hate that I have to pay so many taxes." [ut53441]	"What factors should I consider to understand the amount of taxes that I have to pay?"

Note. These statements do not represent the views of the author.

Appendix M. Study 2: API providers, developers, and LLM versions used.

API Provider	LLM Developer	LLM Version
Together AI API	Meta	Llama-3.3-70B-Instruct-Turbo
Together AI API	Meta	Llama-4-Scout-17B-16E-Instruct
Together AI API	Meta	Meta-Llama-3-8B-Instruct-Lite
Mistral API	Mistral AI	mistral-small-2506
Mistral API	Mistral AI	mistral-medium-2505
Mistral API	Mistral AI	magistral-medium-2506
Together AI API	Google	gemma-3n-E4B-it
Together AI API	Google	gemma-2-27b-it
Google API	Google	gemini-2.5-flash
Google API	Google	gemini-1.5-pro
Anthropic API	Anthropic	claude-3-opus-20240229
Anthropic API	Anthropic	claude-3-haiku-20240307
Anthropic API	Anthropic	claude-sonnet-4-20250514
OpenAI API	OpenAI	gpt-4.1-nano
OpenAI API	OpenAI	04-mini
OpenAI API	OpenAI	gpt-4.1
OpenAI API	OpenAI	gpt-4.1-mini
Together AI API	Qwen	Qwen2.5-72B-Instruct-Turbo
Together AI API	Qwen	QwQ-32B
Together AI API	DeepSeek	DeepSeek-R1
Together AI API	DeepSeek	DeepSeek-R1-0528-tput

Appendix N. Study 2: LLM response length differences to experimental manipulation.

LLM Provider	Model Name	What/How Mean (SD)	Hobson's C. Mean (SD)	*t*-value	*p*-value	Cohen's *d*
Qwen	QwQ 32B	874.4 (759.7)	585.6 (298.0)	7.87	$<$ .001	0.38
Qwen	Qwen2.5 72B Instruct Turbo	591.8 (279.2)	371.6 (248.6)	18.00	$<$ .001	0.87
Anthropic	claude 3 haiku 20240307	248.1 (101.0)	187.8 (99.1)	12.33	$<$ .001	0.60
Anthropic	claude 3 opus 20240229	260.5 (81.8)	214.4 (95.6)	10.69	$<$ .001	0.52
Anthropic	claude sonnet 4 20250514	266.8 (74.1)	194.3 (77.4)	19.05	$<$ .001	0.92
Deepseek-Ai	DeepSeek R1	572.5 (134.1)	587.3 (161.3)	-1.86	0.063	-0.09
Deepseek-Ai	DeepSeek R1 0528 tput	941.5 (400.0)	737.3 (353.9)	8.75	$<$ .001	0.42
Google	gemini 1.5 pro	526.5 (216.0)	308.1 (233.4)	20.10	$<$ .001	0.97
Google	gemini 2.5 flash	1110.2 (521.6)	675.5 (507.6)	17.82	$<$ .001	0.87
Google	gemma 2 27b it	402.2 (162.7)	280.0 (161.6)	16.22	$<$ .001	0.79
Google	gemma 3n E4B it	1224.9 (527.2)	698.7 (498.0)	21.84	$<$ .001	1.06
Meta-Llama	Llama 3.3 70B Instruct Turbo	568.9 (178.2)	398.0 (192.2)	19.84	$<$ .001	0.96
Meta-Llama	Llama 4 Scout 17B 16E Instruct	552.0 (192.7)	377.3 (204.6)	17.53	$<$ .001	0.85
Meta-Llama	Meta Llama 3 8B Instruct Lite	499.5 (167.7)	355.6 (192.9)	15.30	$<$ .001	0.74
Mistral	magistral medium 2506	464.0 (414.1)	541.2 (440.1)	-2.38	0.018	-0.12
Mistral	mistral medium latest	609.2 (210.0)	432.4 (214.8)	17.26	$<$ .001	0.84
Mistral	mistral small	325.2 (137.4)	244.0 (121.6)	12.06	$<$ .001	0.58
Openai	gpt 4.1	500.8 (216.6)	272.3 (185.3)	23.21	$<$ .001	1.12
Openai	gpt 4.1 mini	374.9 (207.9)	172.4 (158.3)	20.86	$<$ .001	1.01
Openai	gpt 4.1 nano	327.1 (169.1)	153.7 (138.1)	21.77	$<$ .001	1.05
Openai	o4 mini	680.0 (282.3)	515.1 (261.7)	12.90	$<$ .001	0.62

Note. Hobson's C (Hobson's Choice) is the most closed-ended type of interrogative, and What/How is the most open-ended form of interrogative according to the taxonomy of interrogatives by Belnap & Steel (1976). All LLM responses were tokenized using the same tokenizer (via the `tiktoken` library), making lengths comparable across models. *t*-values are from paired t-tests; Cohen's *d* quantifies the difference in response length between What/How and Hobson's prompts, with positive values indicating longer responses to What/How.

Appendix O. Average treatment effects and Levene's test results across LLMs.

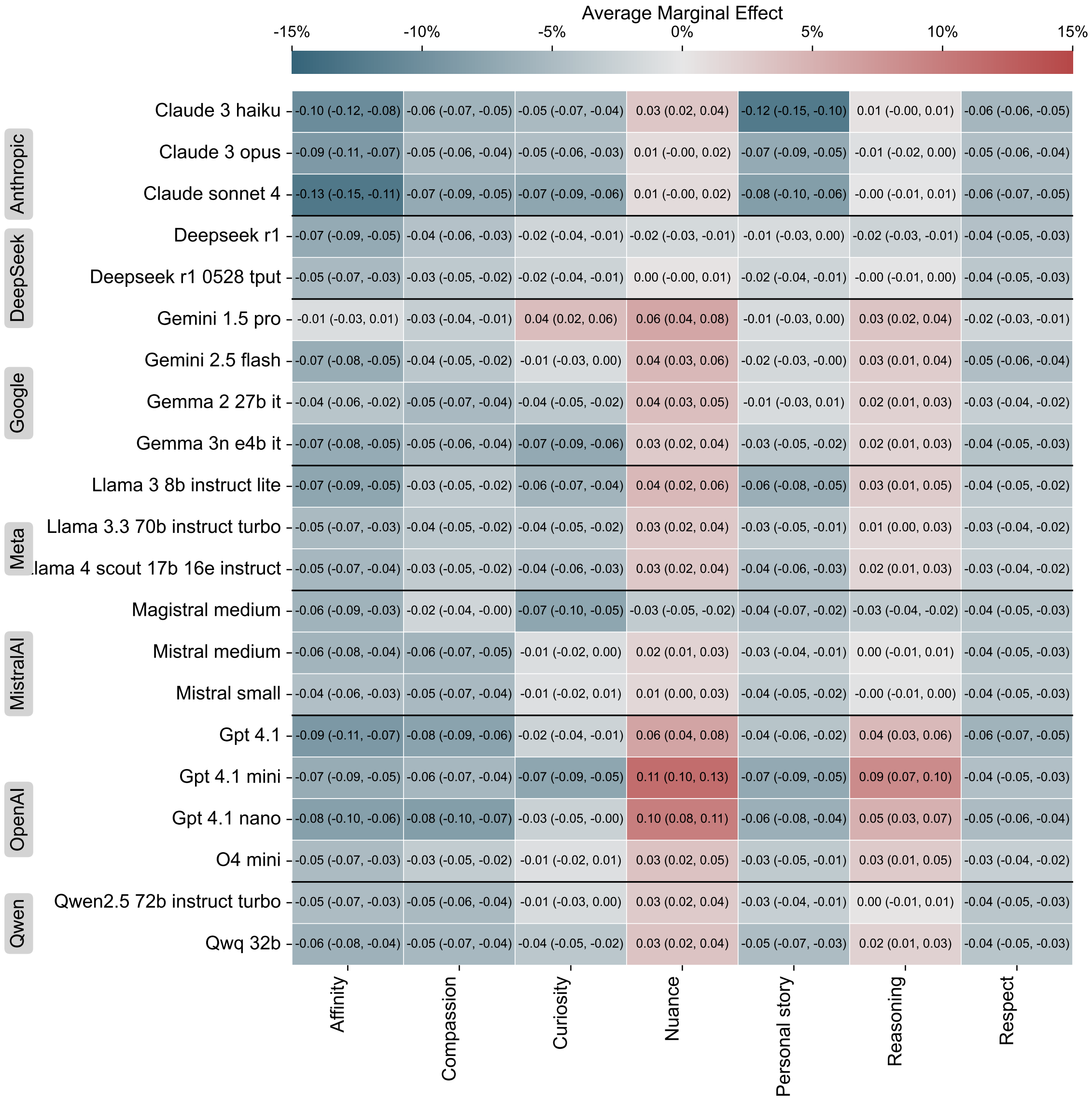
Model	Affinity		Compassion		Curiosity		Nuance	
	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)
Full Sample	-0.07 (0.00)	76.79 ($<$ .001)	-0.05 (0.00)	34.95 ($<$ .001)	-0.03 (0.00)	0.10 (0.756)	0.03 (0.00)	110.83 ($<$ .001)
Claude-3-Haiku-20240307	-0.10 (0.02)	13.07 ($<$ .001)	-0.06 (0.02)	4.45 (0.035)	-0.05 (0.01)	6.17 (0.013)	0.03 (0.01)	8.03 (0.005)
Claude-3-Opus-20240229	-0.09 (0.02)	8.92 (0.003)	-0.05 (0.01)	4.44 (0.035)	-0.05 (0.01)	5.02 (0.025)	0.01 (0.01)	0.00 (0.957)
Claude-Sonnet-4-20250514	-0.13 (0.02)	6.78 (0.009)	-0.07 (0.02)	2.06 (0.152)	-0.07 (0.01)	2.12 (0.146)	0.01 (0.01)	1.85 (0.174)
Deepseek-R1	-0.07 (0.01)	5.03 (0.025)	-0.04 (0.02)	1.69 (0.194)	-0.02 (0.01)	4.07 (0.044)	-0.02 (0.01)	0.00 (0.948)
Deepseek-R1-0528-Tput	-0.05 (0.01)	7.59 (0.006)	-0.03 (0.02)	0.03 (0.853)	-0.02 (0.01)	3.00 (0.084)	0.00 (0.01)	0.02 (0.887)
Gemini-1.5-Pro	-0.01 (0.02)	0.39 (0.535)	-0.03 (0.02)	1.17 (0.279)	0.04 (0.01)	3.08 (0.080)	0.06 (0.01)	10.50 (0.001)
Gemini-2.5-Flash	-0.07 (0.01)	8.14 (0.004)	-0.04 (0.02)	2.58 (0.109)	-0.01 (0.01)	0.86 (0.354)	0.04 (0.01)	8.87 (0.003)
Gemma-2-27B-It	-0.04 (0.02)	1.45 (0.229)	-0.05 (0.01)	6.92 (0.009)	-0.04 (0.01)	0.13 (0.714)	0.04 (0.01)	5.60 (0.018)
Gemma-3N-E4B-It	-0.07 (0.01)	15.22 ($<$ .001)	-0.05 (0.01)	3.98 (0.046)	-0.07 (0.01)	0.44 (0.507)	0.03 (0.01)	3.26 (0.071)
Gpt-4.1	-0.09 (0.02)	3.00 (0.084)	-0.08 (0.02)	5.18 (0.023)	-0.02 (0.01)	0.10 (0.749)	0.06 (0.01)	11.53 ($<$ .001)
Gpt-4.1-Mini	-0.07 (0.02)	0.04 (0.849)	-0.06 (0.02)	0.13 (0.718)	-0.07 (0.01)	1.30 (0.255)	0.11 (0.01)	21.00 ($<$ .001)
Gpt-4.1-Nano	-0.08 (0.02)	2.62 (0.106)	-0.08 (0.02)	1.72 (0.191)	-0.03 (0.01)	0.42 (0.516)	0.10 (0.01)	6.55 (0.011)
Llama-3.3-70B-Instruct-Turbo	-0.05 (0.01)	11.52 ($<$ .001)	-0.04 (0.01)	2.21 (0.137)	-0.04 (0.01)	0.20 (0.654)	0.03 (0.01)	4.30 (0.038)
Llama-4-Scout-17B-16E-Instruct	-0.05 (0.01)	3.72 (0.054)	-0.03 (0.01)	0.20 (0.652)	-0.04 (0.01)	0.07 (0.787)	0.03 (0.01)	3.09 (0.079)
Magistral-Medium-2506	-0.06 (0.02)	5.33 (0.021)	-0.02 (0.02)	0.22 (0.640)	-0.07 (0.01)	1.26 (0.262)	-0.03 (0.01)	1.94 (0.164)
Meta-Llama-3-8B-Instruct-Lite	-0.07 (0.01)	15.04 ($<$ .001)	-0.03 (0.01)	1.78 (0.183)	-0.06 (0.01)	0.89 (0.344)	0.04 (0.01)	12.08 ($<$ .001)
Mistral-Medium-Latest	-0.06 (0.02)	0.81 (0.369)	-0.06 (0.02)	0.55 (0.460)	-0.01 (0.01)	2.27 (0.132)	0.02 (0.01)	0.63 (0.429)
Mistral-Small	-0.04 (0.02)	0.99 (0.321)	-0.05 (0.01)	2.78 (0.096)	-0.01 (0.01)	0.87 (0.351)	0.01 (0.01)	0.13 (0.718)
O4-Mini	-0.05 (0.02)	0.10 (0.751)	-0.03 (0.02)	5.86 (0.016)	-0.01 (0.01)	0.63 (0.428)	0.03 (0.01)	3.24 (0.072)
Qwen2.5-72B-Instruct-Turbo	-0.05 (0.02)	1.90 (0.169)	-0.05 (0.01)	1.27 (0.261)	-0.01 (0.01)	0.02 (0.894)	0.03 (0.01)	0.82 (0.366)
Qwq-32B	-0.06 (0.01)	0.16 (0.686)	-0.05 (0.02)	0.37 (0.542)	-0.04 (0.01)	0.01 (0.905)	0.03 (0.01)	5.34 (0.021)

Note. The table reports the average treatment effect (ATE) and standard error for each Jigsaw attribute, comparing responses to What/How versus Hobson's Choice interrogatives. Levene's test F-statistics and p-values assess equality of variances between the two prompt types.

Appendix O. (continued)

Model	Personal Story		Reasoning		Respect	
	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)	ATE (SE)	Levene F (p)
Full Sample	-0.04 (0.00)	6.61 (0.010)	0.02 (0.00)	79.78 ($<$ .001)	-0.04 (0.00)	63.05 ($<$ .001)
Claude-3-Haiku-20240307	-0.12 (0.01)	5.12 (0.024)	0.01 (0.01)	3.53 (0.060)	-0.06 (0.01)	15.09 ($<$ .001)
Claude-3-Opus-20240229	-0.07 (0.01)	2.36 (0.125)	-0.01 (0.01)	0.00 (0.993)	-0.05 (0.01)	9.13 (0.003)
Claude-Sonnet-4-20250514	-0.08 (0.02)	1.30 (0.255)	-0.00 (0.01)	2.22 (0.136)	-0.06 (0.01)	7.61 (0.006)
Deepseek-R1	-0.01 (0.02)	0.03 (0.853)	-0.02 (0.01)	0.26 (0.610)	-0.04 (0.01)	4.25 (0.040)
Deepseek-R1-0528-Tput	-0.02 (0.02)	0.64 (0.422)	-0.00 (0.01)	0.01 (0.917)	-0.04 (0.01)	4.55 (0.033)
Gemini-1.5-Pro	-0.01 (0.02)	0.06 (0.799)	0.03 (0.01)	6.05 (0.014)	-0.02 (0.01)	0.02 (0.884)
Gemini-2.5-Flash	-0.02 (0.02)	0.08 (0.779)	0.03 (0.01)	5.83 (0.016)	-0.05 (0.01)	1.49 (0.223)
Gemma-2-27B-It	-0.01 (0.02)	0.03 (0.865)	0.02 (0.01)	4.08 (0.044)	-0.03 (0.01)	1.57 (0.210)
Gemma-3N-E4B-It	-0.03 (0.01)	0.01 (0.921)	0.02 (0.01)	3.85 (0.050)	-0.04 (0.01)	10.82 (0.001)
Gpt-4.1	-0.04 (0.02)	2.22 (0.136)	0.04 (0.01)	10.36 (0.001)	-0.06 (0.01)	7.83 (0.005)
Gpt-4.1-Mini	-0.07 (0.01)	3.76 (0.053)	0.09 (0.01)	22.00 ($<$ .001)	-0.04 (0.01)	5.62 (0.018)
Gpt-4.1-Nano	-0.06 (0.01)	0.25 (0.617)	0.05 (0.01)	7.96 (0.005)	-0.05 (0.01)	8.13 (0.004)
Llama-3.3-70B-Instruct-Turbo	-0.03 (0.01)	0.96 (0.328)	0.01 (0.01)	3.87 (0.050)	-0.03 (0.01)	6.16 (0.013)
Llama-4-Scout-17B-16E-Instruct	-0.04 (0.01)	0.07 (0.795)	0.02 (0.01)	6.03 (0.014)	-0.03 (0.01)	0.91 (0.339)
Magistral-Medium-2506	-0.04 (0.02)	6.89 (0.009)	-0.03 (0.01)	3.05 (0.081)	-0.04 (0.01)	5.23 (0.022)
Meta-Llama-3-8B-Instruct-Lite	-0.06 (0.01)	4.11 (0.043)	0.03 (0.01)	11.71 ($<$ .001)	-0.04 (0.01)	3.80 (0.051)
Mistral-Medium-Latest	-0.03 (0.02)	0.01 (0.914)	0.00 (0.01)	0.16 (0.691)	-0.04 (0.01)	0.35 (0.552)
Mistral-Small	-0.04 (0.01)	0.00 (0.960)	-0.00 (0.01)	0.13 (0.717)	-0.04 (0.01)	1.19 (0.277)
O4-Mini	-0.03 (0.01)	2.59 (0.108)	0.03 (0.01)	4.52 (0.034)	-0.03 (0.01)	1.70 (0.193)
Qwen2.5-72B-Instruct-Turbo	-0.03 (0.01)	0.16 (0.687)	0.00 (0.01)	0.58 (0.447)	-0.04 (0.01)	1.28 (0.258)
Qwq-32B	-0.05 (0.02)	0.53 (0.466)	0.02 (0.01)	4.65 (0.031)	-0.04 (0.01)	0.49 (0.483)

Note. The table reports the average treatment effect (ATE) and standard error for each Jigsaw attribute, comparing responses to What/How versus Hobson's Choice interrogatives. Levene's test F-statistics and p-values assess equality of variances between the two prompt types.



		Affinity	Compassion	Curiosity	Nuance	Personal story	Reasoning	Respect
Anthropic	Claude 3 haiku	-0.10 (-0.12, -0.08)	-0.06 (-0.07, -0.05)	-0.05 (-0.07, -0.04)	0.03 (0.02, 0.04)	-0.12 (-0.15, -0.10)	0.01 (-0.00, 0.01)	-0.06 (-0.06, -0.05)
	Claude 3 opus	-0.09 (-0.11, -0.07)	-0.05 (-0.06, -0.04)	-0.05 (-0.06, -0.03)	0.01 (-0.00, 0.02)	-0.07 (-0.09, -0.05)	-0.01 (-0.02, 0.00)	-0.05 (-0.06, -0.04)
	Claude sonnet 4	-0.13 (-0.15, -0.11)	-0.07 (-0.09, -0.05)	-0.07 (-0.09, -0.06)	0.01 (-0.00, 0.02)	-0.08 (-0.10, -0.06)	-0.00 (-0.01, 0.01)	-0.06 (-0.07, -0.05)
DeepSeek	Deepseek r1	-0.07 (-0.09, -0.05)	-0.04 (-0.06, -0.03)	-0.02 (-0.04, -0.01)	-0.02 (-0.03, -0.01)	-0.01 (-0.03, 0.00)	-0.02 (-0.03, -0.01)	-0.04 (-0.05, -0.03)
	Deepseek r1 0528 tput	-0.05 (-0.07, -0.03)	-0.03 (-0.05, -0.02)	-0.02 (-0.04, -0.01)	0.00 (-0.00, 0.01)	-0.02 (-0.04, -0.01)	-0.00 (-0.01, 0.00)	-0.04 (-0.05, -0.03)
Google	Gemini 1.5 pro	-0.01 (-0.03, 0.01)	-0.03 (-0.04, -0.01)	0.04 (0.02, 0.06)	0.06 (0.04, 0.08)	-0.01 (-0.03, 0.00)	0.03 (0.02, 0.04)	-0.02 (-0.03, -0.01)
	Gemini 2.5 flash	-0.07 (-0.08, -0.05)	-0.04 (-0.05, -0.02)	-0.01 (-0.03, 0.00)	0.04 (0.03, 0.06)	-0.02 (-0.03, -0.00)	0.03 (0.01, 0.04)	-0.05 (-0.06, -0.04)
	Gemma 2 27b it	-0.04 (-0.06, -0.02)	-0.05 (-0.07, -0.04)	-0.04 (-0.05, -0.02)	0.04 (0.03, 0.05)	-0.01 (-0.03, 0.01)	0.02 (0.01, 0.03)	-0.03 (-0.04, -0.02)
	Gemma 3n e4b it	-0.07 (-0.08, -0.05)	-0.05 (-0.06, -0.04)	-0.07 (-0.09, -0.06)	0.03 (0.02, 0.04)	-0.03 (-0.05, -0.02)	0.02 (0.01, 0.03)	-0.04 (-0.05, -0.03)
Meta	Llama 3 8b instruct lite	-0.07 (-0.09, -0.05)	-0.03 (-0.05, -0.02)	-0.06 (-0.07, -0.04)	0.04 (0.02, 0.06)	-0.06 (-0.08, -0.05)	0.03 (0.01, 0.05)	-0.04 (-0.05, -0.02)
	Llama 3.3 70b instruct turbo	-0.05 (-0.07, -0.03)	-0.04 (-0.05, -0.02)	-0.04 (-0.05, -0.02)	0.03 (0.02, 0.04)	-0.03 (-0.05, -0.01)	0.01 (0.00, 0.03)	-0.03 (-0.04, -0.02)
	Llama 4 scout 17b 16e instruct	-0.05 (-0.07, -0.04)	-0.03 (-0.05, -0.02)	-0.04 (-0.06, -0.03)	0.03 (0.02, 0.04)	-0.04 (-0.06, -0.03)	0.02 (0.01, 0.03)	-0.03 (-0.04, -0.02)
MistralAI	Magistral medium	-0.06 (-0.09, -0.03)	-0.02 (-0.04, -0.00)	-0.07 (-0.10, -0.05)	-0.03 (-0.05, -0.02)	-0.04 (-0.07, -0.02)	-0.03 (-0.04, -0.02)	-0.04 (-0.05, -0.03)
	Mistral medium	-0.06 (-0.08, -0.04)	-0.06 (-0.07, -0.05)	-0.01 (-0.02, 0.00)	0.02 (0.01, 0.03)	-0.03 (-0.04, -0.01)	0.00 (-0.01, 0.01)	-0.04 (-0.05, -0.03)
	Mistral small	-0.04 (-0.06, -0.03)	-0.05 (-0.07, -0.04)	-0.01 (-0.02, 0.01)	0.01 (0.00, 0.03)	-0.04 (-0.05, -0.02)	-0.00 (-0.01, 0.00)	-0.04 (-0.05, -0.03)
OpenAI	Gpt 4.1	-0.09 (-0.11, -0.07)	-0.08 (-0.09, -0.06)	-0.02 (-0.04, -0.01)	0.06 (0.04, 0.08)	-0.04 (-0.06, -0.02)	0.04 (0.03, 0.06)	-0.06 (-0.07, -0.05)
	Gpt 4.1 mini	-0.07 (-0.09, -0.05)	-0.06 (-0.07, -0.04)	-0.07 (-0.09, -0.05)	0.11 (0.10, 0.13)	-0.07 (-0.09, -0.05)	0.09 (0.07, 0.10)	-0.04 (-0.05, -0.03)
	Gpt 4.1 nano	-0.08 (-0.10, -0.06)	-0.08 (-0.10, -0.07)	-0.03 (-0.05, -0.00)	0.10 (0.08, 0.11)	-0.06 (-0.08, -0.04)	0.05 (0.03, 0.07)	-0.05 (-0.06, -0.04)
	O4 mini	-0.05 (-0.07, -0.03)	-0.03 (-0.05, -0.02)	-0.01 (-0.02, 0.01)	0.03 (0.02, 0.05)	-0.03 (-0.05, -0.01)	0.03 (0.01, 0.05)	-0.03 (-0.04, -0.02)
Qwen	Qwen2.5 72b instruct turbo	-0.05 (-0.07, -0.03)	-0.05 (-0.06, -0.04)	-0.01 (-0.03, 0.00)	0.03 (0.02, 0.04)	-0.03 (-0.04, -0.01)	0.00 (-0.01, 0.01)	-0.04 (-0.05, -0.03)
	Qwq 32b	-0.06 (-0.08, -0.04)	-0.05 (-0.07, -0.04)	-0.04 (-0.05, -0.02)	0.03 (0.02, 0.04)	-0.05 (-0.07, -0.03)	0.02 (0.01, 0.03)	-0.04 (-0.05, -0.03)

Appendix P. Average Marginal Effects with 95% Confidence Intervals of Interrogative Type on Bridging Attribute Scores by LLM

Note. Cell values show the average marginal effect with 95% confidence intervals in parentheses (lower bound, upper bound). Color indicates direction: red hues represent positive effects, indicating stronger expression of the attribute in responses to What/How interrogatives (compared to Hobson's Choice), while blue hues represent negative effects, indicating reduced expression. Confidence intervals were derived via parametric Monte Carlo simulation (1,000 draws). Average marginal effects represent the expected change in the attribute score when changing the question type, holding LLM constant.

References

Aoki, G. (2024). *Large Language Models in Politics and Democracy: A Comprehensive Survey*. arXiv. https://doi.org/10.48550/arXiv.2412.04498

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, *111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Barrie, C., Palmer, A., & Spirling, A. (2025). *Replication for Language Models*.

Belnap, N. D., & Steel, T. B. (1976). *The logic of questions and answers*. Yale University Press.

Bick, A., Blandin, A., & Deming, D. J. (2024). *The Rapid Adoption of Generative AI*. https://doi.org/10.20955/wp.2024.027

Bloom, B. S. (1986). *Taxonomy of educational objectives.* (29. print). Longman.

Brachman, M., El-Ashry, A., Dugan, C., & Geyer, W. (2025). *Current and Future Use of Large Language Models for Knowledge Work*. arXiv. https://doi.org/10.48550/arXiv.2503.16774

Caluwaerts, D., Bernaerts, K., Kesberg, R., Smets, L., & Spruyt, B. (2023). Deliberation and polarization: A multi-disciplinary review. *Frontiers in Political Science*, *5*. https://doi.org/10.3389/fpos.2023.1127372

Chen, D., Parsa, R., Swanson, K., Nunez, J.-J., Critch, A., Bitterman, D. S., Liu, F.-F., & Raman, S. (2025). Large language models in oncology: A review. *BMJ Oncology*, *4*(1), e000759. https://doi.org/10.1136/bmjonc-2025-000759

Chen, X., He, B., Lin, H., Han, X., Wang, T., Cao, B., Sun, L., & Sun, Y. (2024). *Spiral of Silence: How is Large Language Model Killing Information Retrieval? – A Case Study on Open Domain Question Answering*. arXiv. https://doi.org/10.48550/arXiv.2404.10496

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training* *of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://doi.org/10.48550/ARXIV.1810.04805

Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2023). *Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2306.04746

Egami, N., Hinck, M., Stewart, B. M., & Wei, H. (2025). *Dsl: Design-based supervised learning*. http://naokiegami.com/dsl/

Elim, E. H. S. Y. (2024). Promoting cognitive skills in AI-supported learning environments: The integration of bloom's taxonomy. *Education 3-13*, 1–11. https://doi.org/10.1080/03004279.2024.2332469

Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *Npj Digital Medicine*, *8*(1), 263. https://doi.org/10.1038/s41746-025-01684-1

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. arXiv. https://doi.org/10.48550/arXiv.1506.02142

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2022). *A Survey of Uncertainty in Deep Neural Networks*. arXiv. https://doi.org/10.48550/arXiv.2107.03342

Grice, H. P. (1991). *Studies in the way of words*. Harvard university press.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks*. arXiv. https://doi.org/10.48550/arXiv.1706.04599

Hakkarainen, K., & Sintonen, M. (2002). The Interrogative Model of Inquiry and Computer-Supported Collaborative Learning. *Science and Education*, *11*(1), 25–43. https://doi.

org/10.1023/a:1013076706416

Huang, H.-Y., Yang, Y., Zhang, Z., Lee, S., & Wu, Y. (2024). *A Survey of Uncertainty Estimation in LLMs: Theory Meets Practice*. arXiv. https://doi.org/10.48550/arXiv.2410.15326

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, *8*, 423–438. https://doi.org/10.1162/tacl_a_00324

Kadt, D. de, & Grzymala-Busse, A. (2025). *Good description*. https://github.com/ddekadt/good_description/blob/main/good_description_ddk_agb.pdf

Kasirzadeh, A., & Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, *36*(2). https://doi.org/10.1007/s13347-023-00606-x

Kharchenko, J., Roosta, T., Chadha, A., & Shah, C. (2025). *How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions*. arXiv. https://doi.org/10.48550/arXiv.2406.14805

Kim, Y., Chin, B., Son, K., Kim, S., & Kim, J. (2025). *Applying the Gricean Maxims to a Human-LLM Interaction Cycle: Design Insights from a Participatory Approach*. arXiv. https://doi.org/10.48550/ARXIV.2503.00858

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). *The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models*. https://doi.org/10.48550/ARXIV.2404.16019

Koralus, P. (2023). *Reason and Inquiry: The Erotetic Theory*. Oxford University Press, Incorporated.

Koralus, P., Moss, S., & Todd, M. (2025). *PyETR*. University of Oxford; University of Birmingham; Dreaming Spires.

Krause, L., & Vossen, P. T. J. M. (2024). The Gricean Maxims in NLP - A Survey. *Proceedings of the 17th International Natural Language Generation Conference*, 470–485. https://doi.org/10.18653/v1/2024.inlg-main.39

Laux, J., Stephany, F., & Liefgreen, A. (2023). *Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy*. arXiv. https://doi.org/10.48550/ARXIV.2312.14565

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). *A New Generation of Perspective API: Efficient Multilingual Character-level Transformers*. arXiv. https://doi.org/10.48550/arXiv.2202.11176

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large Language Models Understand and Can be Enhanced by Emotional Stimuli*. arXiv. https://doi.org/10.48550/arXiv.2307.11760

Luo, Y., Liu, T., Pang, P. C.-I., McKay, D., Chen, Z., Buchanan, G., & Chang, S. (2025). *Enhanced Bloom's Educational Taxonomy for Fostering Information Literacy in the Era of Large Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2503.19434

Massing, N., & Schneider, S. L. (2017). Degrees of competency: The relationship between educational qualifications and adult skills across countries. *Large-Scale Assessments in Education*, *5*(1), 6. https://doi.org/10.1186/s40536-017-0041-y

Morucci, M., & Spirling, A. (2024). *Model Complexity for Supervised Learning: Why Simple Models Almost Always Work Best, And Why It Matters for Applied Research*.

Novoa, G., Echelbarger, M., Gelman, A., & Gelman, S. A. (2023). Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, *120*(47), e2309361120. https://doi.org/10.1073/pnas.2309361120

Ovadya, A., & Thorburn, L. (2023). *Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance*. arXiv. https://doi.org/10.48550/arXiv.2301.09976

Panzeri, F., & Foppolo, F. (2021). Children's and adults' sensitivity to Gricean maxims and to the maximize presupposition principle. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.624628

Pavlovic, M. (2025). *Understanding Model Calibration – A gentle introduction and visual exploration of calibration and the expected calibration error (ECE)*. arXiv. https://doi.org/10.48550/arXiv.2501.19047

Pearl, J. (2013). *Understanding Simpson's Paradox* [{SSRN} {Scholarly} {Paper}]. Social Science Research Network. https://doi.org/10.2139/ssrn.2343788

Pereira, A., & Ortiz, K. Z. (2022). Language skills differences between adults without formal education and low formal education. *Psicologia: Reflexão e Crítica*, *35*(1), 4. https://doi.org/10.1186/s41155-021-00205-9

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N., & Lewis, M. (2023). Measuring and Narrowing the Compositionality Gap in Language Models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5687–5711). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.378

Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M., & Bagheri, E. (2025). *Benchmarking Prompt Sensitivity in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2502.06065

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. (2022). Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/2022.naacl-main.13

Saad, F., Murukannaiah, P. K., & Singh, M. P. (2025). *Gricean Norms as a Basis for Effective Collaboration*. arXiv. https://doi.org/10.48550/ARXIV.2503.14484

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The Prompt Report: A Systematic*

Survey of Prompt Engineering Techniques. arXiv. https://doi.org/10.48550/arXiv.2406.06608

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting*. arXiv. https://doi.org/10.48550/arXiv.2310.11324

Scorzato, L. (2024). Reliability and Interpretability in Science and Deep Learning. *Minds and Machines*, *34*(3), 27. https://doi.org/10.1007/s11023-024-09682-0

Shen, K., & Kejriwal, M. (2023). An experimental study measuring the generalization of fine-tuned language representation models across commonsense reasoning benchmarks. *Expert Systems*, *40*(5), e13243. https://doi.org/10.1111/exsy.13243

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2102.02503

Thapa, S., Shiwakoti, S., Shah, S. B., Adhikari, S., Veeramani, H., Nasim, M., & Naseem, U. (2025). Large language models (LLM) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining*, *15*(1), 4. https://doi.org/10.1007/s13278-025-01428-9

Tran, T. T. H. T., Nguyen Thi Quynh Hoa. (2020). An investigation into the flouting of conversational maxims employed by male and female guests in the American talk show ”The Ellen Show”. *Journal of Science and Technology Issue on Information and Communications Technology*, 117–122. https://doi.org/10.31130/jst-ud2020-069e

Vassimon Manela, D. de, Errington, D., Fisher, T., Breugel, B. van, & Minervini, P. (2021). Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2232–2242). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.190

Viveros-Muñoz, R., Carrasco-Sáez, J., Contreras-Saavedra, C., San-Martín-Quiroga, S., & Contreras-Saavedra, C. E. (2025). Does the Grammatical Structure of Prompts Influence the Responses of Generative Artificial Intelligence? An Exploratory Analysis in Spanish. *Applied Sciences*, *15*(7), 3882. https://doi.org/10.3390/app15073882

Wang, T., Wang, Y., Zhou, J., Peng, B., Song, X., Zhang, C., Sun, X., Niu, Q., Liu, J., Chen, S., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Fei, C., Yin, C. H., & Yan, L. K. (2025). *From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence*. arXiv. https://doi.org/10.48550/arXiv.2501.03282

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. arXiv. https://doi.org/10.48550/arXiv.2112.04359

Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). *Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment*. arXiv. https://doi.org/10.48550/ARXIV.2504.14232

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). *Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance*. arXiv. https://doi.org/10.48550/arXiv.2402.14531

Zheng, Z., Qiu, S., Fan, L., Zhu, Y., & Zhu, S.-C. (2021). GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2074–2085). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.182

Zhu, J., Kempermann, M., Cannanure, V. K., Hartland, A., Navarrete, R. M., Carteny, G., Braun, D., & Weber, I. (2025). *Learn, Explore and Reflect by Chatting: Understanding the Value of an LLM-Based Voting Advice Application Chatbot*. arXiv. https://doi.org/

10.48550/arXiv.2505.09806