Bachelor's Thesis

Consistency of LLMs in Multiple-Choice Question Answering: An Empirical Evaluation across Varying Scales

Department of Statistics
Ludwig-Maximilians-Universität München

Martin Kandlinger

Munich, December 26th, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Dr. Matthias Aßenmacher and Bolei Ma

Abstract

Large Language Models are increasingly deployed as simulated respondents for social science research, yet their reliability in maintaining consistent logical stances remains under-examined. This thesis evaluates the internal consistency of four major open-weights model families, comparing Base and Instruction-Tuned variants across survey-style subjective multiple-choice tasks. By subjecting models to systematic perturbations, including semantic rewording of answer options, randomized option shuffling, and prompt reformatting, this study isolates genuine semantic stability from mechanical generation artifacts.

The results reveal a fundamental divergence driven by the alignment process. Base models demonstrate high volatility and susceptibility to choice architecture, often defaulting to positional biases or collapsing into random guessing when structural anchors are removed. In contrast, Instruction Tuning acts as a critical reliability filter, reducing response volatility by approximately 50 % and effectively grounding model outputs in semantic content. While models generally tolerate linguistic rewording, randomized option ordering proves to be a significantly more rigorous stress test. These findings imply that for synthetic survey research, the use of Instruction-Tuned models and ensemble verification strategies is mandatory to distinguish the signal of synthetic opinion from the noise of stochastic text generation.

Contents

1 Introduction

Large Language Models (LLMs) have become essential in natural language processing, powering a wide range of applications from information retrieval to conversational agents and increasingly serving as automated evaluators in academic, medical, and social science domains. Among the arsenal of evaluation methods, multiple-choice question answering (MCQA) occupies a central role. As Balepur et al. (2025) note, MCQA is popular for LLM evaluation due to its simplicity and similarity to human testing in educational and survey settings. Furthermore, Mucciaccia et al. (2025) observe that MCQA provides a method to objectively and efficiently assess a diverse set of underlying abilities, from recalling factual information and reasoning to judgment and preference measurement.
However, the reliability of this method is increasingly questioned. Recent research reveals that the outcomes of these probabilistic, next-token predictors are highly sensitive to seemingly minor choices, such as option order or prompt design (Zheng et al., 2024, Molfese et al., 2025). Unlike human respondents who typically parse semantic meaning, LLMs may rely on surface-level statistical patterns, raising concerns about the robustness and reproducibility of studies that rely on them as simulated subjects.

When multiple-choice question formats are adapted for survey-style settings, one conceptual question may be presented through many different Likert-type scales or response phrasings. Yet it remains uncertain to what extent LLMs maintain consistent judgments when the scale design, phrasing, or polarity changes, posing challenges for the validity and reproducibility of evaluations (Lee et al., 2025). A recurring theme in recent discussions is the need to assess not only pointwise accuracy but also the consistency of LLMs, both when tested repeatedly with the same prompt (self-consistency), and when answer scales, labels, or polarities are varied (inter-scale consistency).
Building on these concerns, this study investigates the consistency of LLMs in MCQA, specifically examining how model responses shift when Likert-type answer options are varied in number, reworded to have the same meaning, or presented in unipolar or bipolar formats. Such variation in scale design is common in surveys and known to influence how humans interpret and select answers (see, e.g., Menold, 2021), making it critical to understand whether LLMs are also sensitive to these factors, especially since even small changes in answer wording or prompt formulation can result in different model responses (Lunardi et al., 2025).

The empirical investigation uses a curated set of multiple-choice questions from the OpinionQA corpus (Santurkar et al., 2023) to probe whether, and how, LLM answers change under varying scale formulations and response phrasings. To guide this investigation, this thesis addresses the following research questions:

1. To what extent does the rewording of answer options, while maintaining semantic equivalence, impact the consistency of LLM responses in subjective MCQA tasks?

2. How does the permutation of answer option order influence the stability of model decisions, and does it reveal positional bias?

3. How do structural anchors (e.g., numeric vs. alphabetic vs. no option labels) and variation in option availability (e.g., inclusion of non-substantive options like

"Refused" option) influence the consistency of LLM predictions?

The work is structured as follows. Section 2 situates the study in the current literature, detailing the theoretical foundations of Transformer-based models, general survey methodology, and existing findings on LLM biases. Section 3 introduces the general methodological framework, outlining the end-to-end pipeline employed to process raw data into analyzable results. Section 4 details the experimental design, including data curation and adaptation, consistency metrics, scale-variant generation, and evaluation protocols. Section 5 presents the findings on how variations in answer scales and phrasings impact response consistency across different LLMs. Section 6 analyzes trade-offs, threats to validity, and implications for future evaluator design. Finally, Section 7 summarizes key findings and their implications.

2 Related Work

2.1 Theoretical Foundations of Transformer-Based LLMs

Prior to the current generation of LLMs, natural language processing relied heavily on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997). These architectures processed text sequentially, ingesting one word (or "token") at a time. This process is analogous to reading a text one word at a time without grasping the "bigger picture." As a result, RNNs struggled with long-range dependencies, often forgetting context from the beginning of a paragraph by the time they reached the end, and were computationally inefficient, as the processing of the n-th word depended on the completion of the $(n-1)$-th word.
The introduction of the *Transformer* architecture by Vaswani et al. (2017) fundamentally shifted this paradigm. Instead of processing words sequentially, Transformers process the entire input sequence in parallel. This allows the model to "look at" all words simultaneously, determining their contextual importance and connections between them regardless of their distance in the text. This capability relies on a mechanism known as *Self-Attention*.

2.1.1 The Mechanism of Self-Attention

Self-Attention describes the process where the model determines the relevance of words in a text using a mathematical system of weights. This allows the model to compare the relevance of different words relative to one another. The Transformer consists of a stack of layers, each layer contains two primary sub-components: a Self-Attention mechanism (where tokens share information) followed by a Feed-Forward Network (which processes each token individually).
To perform self-attention, the model assigns three learned vectors to every input token: a **Query** (Q), a **Key** (K), and a **Value** (V):

- The **Query** represents what the current token is looking for (e.g., a pronoun looking for its noun or a noun looking for adjectives that further describe it).
- The **Key** acts as a label or identifier for every other token in the sequence (e.g., identifying a token as a "subject" or "object").
- The **Value** contains the actual semantic content of the token.

Keys that correspond to specific Queries are closely aligned in the vector space (meaning they point in similar directions). To measure how well each Key matches the current Query, the model computes the dot product across every possible pair. A high dot product indicates a strong semantic relationship (e.g., connecting "he" with "the man" earlier in the sentence); in such cases, the Query is said to "attend to" the Key. These raw scores are then normalized into probabilities using a Softmax function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k represents the dimension of the key vectors. The scaling factor $\sqrt{d_k}$ is critical for numerical stability; without it, the dot products could grow extremely large, pushing the softmax function into regions with extremely small gradients (vanishing gradients), which would effectively halt the training process (Vaswani et al., 2017).

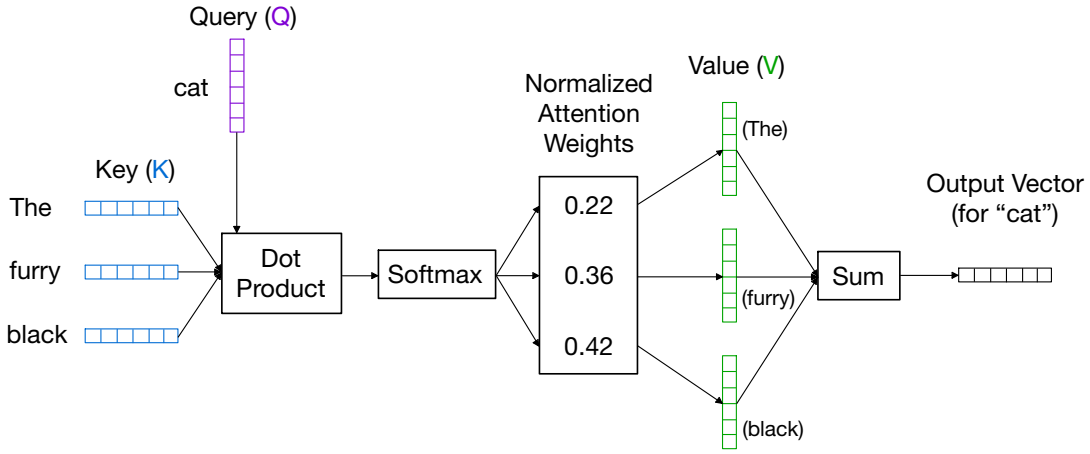


Figure 1: Simplified Self-Attention for "The furry black cat". The Query ("cat") computes weights via dot products with all Keys. These weights scale the Value vectors to prioritize information relevant to the Query.

The resulting output is a weighted sum of the *Values* (visualized in Figure 1), where irrelevant words are drowned out and relevant words are amplified. By stacking multiple such layers (Multi-Head Attention), the model learns complex linguistic nuances, such as syntax, sentiment, and coreference, far more effectively than sequential models. Furthermore, because this operation is matrix-based rather than recurrent, it allows for parallelization on modern GPU hardware, enabling the training of models on large datasets.

2.1.2 Architectures: Encoders and Decoders

The original Transformer consisted of two components: an *Encoder* (which processes the input) and a *Decoder* (which generates output). Modern LLMs typically specialize in one:

- **Encoder Models (e.g., BERT):** The Encoder processes the input sequence bidirectionally, allowing every token to attend to both past and future context. This architecture, popularized by BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), excels at discriminative tasks like sentiment analysis or classification because it builds a comprehensive representation of the entire sentence context, but is less suitable for text generation.
- **Decoder Models (e.g., GPT, Llama):** The Decoder is unidirectional. It employs a masking mechanism (often called a "causal mask") that prevents a token from attending to future tokens. The model predicts the next token t_i based solely on the history $t_1 \dots t_{i-1}$ (Radford et al., 2019). This autoregressive property, meaning the model generates output iteratively, consuming its own previous predictions as

input for the next step, makes Decoder architectures (such as the Generative Pre-trained Transformer, or GPT) inherently generative.

The models examined in this thesis, Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Qwen (Yang et al., 2024), and Gemma (Gemma Team et al., 2024), are all *Decoder-only* architectures. They function as probabilistic engines, trained to predict the most likely continuation of a text sequence (Brown et al., 2020).

2.1.3 Alignment and Post-Training

A raw, pre-trained Decoder model (often referred to as a "Base" model) is primarily a text completer. If prompted with "What is the capital of France?", a Base model might continue with "and what is its population?" rather than answering the question, as it is merely mimicking the pattern of a list of questions found in its training data.
To transform these predictors into helpful assistants, models undergo a secondary process known generally as *Alignment* or *Post-Training*. This typically involves two steps:

1. **Instruction Tuning:** The model undergoes Supervised Fine-Tuning (SFT) on a dataset of formatted (Instruction, Response) pairs. This stage teaches the model to recognize and execute tasks rather than just completing text patterns (Wei et al., 2022).

2. **Reinforcement Learning from Human Feedback (RLHF):** To further align the model with human intent, Ouyang et al. (2022) introduced a preference optimization stage. The model generates multiple candidate answers, which are ranked by humans to train a *reward model* that serves as a proxy for human preference. The LLM is then updated via reinforcement learning algorithms to maximize the score from this reward model, optimizing for abstract qualities like helpfulness and safety.

This distinction is central to this thesis, which compares Base models against their Aligned counterparts. Note that while "Instruction Tuning" technically refers only to the first stage (SFT), this thesis follows the naming convention of model providers (e.g., *Llama-3-Instruct*) and uses the term "Instruction-Tuned" (IT) to refer to models that have undergone the full alignment pipeline (SFT + RLHF).

2.2 General Survey Methodology and Psychometrics

The reliability of survey data is fundamentally determined by the quality of the measurement instrument. Within Classical Test Theory (CTT), an observed response is conceptualized as the sum of a latent *true score*, representing the respondent's underlying attitude or opinion, and a *measurement error* component that captures random or systematic noise introduced by the instrument, the context, or the respondent (Lord et al., 1968). From this perspective, the central objective of rigorous survey design is to minimize measurement error, such that observed variation in responses reflects meaningful differences in latent opinions rather than artifacts induced by question wording, response scales, or presentation format.

Reliability and Consistency. Psychometric theory distinguishes between several forms of reliability, each capturing a different aspect of measurement stability. Of particular relevance to this study is *parallel-forms reliability*, which assesses the degree to which two equivalent versions of a measurement instrument yield consistent results (Cronbach, 1951). Traditionally, parallel forms are constructed to differ only in superficial characteristics while targeting the same underlying construct. In the context of Large Language Models, robustness to rewording can be interpreted analogously: a reliable respondent should map semantically equivalent formulations (e.g., "Do you agree?" versus "What is your stance?") to comparable outputs. Substantial variation in responses across such rewordings therefore suggests a lack of internal consistency in the measurement process.

Response Styles and Biases. Human respondents frequently exhibit systematic biases, known as "response styles".

- **Satisficing:** As noted by Krosnick (1991), respondents may avoid the cognitive effort required to optimally answer a question. This often manifests as a tendency to select neutral or safe middle options to avoid taking a strong stance.

- **Acquiescence Bias:** This is defined as the tendency for respondents to agree with statements regardless of their substantive content (Couch and Keniston, 1960). This bias is often attributed to social desirability concerns, satisficing behavior, or a general inclination to appear cooperative. When LLMs are deployed as virtual survey respondents, acquiescence bias becomes particularly salient, as many models are explicitly trained to adopt a helpful and agreeable interaction style. As a result, apparent agreement in model responses may reflect alignment-induced priors rather than stable underlying judgments.

- **Order Effects (Primacy and Recency):** The ordering of response options is another well-established source of measurement error in surveys. Research on response-order effects distinguishes between *primacy effects*, in which respondents disproportionately select options presented early in a list, and *recency effects*, in which later options are favored (Krosnick and Alwin, 1987). Primacy effects are more commonly observed in visually presented questionnaires, where respondents scan options sequentially, whereas recency effects tend to arise in auditory or sequential presentation formats, where later options remain more accessible in working memory. Such effects demonstrate that response distributions can be shaped by presentation order independently of substantive preference, highlighting the importance of careful scale design.

Scale Design Impact. Social science research has long established that the structural properties of the answer scale itself affect responses. Menold (2021) find that verbal agreement scales produce less bias and higher reliability when using unipolar wording (e.g., "do not agree" to "agree") rather than bipolar ("disagree" to "agree"). Understanding these human cognitive mechanisms provides a critical baseline: it allows researchers to determine whether LLM instability is mimicking human-like cognitive biases or resulting from entirely distinct, mechanical artifacts of the attention mechanism.

Taken together, these psychometric principles underscore that variability in responses can arise from properties of the measurement instrument itself, motivating an examination of whether analogous sources of measurement error affect LLMs when they are evaluated using survey-style prompts and response scales.

2.3 LLMs in Survey Research and Evaluation

MCQA Evaluation Biases and Prompt Sensitivity. Multiple-choice question answering has become a cornerstone evaluation method for LLMs, but accumulating evidence demonstrates that model performance is highly sensitive to superficial format features. A primary concern is *selection bias* (often referred to as positional bias), where the model's choice is systematically distorted by the presentation order, regardless of semantic content. Wang et al. (2024) identify a pervasive *primacy bias* in models like GPT-4, where the model preferentially selects the first option presented. Similarly, Zheng et al. (2024) show that simply reordering answer choices can substantially swing model accuracy. For instance, GPT-3.5's score on the MMLU (Massive Multitask Language Understanding) benchmark drops from 67.2 % to 60.9 % when the correct answer is consistently placed last.

Conversely, Wang et al. (2024) also observe that other models, particularly those fine-tuned on instruction-following data, may exhibit *recency bias*, favoring the last option due to the autoregressive nature of generation. This bias often arises because LLMs learn statistical priors for option labels (e.g., preferring "A" over "D") rather than making purely semantic judgments. Consequently, this sensitivity creates a "random guess" baseline that is not uniform, e.g., if an LLM has a high intrinsic preference for option "A", a binary choice question is practically invalid unless the options are permuted and averaged. The problem extends beyond option order to what Ngweta et al. (2025) term *prompt brittleness*, where minor changes in prompt formatting, punctuation, or spacing can produce fluctuations in model accuracy exceeding 76 % between semantically equivalent prompts. Meanwhile, Balepur et al. (2025) critique the MCQ format itself as artificially forced, arguing it oversimplifies subjective tasks and primarily tests answer-selection ability rather than generative reasoning. They note that while MCQs dominate popular benchmarks (appearing in 71 % of GPT-4's evaluation tasks), over 90 % of real user queries are free-form, potentially misleading evaluations by rewarding test-taking skill over general capability.

Further complicating matters, Molfese et al. (2025) demonstrate that evaluation procedures (regex vs. log-prob matching, constrained vs. free-form prompts) can under- or over-estimate performance. In their study, enforcing strict answer formats caused lower measured accuracy, whereas allowing free-form answers improved correctness but required brittle parsing routines.

Consistency and Persona Adoption. Given these format sensitivities, recent work has emphasized measuring consistency alongside accuracy. Lee et al. (2025) formalize this in terms of *self-consistency*, which describes the stability of a model's outputs under repeated identical prompts and *inter-scale consistency*, the stability of outputs when evaluation scales or phrasing are varied. Their findings reveal that high-performing models are not inherently consistent; using different rating formats (e.g. 5-point vs. 10-point

scales, numeric vs. verbal labels) often yields divergent evaluations for the same input. Sampling randomness introduced by different decoding seeds produces nontrivial variability, with non-numerical verbal scales causing larger judgment swings than purely interval scales.

These consistency challenges extend to input formulation. Lunardi et al. (2025) systematically paraphrase thousands of benchmark questions and find that while model rankings remain roughly stable, absolute accuracy drops markedly under paraphrasing. Even top-ranked models can lose significant percentage points when encountering synonymous wording, suggesting that minor phrasing changes can yield "wrong scores" despite the model possessing the requisite knowledge.

Applying these concepts to survey simulation, Rupprecht et al. (2025) argue that response patterns depend heavily on the "persona" the model adopts; an LLM prompted to act as a helpful assistant may gravitate toward agreeableness (Acquiescence Bias), whereas a neutral prompt might result in a different distributional prior. Understanding these priors is crucial, as they can distort the interpretation of Likert-scale results: a "Neutral" response from an LLM might not indicate a lack of opinion, but rather a failure of the model to map its internal probability distribution to the extreme ends of the provided scale.

Emerging evidence indicates LLMs exhibit the same, and sometimes amplified, sensitivities compared to humans. Rupprecht et al. (2025) systematically apply perturbations to World Values Survey items and find consistent human-like biases, such as *recency bias*, where models heavily favor the last-presented option (in some cases approximately 20 times more frequently) and strong sensitivity to semantic paraphrasing. Notably, while higher-capacity models show somewhat greater stability, no model proves robust to scale or phrasing changes. Khan et al. (2025) similarly report large instabilities in LLM answers to "cultural alignment" surveys, where results swing dramatically with minor prompt variations, suggesting that LLM judgments often reflect methodological artifacts rather than true model traits.

Symbol Binding and Mitigation Strategies. The underlying cause of these instabilities may involve inadequate *symbol binding*, the model's ability to associate option content with option labels. Xue et al. (2024) demonstrate that selection bias persists during supervised fine-tuning because models fail to reliably bind answer symbols to their semantic content. Their proposed solution greatly reduces positional bias and substantially improves MCQ accuracy by training with a contrastive "Point-wise Intelligent Feedback" loss that forces the model to match each option's content to its label. However, this approach requires access to model weights and retraining, limiting its applicability to closed-source models.

Alternative mitigation strategies include prompt engineering and evaluation design adjustments. Ngweta et al. (2025) show that training or prompting models with a mixture of formats can reduce brittleness, while careful evaluation protocol design, as emphasized by Molfese et al. (2025), can help ensure that scoring methods accurately capture model capabilities rather than format artifacts.

2.4 The OpinionQA Corpus and Survey Simulation

The OpinionQA corpus (Santurkar et al., 2023) provides a natural testbed for investigating these issues in survey-style settings. The dataset comprises 1,498 opinion-based multiple-choice questions derived from the American Trends Panel surveys conducted by the Pew Research Center (Pew Research Center, 2024). These questions cover a broad spectrum of highly subjective topics, ranging from political ideology and social issues to consumer preferences.
Originally, the corpus was designed to assess the alignment between LLM "views" and human demographic groups. Santurkar et al. (2023) found that base models often exhibit distinct "liberal" tendencies, aligning more closely with left-leaning demographic groups on topics like climate change and gun control. However, while previous work has explored whose opinions LLMs reflect, the question of how consistently they express those opinions across different scale formulations remains underexplored. This thesis leverages OpinionQA not to measure demographic alignment, but as a source of validated, subjective prompts to stress-test the structural stability of LLM evaluations.

In summary, the related literature establishes that LLMs exhibit substantial sensitivity to MCQA format choices, that consistency is a critical but often overlooked dimension of evaluation quality, and that existing mitigation strategies provide only partial solutions. This study builds on these insights by systematically manipulating Likert-type scale designs and measuring the resulting impact on LLM answer consistency, contributing to a more nuanced understanding of LLM robustness in survey-style evaluation tasks.

3 Experimental Framework

This chapter introduces the general methodological framework developed to evaluate the consistency of Large Language Models in multiple-choice settings. While the subsequent chapter details the specific datasets and models used in this study, the framework described here serves as a generalized pipeline for stress-testing LLM decision-making.
The core objective of this framework is to separate an LLM's understanding of a concept from its sensitivity to the format in which that concept is presented. To achieve this, the pipeline takes a standardized survey question and systematically pairs it with different sets of answer options. While the question text remains constant, the answer options vary in wording, granularity, and order. By comparing the model's responses across these variations, the framework allows for a mathematical assessment of consistency.
The workflow consists of four distinct stages, illustrated in Figure 2: *Input Standardization and Classification*, *Scale Generation*, *Standardized Inference*, and *Unified Analysis*.

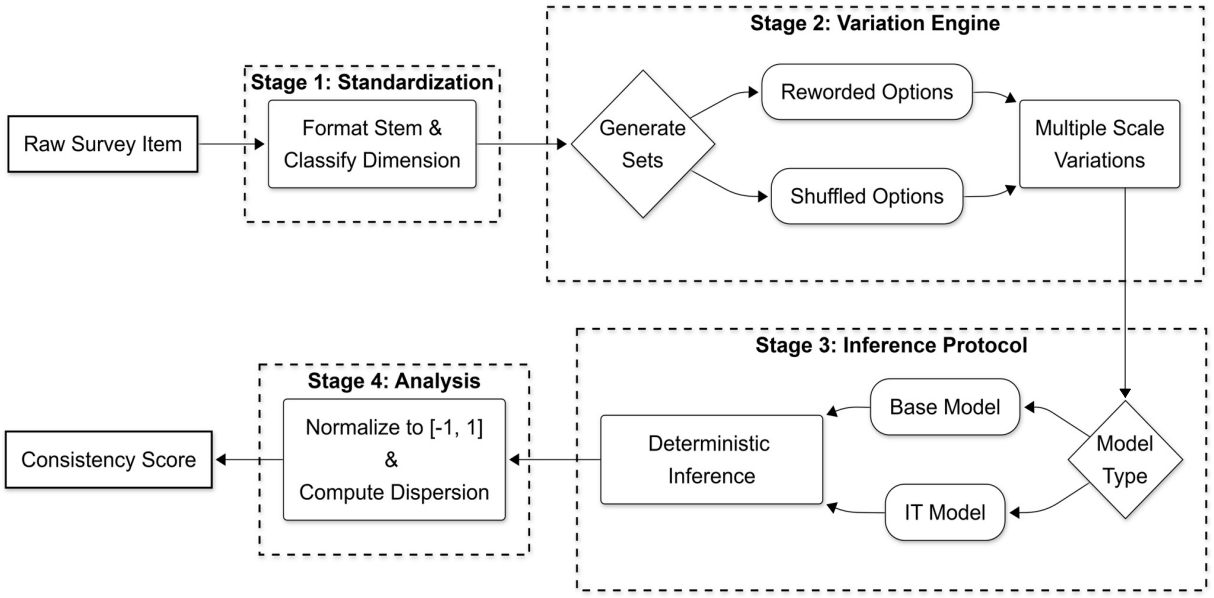


Figure 2: Overview of the experimental framework. The pipeline standardizes a raw question, classifies its underlying dimension (e.g., Agreement), and attaches multiple distinct answer scales. The model's responses to these variations are then normalized into a shared numerical space.

3.1 Stage 1: Input Standardization and Classification

The first stage involves preparing the raw input data to ensure it is compatible with a systematic variation engine (details on the specific implementation are provided in Section 4.1). This process consists of two substeps:

1. Stem Standardization. To isolate the answer options as a variable, the framework relies on the structure of Likert scales. Likert scales are psychometric scales where responses are presented along a continuum rather than as distinct categorical choices (Likert, 1932). Respondents are asked to specify their position on the provided range, for

example, by rating their level of agreement with a statement or their sentiment towards a political party.
Raw survey items often embed the substantive statement into the answer choices to keep the question stem short (e.g., "Which statement do you agree with?"). Because such formats tightly couple the question to specific options, they must be reformatted to include the substantive content entirely in the question stem (e.g., "How much do you agree with the following statement..."). This ensures that the question stem serves as a fixed anchor, allowing the answer options to be swapped out modularly without altering the core inquiry.

2. Dimensional Classification. Once standardized, the underlying latent variable of the question, referred to here as the "Content Dimension", is identified. All questions are classified into a semantic type, such as *Agreement*, *Importance*, or *Frequency*. This classification is critical because it determines which set of answer scales (e.g., "Agree/Disagree" vs. "Important/Unimportant") is semantically appropriate for generating variations in the next stage.

3.2 Stage 2: The Scale Variation Engine

With the question stem fixed and the content dimension identified, the "Variation Engine" generates the specific answer option sets that will be presented to the model. Rather than altering the question, this stage attaches different response scales to the standardized stem. The framework utilizes two distinct strategies for variation:

Reworded Variations (Semantic & Structural Changes). For a given Content Dimension (e.g., Agreement), the engine retrieves multiple pre-defined sets of answer options that represent that dimension. These sets vary systematically in:

- **Cardinality:** The number of options (e.g., from a 4-point to a 5- or 6-point scale).
- **Phrasing:** Using synonymous labels (e.g., substituting "Fully" with "Strongly").
- **Polarity:** The directionality of the scale (Höhne et al., 2022). This includes *bipolar* scales, which measure a construct from negative to positive with a neutral or zero point (e.g., "Strongly agree" to "Strongly disagree"), and *unipolar* scales, which measure the magnitude of a single attribute from zero to high (e.g., "Extremely important" to "Not at all important").

This process results in a "Reworded Dataset" where the same question stem is paired with multiple valid, semantically equivalent answer scales.

Shuffled Variations (Positional Changes). Permutation operations are performed to isolate the effect of option order (Positional Bias). For every question, a single canonical set of options is selected, and random permutations of the answer sequence are generated. This results in a "Shuffled Dataset" where the wording and scale properties remain identical to the source, but the position of each option (e.g., 1st vs. last) changes.

3.3 Stage 3: Standardized Inference Protocol

The third stage manages the interaction with the subject models. While the Variation Engine (Section 3.2) produces the content of the query, the Inference Protocol determines the syntax used to present it. To ensure comparability across different model architectures (e.g., base vs. instruction-tuned), the framework employs a modular system that handles two functions:

Model-Specific Encapsulation. Different model families require distinct input formats to function correctly. The driver encapsulates the standardized question and options into the appropriate template:

- **Instruction-Tuned Models:** Inputs are wrapped in chat-specific structures (e.g., `user` tags) to trigger the model's aligned response behavior.
- **Base Models:** Inputs are formatted as text-completion tasks, often utilizing a forcing suffix (e.g., "The answer is:") to prime the model for an immediate response.

Syntactic Variation and Determinism. Beyond architectural adaptation, this stage allows for the injection of syntactic variations to test robustness against prompt formatting. This includes altering option label formats (e.g., testing "1, 2, 3" vs. "A, B, C"), comparing performance under zero-shot conditions versus few-shot setups, or altering the system instructions to impose specific personas or constraints (e.g., "Answer as a conservative policy expert" vs. "You are taking part in a survey"). Finally, to ensure that the measured consistency reflects the model's internal representation rather than sampling noise, the inference is executed deterministically (typically using greedy decoding with temperature set to zero).

3.4 Stage 4: Unification and Consistency Analysis

The final stage addresses the analytical challenge of comparing outputs drawn from heterogeneous scales. Since the Variation Engine (Section 3.2) produces answers using differing cardinalities (e.g., 4-point vs. 5-point scales) and labels, the raw tokens cannot be directly compared.

Normalization Strategy. To enable cross-scale comparison, the framework maps all valid substantive outputs to a unified latent interval $I \in [-1, 1]$. This mapping assumes that Likert response options represent equidistant points along the underlying semantic continuum. For any scale of cardinality N, the options are projected onto I such that the most positive option maps to 1.0 and the most negative to -1.0. This normalization allows for the calculation of distance metrics between different scales and formulations.

Consistency Quantification. With all responses mapped to a shared numerical space, the framework quantifies stability using distance-based metrics. Rather than measuring accuracy against a ground truth (which is undefined for opinion-based questions), the framework measures "Self-Consistency": the dispersion of the projected values across the generated variations for each individual question, calculated within a fixed prompting configuration.

Computing dispersion at the question level is essential to distinguish instability from opinion diversity. Without this isolation, high variance could simply reflect that the model holds strong but opposing views on different topics (e.g., highly positive for Question A and highly negative for Question B). By restricting the analysis to individual questions, low dispersion implies that the model possesses a robust internal understanding of the specific concept, whereas high dispersion indicates that the model's judgment is brittle, fluctuating based on the presentation format or option order.

4 Experimental Setups

This section details the experimental methodology, beginning with the data preparation and the systematic introduction of Likert-scale variations. It then outlines the prompting strategies, model specifications, and the metrics used to quantify answer consistency. The goal is to standardize questions and response scales so that the analysis can isolate the effect of scale wording, polarity, and cardinality on model choices.

4.1 Data Preparation

The foundation of this study is a carefully curated dataset derived from the OpinionQA corpus (Santurkar et al., 2023) (see Section 2.4 for details on the original corpus). The preparation process involved multiple stages of automated and manual refinement to ensure the questions were suitable for investigating variation sensitivity in LLMs.

Initial Data Unification and Filtering. The original OpinionQA data, comprising questions from the Pew American Trends Panel, was first consolidated into a unified CSV (comma-separated values) format using an automated Python script. This initial dataset then underwent rigorous manual review. Each question was evaluated based on its suitability for LLM administration, specifically checking whether the question implicitly assumed a specific respondent background (e.g., questions asking "While growing up, how often did you..." or "How much ... would you say is in your neighborhood"). Questions deemed inappropriate or unanswerable without giving the models a specific persona were excluded to ensure the final dataset assessed the models' generalized opinions rather than role-playing capabilities.

Adaptation for Likert-Scale Administration. A second manual pass ensured all remaining questions could be coherently answered using Likert-scale response options (as defined in Section 3.1). This required reformatting questions where the substantive content resided in the answer choices rather than the question stem. For instance, questions following the format *"Which statement do you agree with the most?"* with multiple substantive statements as options were transformed into multiple individual questions using the stem *"How much do you agree or disagree with the following statement? ..."*, followed by each respective statement. This adaptation increased the total number of questions while ensuring consistent Likert-scale applicability across the dataset.

Question Typology and Scale Standardization. To enable systematic scale variation, questions were automatically classified based on their Content Dimension (as defined in Section 3.1). This operational categorization, referred to hereafter as the question type, groups items into 15 distinct classes (e.g., Agreement, Importance, Quantity), as detailed in Table 1. The classification was performed using a Python script employing keyword matching and Regular Expressions (RegEx). This grouping allows for consistent answer option sets across questions of the same type. For example, all questions classified as "Agreement" are answered using variants of *["Fully agree", "Somewhat agree", "Somewhat disagree", "Fully disagree"]*.

Table 1: Distribution of Content Dimensions in the final dataset ($N = 1,235$). The "Agreement" type is the most prevalent, reflecting the common structure of the source survey data.

Content Dimension	**Count**	**Percentage**
Agreement	404	32.7 %
Quantity	214	17.3 %
Better / Worse	101	8.2 %
Importance	91	7.4 %
Likelihood	85	6.9 %
Positive / Negative	67	5.4 %
Good / Bad	50	4.0 %
Magnitude of Problem	48	3.9 %
Frequency	38	3.1 %
Performance ("How Well")	38	3.1 %
Priority	35	2.8 %
Acceptance	31	2.5 %
Reason	15	1.2 %
Concern	11	0.9 %
Increase / Decrease	7	0.6 %
Total	**1,235**	**100.0 %**

It is recognized in the survey methodology literature that such Agree-Disagree (AD) scales require respondents to perform an additional cognitive step of mapping their judgment onto an agreement dimension, which can increase cognitive burden and introduce measurement error compared to item-specific questions that directly query the underlying dimension of interest (Dykema et al., 2022). Nonetheless, the AD format is used in this study because the primary goal is to investigate the effect of systematic changes in answer scale properties, rather than to create a methodologically perfect opinion questionnaire. By using a uniform scale structure across diverse topics, the study effectively isolates the impact of the scale parameters (cardinality, wording, polarity) from the semantics of the questions themselves. This approach preserves the spirit of the original survey design (where respondents were asked to select the statement they agree with the most) and maintains a manageable number of question types.

Generation of the "Reworded Dataset" (Scale Variations). To address the research questions regarding phrasing and scale design, five semantically equivalent but differently worded answer option sets were developed for each question type. These systematically vary along three dimensions:

1. **Cardinality:** The number of response options (4, 5, or 6 points).
2. **Phrasing:** Synonymous rewording of labels while maintaining the same semantic meaning (e.g., changing "Strongly agree" to "Fully agree").
3. **Polarity:** The directionality of the scale (unipolar vs. bipolar) as defined in Section 3.2.

Each variation was annotated with its attributes (e.g., "5-point bipolar") and mapped to a numerical scale using equidistant points along the interval $[-1, 1]$. For example, a 5-point scale is mapped to $\{1.0, 0.5, 0.0, -0.5, -1.0\}$. This normalization facilitates the calculation of distance-based consistency metrics across scales of differing lengths.

Generation of the "Shuffled Dataset" (Order Permutations). To investigate positional bias, a second distinct dataset was generated. For the answer options, the original wording was retained if it fit a Likert scale structure, or otherwise adapted to a Likert scale, before generating permuted versions. For every question, five random permutations were created using fixed random seeds $(2, 7, 10, 67, 101)$. Combined with the original order, this results in six unique orderings per question. This dataset isolates the effect of option position while holding the question semantics and answer wording constant.

Final Dataset Composition. The data preparation process resulted in a final curated set of 1235 unique questions, organized into two experimental datasets for analysis. The Reworded Dataset comprises 6175 items (1235 questions $\times$ 5 wording variations), while the Shuffled Dataset comprises 7410 items (1235 questions$\times$6 order permutations). These comprehensive datasets enable a robust analysis of LLM consistency when answer options change in different ways.

4.2 Model Configurations

To examine the consistency of LLM responses across both architecture families and alignment regimes, this study evaluates four open-weight model families in two configurations each: a base (pretrained) model and its corresponding instruction-tuned variant. All models are medium-sized (7 – 9B parameters), which permits running multiple evaluation passes per item while remaining representative of contemporary, general-purpose LLM architectures.

All models were sourced from the Hugging Face Hub and inference was performed using the `transformers` library on the GPU cluster of the Leibniz Supercomputing Centre (LRZ). Models were loaded in their native precision (typically bfloat16 or float16) using the `dtype='auto'` setting, ensuring that the results reflect the exact model weights without performance degradation from quantization.

Selected Model Families. The study includes four distinct model families to ensure that observed consistency patterns are not specific to a certain training recipe or tokenizer. Table 2 summarizes the exact versions used.

Llama 3.1 (8B) represents a widely adopted decoder-only transformer family from Meta, known for strong performance on English-centric benchmarks (Grattafiori et al., 2024). **Mistral v0.3 (7B)** is a compact model family from Mistral AI optimized for efficiency, serving as a strong baseline in open-source evaluations (Jiang et al., 2023). **Qwen 2.5 (7B)**, developed by Alibaba Cloud, is a multilingual model family that demonstrates competitive performance relative to larger open-weight systems, particularly in reasoning and non-English tasks (Yang et al., 2024). Finally, **Gemma 2 (9B)** from Google is built on the same research and technology as the Gemini models, targeting resource-constrained deployment while retaining strong general capabilities (Gemma Team et al., 2024).

Table 2: Overview of the language models evaluated in this study. Each family is tested in both its base and instruction-tuned configuration.

Family	Specific Version	Params	Type
Llama 3.1	`Meta-Llama-3.1-8B`	8B	Base
	`Meta-Llama-3.1-8B-Instruct`	8B	IT
Mistral	`Mistral-7B-v0.3`	7B	Base
	`Mistral-7B-Instruct-v0.3`	7B	IT
Qwen 2.5	`Qwen2.5-7B`	7B	Base
	`Qwen2.5-7B-Instruct`	7B	IT
Gemma 2	`gemma-2-9b`	9B	Base
	`gemma-2-9b-it`	9B	IT

Base vs. Instruction-Tuned Configurations. For each family, the *Base* model represents the pretrained backbone trained primarily with self-supervised next-token prediction on large-scale text corpora. The corresponding *IT* variant has undergone further post-training, typically involving Supervised Fine-Tuning on instruction datasets and Reinforcement Learning from Human Feedback. Comparing these pairs isolates the effect of alignment training on multiple-choice consistency, holding the underlying architecture and parameter count constant.

Inference and Decoding Strategy. To ensure reproducibility and eliminate variance caused by stochastic sampling, all model outputs were generated using greedy decoding. In the generation configuration, sampling was disabled (`do_sample=False`), which is mathematically equivalent to setting the temperature to 0. This forces the model to deterministically select the token with the highest probability at each step. This approach is preferred for evaluation tasks to ensure that any observed variation in responses is attributable to the changes in the prompt (scale design, wording) rather than the randomness of the generation process. The maximum new token limit was set to 50, sufficient for generating short survey responses.

4.3 Prompting Strategies

The prompting strategy aims to reliably extract a single choice from a set of Likert-type options while minimizing bias, refusal (if refusal is not an answer option) and formatting errors. The strategy employs a zero-shot approach, presenting the task as a neutral survey. This framing aims to reduce the likelihood that the model interprets the task as requiring factual justification or moral reasoning, and discourages content moderation refusals on sensitive subjective items.

Instruction-Tuned vs. Base Model Implementation. Due to the architectural differences between the model types, two distinct prompting paradigms were employed:

1. **Instruction-Tuned Models:** These models were prompted using the standard chat template structure (List of Messages). All prompting content, including the

core task instructions (e.g., "You are taking a survey..."), was delivered within a single message with the role `user`. This approach deliberately omitted a separate `system` prompt to prevent potential inconsistencies in how different models parse such instructions.

2. **Base Models:** As these models are trained for text completion rather than instruction following, they were prompted using a sentence completion strategy. The instruction block was removed, and the prompt consisted solely of the Question, the Options, and a forcing suffix (e.g., "The best answer is option number:") to induce the model to generate the next token as the answer.

Prompt Templates and Constraints. All prompts were constructed to enforce a minimal output format, requesting a single token (number or letter) where possible to facilitate automated parsing.

Template A: Instruction-Tuned Models

> You are taking a survey. There are no right or wrong answers.
> Please choose the one option that best fits your view. The order of the options does not imply any ranking.
> Output format: Provide exactly one `label type` only (no words, no punctuation, no quotes).
>
> Question: `question`
>
> Options:
> `option block`
>
> Please answer with ONLY one of the answer options given above, and with only the `label type` corresponding to the option you choose.

Template B: Base Models

> Question: `question`
>
> Options:
> `option block`
>
> The best answer is option `suffix`:

Experimental Variations. To test the research questions, the `option block`, `label type` and `suffix` components were varied in four configurations.

- **Prompt 1: Numeric Labels.** Options are enumerated as "1) ... 2) ...".
  - `label type`: "number"
  - `suffix`: "number"
- **Prompt 2: Alphabetic Labels.** Options are enumerated as "A) ... B) ...".
  - `label type`: "letter"

  - `suffix`: "letter"

- **Prompt 3: Unlabeled Options.** Options are listed as plain text lines.
  - `label type`: "option"
  - `suffix`: Empty ("The best answer is option:")

- **Prompt 4: Additional Options.** To examine the impact of expanded option availability on response stability, the numeric scale (Prompt 1) was augmented with two additional options: "Don't know" and "Refused". For a scale of length N, these were labeled as $N+1$ and $N+2$. Since these options do not map onto the substantive semantic interval $[-1, 1]$ defined in Section 4.1, they are assigned distinct exclusion codes (-98 and -99) to separate them during analysis.

4.4 Consistency Metrics

To quantify the semantic stability of model responses across variations, the normalized numerical values (mapped to $[-1, 1]$ as defined in Section 4.1) were analyzed. The analysis focuses on measuring the divergence between responses to the same underlying question when the prompt or scale is altered.

Data Filtering and Validity. Prior to calculation, all responses are filtered to separate substantive opinions from invalid outputs. Responses assigned the exclusion codes -98 (Don't Know), -99 (Refused), or those resulting in parsing errors (e.g., if the model's response does not contain any valid answer option) are removed from the set of values.
To ensure a fair comparison and avoid artificially deflating variance through missing data, consistency metrics are computed only for questions where the model provided a valid substantive response for the full set of variations. Specifically, for the Reworded Dataset a question is included only if $n = 5$ valid responses exist, and for the Shuffled Dataset only if $n = 6$ valid responses exist. Questions failing this "completeness check" are excluded from the consistency aggregation.

Metric 1: Average Pairwise Distance (APD). The primary measure of inconsistency is the average absolute difference between all unique pairs of responses for a given question. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the set of valid normalized scores for a single question. The APD is calculated as:

$$APD = \frac{1}{K} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |v_i - v_j| \tag{2}$$

where $K = \frac{n(n-1)}{2}$ represents the total number of unique pairs (10 pairs for $n = 5$, 15 pairs for $n = 6$). An APD of 0 indicates perfect consistency, while higher values indicate semantic drift.

Metric 2: Maximum Pairwise Distance (MPD). To capture the worst-case divergence for a single question (e.g., a model switching from "Strongly Agree" to "Strongly Disagree"), the Maximum Pairwise Distance is calculated:

$$MPD = \max_{1 \le i < j \le n} |v_i - v_j| \tag{3}$$

This metric highlights extreme volatility, identifying cases where minor prompt variations cause the model to span a significant portion of the opinion spectrum.

Metric 3: Mean Standard Deviation (MSD). To quantify the stability of model responses at the question level, the standard deviation of answers is calculated for each unique question q. Let $V_q = \{v_1, \ldots, v_n\}$ be the set of valid numeric projections across the n variations ($n = 5$ for the Reworded dataset and $n = 6$ for the Shuffled dataset). The per-question standard deviation (σ_q) is defined as:

$$\sigma_q = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (v_i - \bar{v})^2} \tag{4}$$

To aggregate this metric for a specific model and prompt, the *Mean Standard Deviation* is used, calculated as the arithmetic mean of σ_q across all valid questions in the dataset. A lower MSD indicates higher consistency, implying the model settles on a specific numeric value regardless of variation.

Metric 4: Pearson Correlation Coefficient (r). To evaluate the structural robustness of a model's logic across different prompt formats (e.g., Numeric vs. Alphabetic), the Pearson correlation coefficient between the response vectors of distinct prompt pairs are calculated. Let $X = \{x_1, \ldots, x_N\}$ be the set of numeric responses for Prompt A across all N valid questions (where valid answers exist for both prompts), and $Y = \{y_1, \ldots, y_N\}$ be the corresponding responses for Prompt B. The correlation r is defined as:

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{5}$$

where $\bar{x}$ and $\bar{y}$ are the means of the response vectors. A high correlation ($r > 0.8$) indicates that the model preserves the relative semantic ordering of questions despite structural changes, implying high stability. Conversely, a low correlation ($r < 0.5$) suggests that the phrasing or option labels significantly alter the model's perception of the questions.

5 Results

This chapter presents the empirical findings regarding the consistency of Large Language Models in survey-style multiple-choice tasks. The analysis is structured to address the research questions defined in Section 1. First, global response patterns and validity rates are examined to establish a baseline. Subsequent sections detail the specific impact of semantic rewording, option ordering, and prompting strategies on response stability.

5.1 Global Response Patterns and Validity

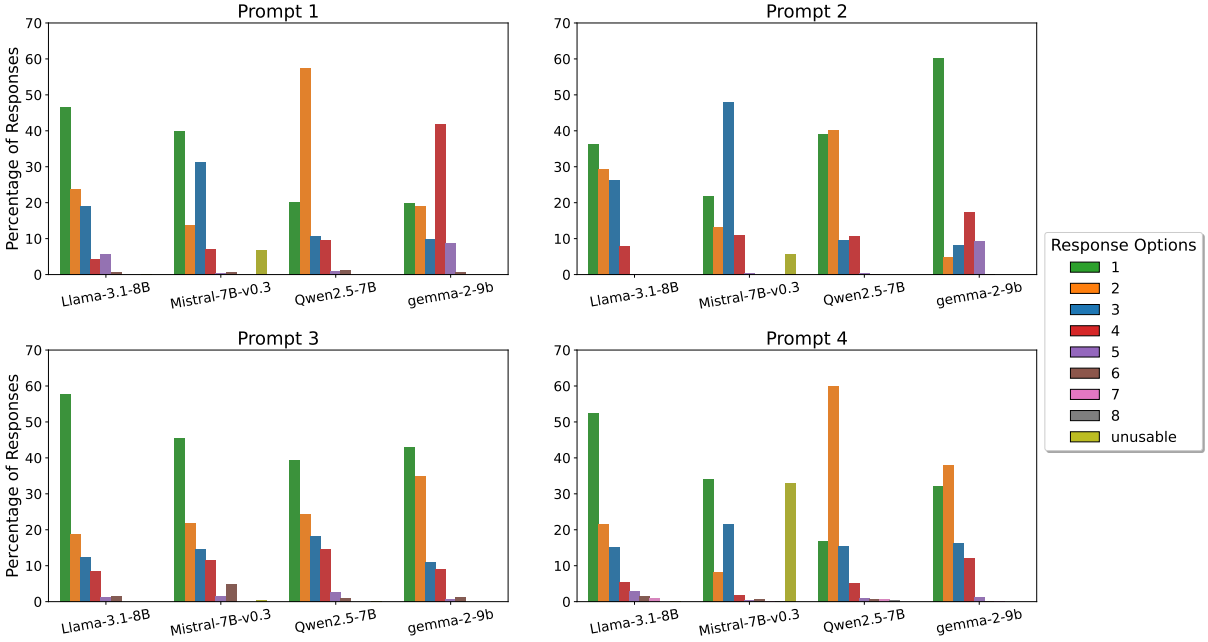
Before analyzing the consistency of model responses, this section first examines whether the models followed formatting instructions and assesses their overall answer tendencies.

5.1.1 Instruction Adherence and Response Validity

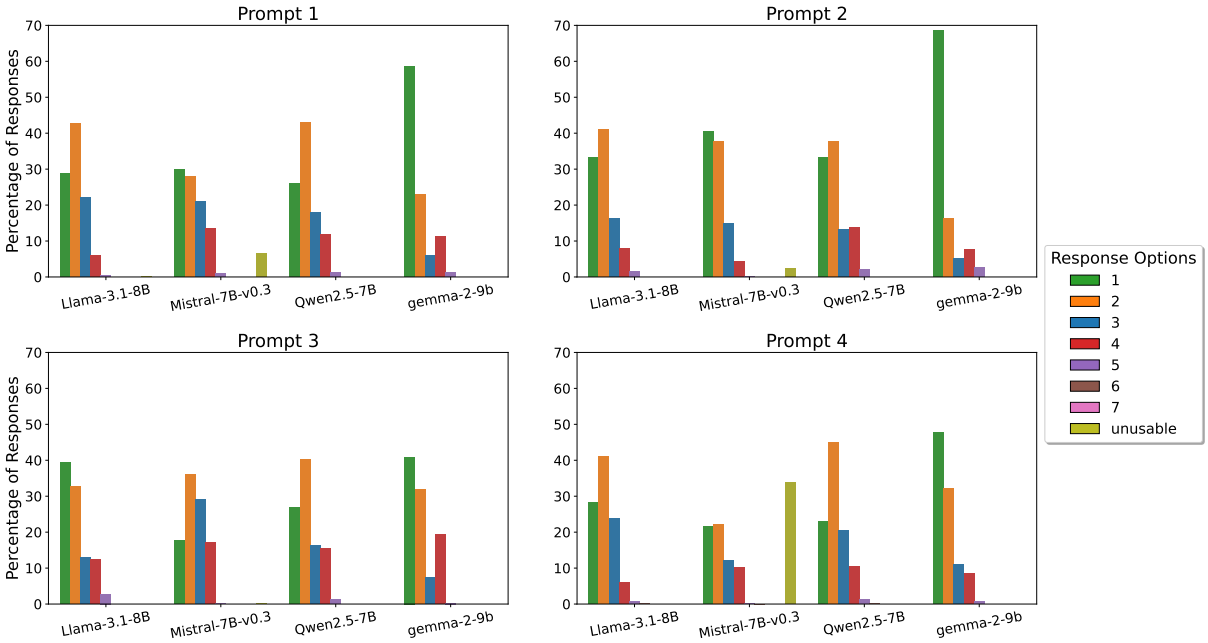
Figure 3 illustrates the distribution of raw response categories across all prompt variations for the base models.
To interpret the response categories, it is important to account for the varying scale lengths (a detailed overview of which is provided in Appendix A.2) and the mapping logic used in the experimental design:

- **Reworded Dataset (Figure 3a):** This dataset includes scale variations with up to 6 options (covering 4-, 5-, and 6-point scales). Consequently, for Prompts 1 – 3, valid responses range from 1 to 6. For Prompt 4, the valid range extends to 8 due to the addition of the "Don't Know" and "Refused" options.
- **Shuffled Dataset (Figure 3b):** This dataset was generated using the original survey wording or its closest Likert adaptation for the answer options. Since these primary scales had a maximum cardinality of 5 points, the valid response range for Prompts 1 – 3 is limited to indices 1 – 5. Accordingly, for Prompt 4, the range extends only to 7 (representing a maximum 5-point scale with two additional options). Crucially, responses in this dataset are mapped back to their original indices. This means that "Option 1" in the plot always refers to the original first option (e.g., "Strongly Agree"), regardless of which position it appeared in during the shuffled trial.

In both cases, the category "Unusable" aggregates all model outputs that failed the strict parsing criteria defined in Section 4.4 (e.g., outputs containing no valid answer options).



(a) Base Models on the Reworded Dataset



(b) Base Models on the Shuffled Dataset

Figure 3: Distribution of response categories for Base models. Subfigure (a) displays the distribution on the Reworded Dataset, while (b) shows the Shuffled Dataset. The Visualization for the Shuffled Dataset uses the original indexing (mapping answers back to their original logical ID), allowing for direct comparison of semantic preferences. Valid indices range from 1 – 8 (Reworded) and 1 – 7 (Shuffled).

Observations: Reworded Dataset (Figure 3a) The analysis of the base models on the Reworded Dataset reveals distinct and often erratic response distributions for each family.

Llama-3.1-8B demonstrates a consistent primacy bias across all four prompt variations. Option 1 is the most frequently selected answer, capturing between 35 % and up to 60 % of the total responses depending on the prompt. Subsequent options (2, 3, etc.) are selected with progressively lower frequency, creating a skewed distribution regardless of the prompting strategy.

Mistral-7B-v0.3 displays sensitivity to prompt formatting and stability issues. While Option 1 is the dominant choice under standard Numeric labels (Prompt 1) and Unlabeled (Prompt 3) formats, the model's preference shifts to Option 3 when Alphabetic labels are used (Prompt 2). Notably, the model struggles with the extended scale in Prompt 4 (Numeric + Don't Know/Refused), where the "Unusable" response rate spikes to nearly 35 %, rivaling the frequency of the most selected valid option (Option 1). Prompts 1 and 2 also exhibit elevated unusable rates ($\approx 8\,\%$) compared to the other models.

Qwen2.5-7B shows distinct preference shifts based on label presentation. Using the Numeric labels of Prompts 1 and 4, the model strongly favors Option 2 ($\approx 60\,\%$). However, in Prompt 2, the preference splits almost equally between Options 1 and 2 ($\approx 40\,\%$ each). In the Unlabeled Prompt 3, the distribution reverts to a standard primacy pattern with Option 1 being the most frequent ($\approx 40\,\%$), followed by Option 2.

Gemma-2-9B also exhibits shifts based on the prompts used. In Prompt 1, it uniquely favors Option 4 ($> 40\,\%$), a pattern not observed in any other model. However, this preference is unstable; in Prompt 2, the distribution shifts dramatically to a massive spike for Option 1 ($> 60\,\%$), while Prompt 3 aligns with the general primacy trend observed in the other models. In Prompt 4, the preference splits, with Option 2 ($\approx 40\,\%$) and Option 1 ($\approx 35\,\%$) being the most selected.

Observations: Shuffled Dataset (Figure 3b) The analysis of the Shuffled Dataset reveals that while shuffling disrupts simple positional preferences, it often exposes latent semantic biases or instability rather than producing balanced distributions.

Llama-3.1-8B maintains a specific preference, but the target shifts. In Prompts 1, 2, and 4, Option 2 emerges as the dominant choice (original index), though with a reduced margin (≈ 10 percentage points above Option 1) compared to the strong primacy observed in the Reworded data. This suggests that when options are shuffled, Llama-3.1 gravitates toward the semantic content of Option 2 (e.g., "Somewhat Agree"), whereas in the fixed order, it defaulted to the position of Option 1.

Mistral-7B-v0.3 continues to display prompt sensitivity and notable stability issues. The unusable response rate remains elevated in Prompt 1 ($\approx 8\,\%$) and spikes in Prompt 4 to $\approx 35\,\%$, where it becomes the highest frequency outcome, while valid options 1 and 2 trail at around 25 % each. Unlike in the Reworded data, Prompt 2 shows a more balanced distribution across Options 1 and 2, while Prompt 3 favors Option 2.

Qwen2.5-7B exhibits strong consistency between the Shuffled and Reworded datasets, particularly in the numeric Prompts 1 and 4. Since the Shuffled plot maps back to canonical indices, the persistence of the spike at Option 2 indicates that Qwen is actively reading and selecting the content of Option 2, regardless of its position. However, the magnitude of this preference is dampened compared to the Reworded data. Prompt 3

diverges most notably compared to the Reworded results, shifting to favor Option 2. This shift results in Qwen displaying the most uniform response patterns across prompts within the Shuffled dataset.
Gemma-2-9B demonstrates a clear and consistent preference for Option 1 across all shuffled prompts, contrasting with its more erratic behavior in the Reworded dataset. In Prompts 1 and 2, Option 1 spikes dramatically ($\approx 60\,\%$ and $\approx 70\,\%$, respectively). While less dominant, Option 1 remains the favorite in Prompt 3 ($\approx 40 - 45\,\%$) and Prompt 4 (just below $50\,\%$).

In contrast to the base models, the instruction-tuned variants demonstrate higher consistency in response distributions across both prompt types and datasets. Furthermore, the rate of unusable responses for these models is negligible. The detailed distributions are provided in Appendix A.3.

5.1.2 Aggregate Distributional Bias

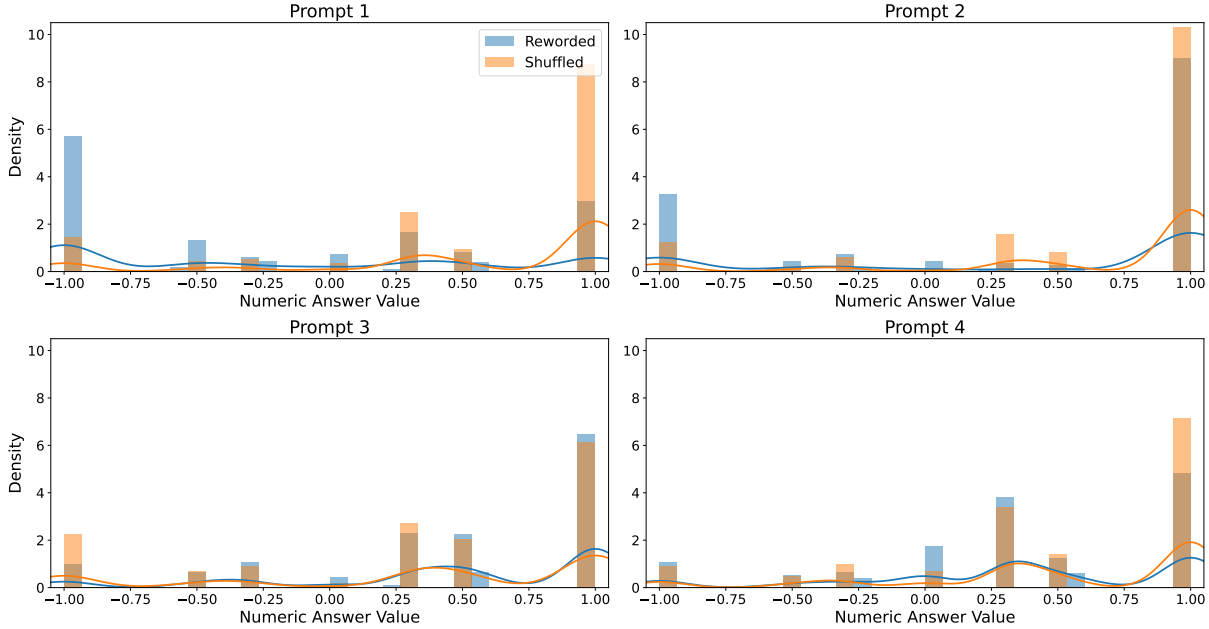
Having assessed the answer distribution and the models' abilities to provide valid responses, the next step examines their aggregate "opinion profile." Figure 4 compares the density of projected numerical responses for the Base and Instruction-Tuned versions of the Gemma model.
To interpret these distributions, three methodological details must be recalled from Section 4.1:

1. **Numerical Mapping:** All valid substantive answers were mapped to equidistant points on the interval $[-1, 1]$, where 1.0 represents maximum affirmation (e.g., "Strongly Agree", "A lot"), 0.0 represents neutrality, and -1.0 represents maximum negation (e.g., "Strongly Disagree", "None at all").

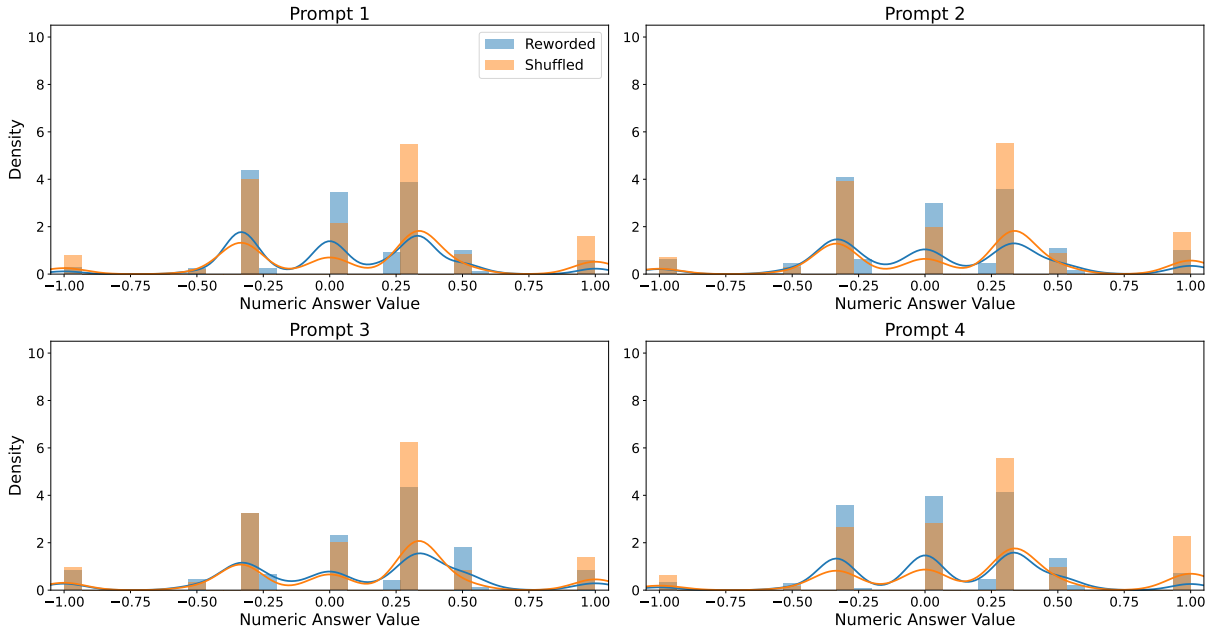
2. **Visual Encoding:** The plots display overlapping density histograms separated by prompt type ($1-4$). The blue distributions represent the Reworded Dataset, while the orange distributions represent the Shuffled Dataset.

3. **Discretization:** Since the source scales have limited cardinality (4, 5, or 6 options), the projected values are discrete rather than continuous. Consequently, the histograms display gaps, as density clusters exclusively around specific mapped values (e.g., $\pm 1, \pm 0.6, \pm 0.2$ for a 6-point scale) rather than spreading across the entire continuum.

This comparison serves as a diagnostic for semantic stability. If a model answers based on semantic content, its aggregate opinion distribution should remain roughly invariant between the Reworded and Shuffled datasets (i.e., the blue and orange areas should overlap).



(a) Base version of Gemma (Gemma-2-9B)



(b) Instruction Tuned version of Gemma (Gemma-2-9B-IT)

Figure 4: Comparative histograms of projected numeric answers for the Reworded (Blue) vs. Shuffled (Orange) datasets. Overlapping distributions indicate that the aggregate sentiment remains stable despite a changing option order.

Observations. The histograms in Figure 4 reveal a fundamental divergence in how Base and Instruction-Tuned models construct their aggregate "opinion profiles."
Gemma-2-9B (Base Model, Figure 4a) exhibits extreme polarity and instability. Across all prompts, the model consistently favors extreme values (± 1.0). The Shuffled distribution

generally skews more positive than the Reworded one. This pattern is most distinct in Prompt 1, where the distributions are inverted: the Reworded dataset (Blue) shows a high density spike at -1.0 (Strongly Negative), whereas the Shuffled dataset (Orange) spikes at $+1.0$ (Strongly Positive). This indicates that the model's opinion is influenced by option order; when the affirmative option moves positions, the model's output flips polarity rather than tracking the semantic content. Prompts 3 and 4 show a lower overall density in the negative extreme, with it mostly distributed in the positive part of the scale, peaking at 1.0. Both, and especially Prompt 3, show a higher degree of overlap between the two histograms. Crucially, the Shuffled distribution (Orange) does not simply flatten into a uniform shape (which would indicate random guessing). Instead, it often retains distinct spikes at the extremes. This indicates that Gemma-2-9B understands the options' meaning and systematically selects the extreme positive option when the order is randomized.
Gemma-2-9B-IT (Instruction-Tuned, Figure 4b) displays a very different and more stable behavior. First, the central tendency is reversed: unlike the Base model, the IT model assigns the highest density to neutral and moderate options (around 0.0 and ± 0.33). Second, the semantic stability is higher. The Reworded (Blue) and Shuffled (Orange) histograms show strong overlap across all prompts, particularly in Prompt 3, where the distributions are nearly identical. Minor deviations persist; for instance, the Shuffled dataset tends to show slightly higher density in the moderately positive region (≈ 0.25), while the Reworded dataset favors the neutral or moderately negative regions slightly more. However, these shifts are subtle magnitude differences compared to the complete polarity inversions observed in the Base model's responses. This indicates that instruction tuning has successfully grounded the model's outputs in the semantic content of the labels, making it largely robust to the reordering of options.

For a comparative analysis of the Qwen family, which exhibits a divergent and more stable pattern, refer to Appendix A.4. Detailed distribution plots for the remaining families (Llama and Mistral) are available in the electronic appendix (Section B).

5.2 Consistency Analysis (Intra-Prompt Robustness)

While the previous section analyzed aggregate behavior, this section evaluates "intra-question consistency". Specifically, the robustness of model responses is measured against variations in answer wording (Reworded Dataset) and option ordering (Shuffled Dataset), focusing on the results from one prompt at a time.

5.2.1 Pairwise Distance Analysis

To assess the magnitude of divergence between variations, the Average Pairwise Distance and Maximum Pairwise Distance are analyzed for each question. The APD provides a general measure of disagreement among the variations, while the MPD captures the "worst-case" scenario (e.g., whether a model ever flips from "Agree" to "Disagree" for the same question).
To ensure direct comparability between the datasets despite their differing sample sizes (6,175 for Reworded vs. 7,410 for Shuffled), the distributions are normalized as probability

densities. In the resulting histograms, the area of each bar represents the proportional frequency of observations (calculated as $\frac{\text{count}}{N \times \text{bin width}}$), ensuring that the total area under each curve sums to 1.

Note on Scale Artifacts. A structural constraint within the Reworded Dataset must be considered: the number of answer options can vary between reworded versions of the same question (ranging from 4 to 6 options). This leads to unavoidable differences in numeric mapping for intermediate options across the different scales:

- **4-point:** $[-1, -0.33, 0.33, 1]$
- **5-point:** $[-1, -0.5, 0.0, 0.5, 1]$
- **6-point:** $[-1, -0.6, -0.2, 0.2, 0.6, 1]$

Consequently, if a model consistently selects the equivalent intermediate option (e.g., the second highest), the pairwise distance will be non-zero purely due to scale artifacts. A distance of exactly 0 is thus mathematically impossible for intermediate options when comparing across differing scales.

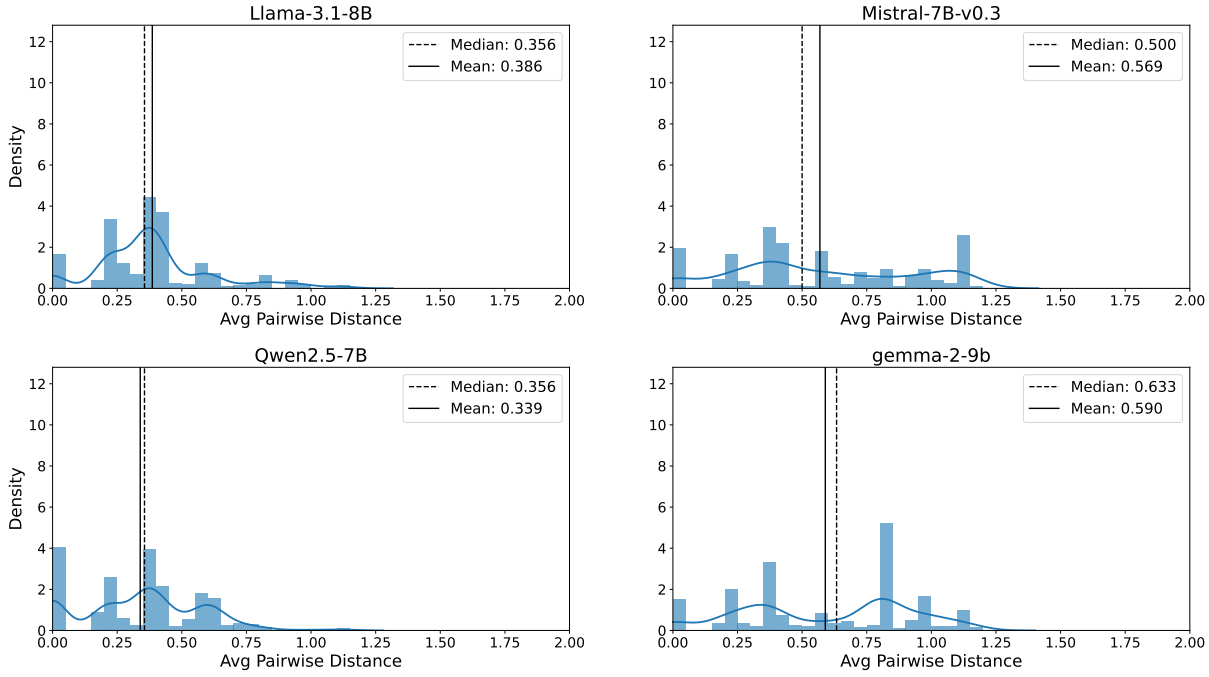


Figure 5: Distribution of Average Pairwise Distances for Base Models using Prompt 1 on the Shuffled Dataset. Lower values indicate tighter clustering of responses.

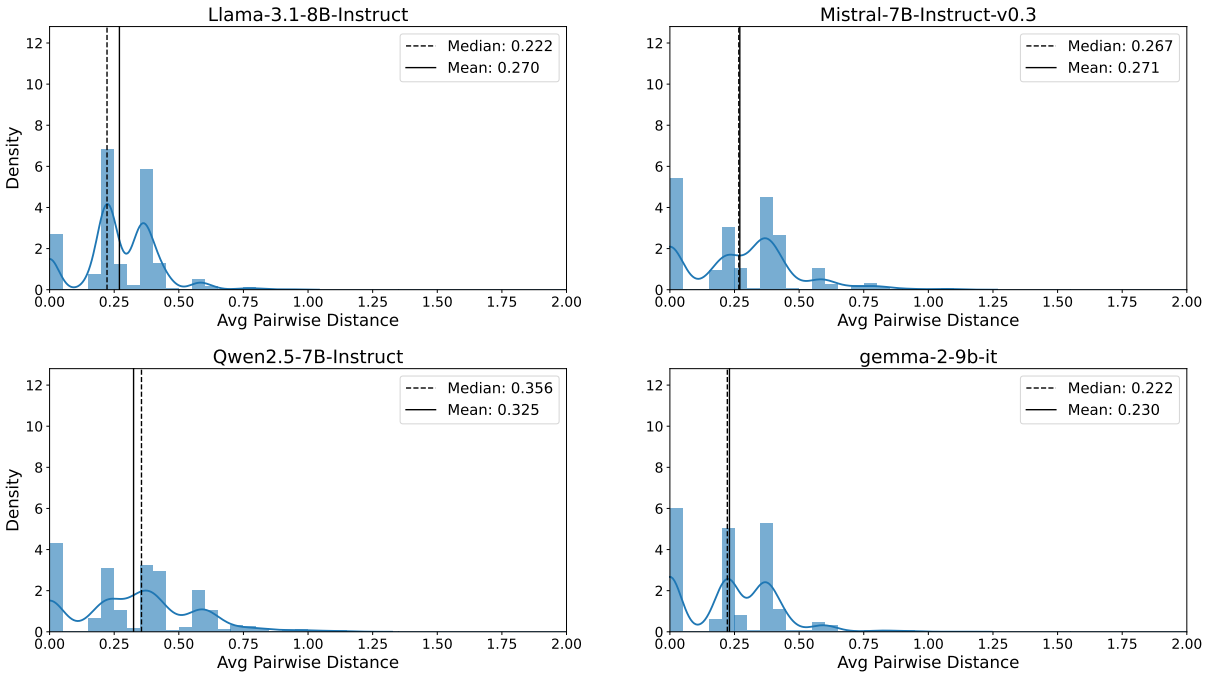


Figure 6: Distribution of Average Pairwise Distances for Instruction-Tuned Models using Prompt 1 on the Shuffled Dataset. Compared to the Base Models (Figure 5), the distributions are shifted closer toward zero.

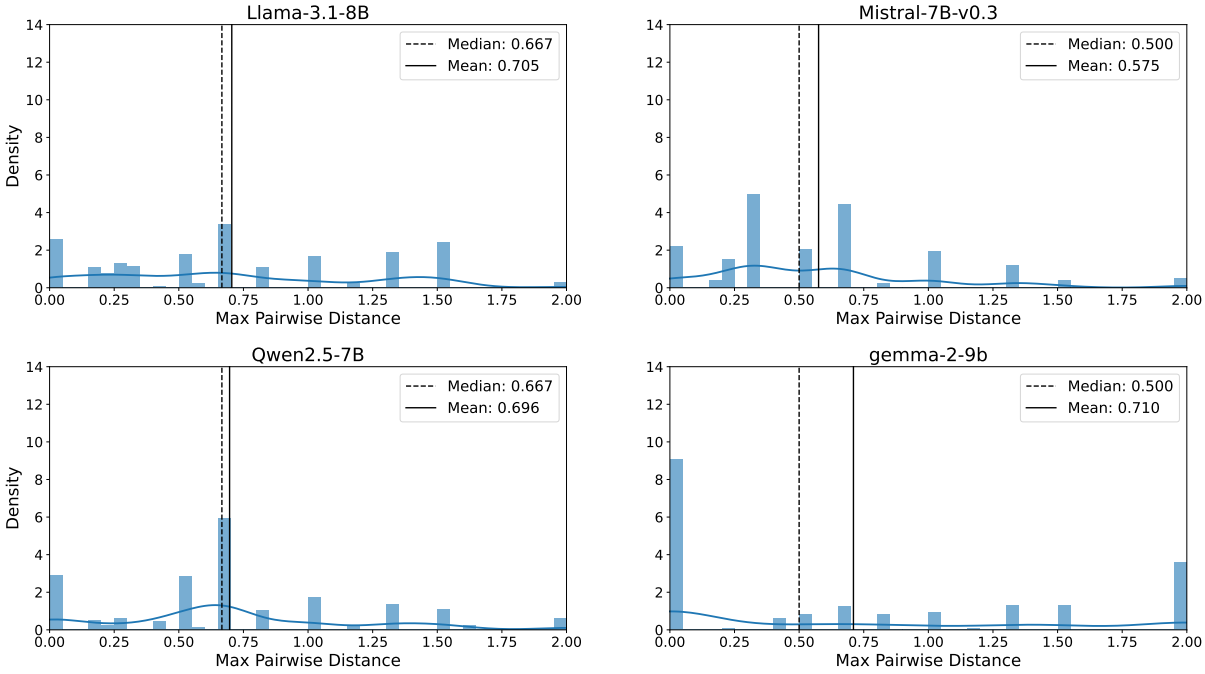


Figure 7: Distribution of Maximum Pairwise Distances for Base Models using Prompt 2 on the Reworded Dataset. A peak at 2.0 indicates full polarity inversions (flipping from -1 to $+1$) within a single question.

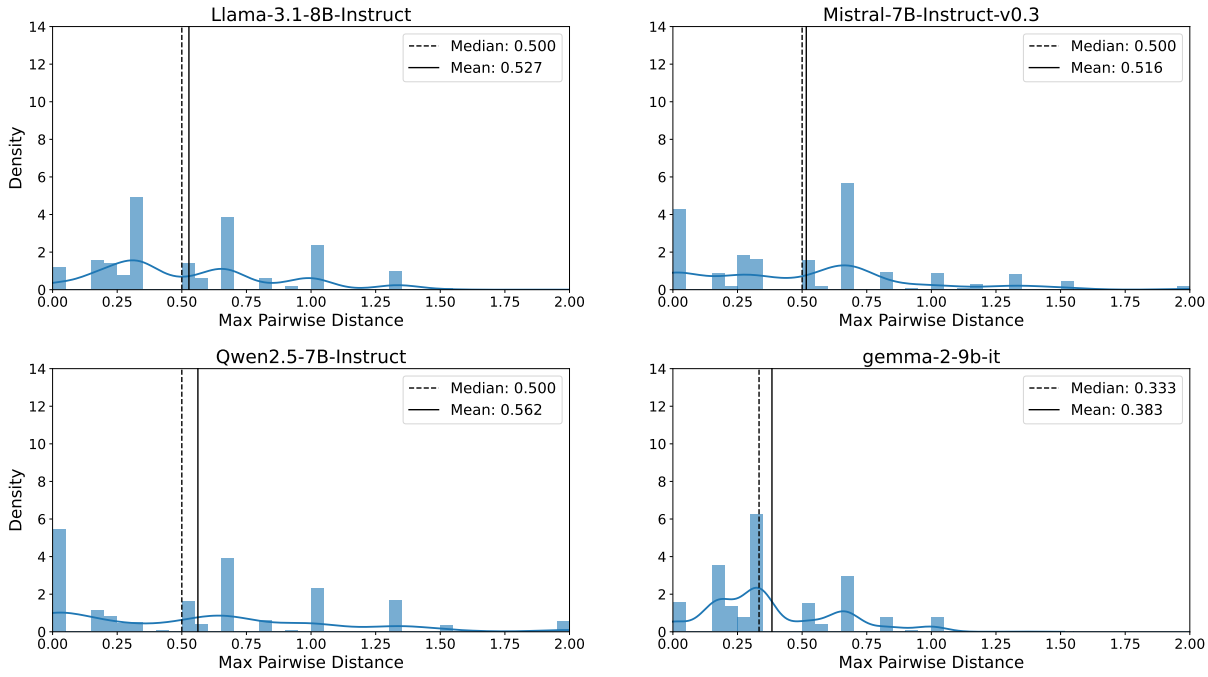


Figure 8: Distribution of Maximum Pairwise Distances for Instruction-Tuned Models using Prompt 2 on the Reworded Dataset. Note the reduction in high-volatility density compared to Figure 7.

Observations. The comparative analysis of Average Pairwise Distance and Maximum Pairwise Distance distributions, visualized in Figures 5 and 6 (APD) and Figures 7 and 8 (MPD), reveals profound differences in consistency between Base and Instruction-Tuned models. In these plots, the x-axis represents the pairwise distance (ranging from 0.0 to 2.0, where 0.0 implies perfect consistency and 2.0 signifies a complete semantic inversion), and the y-axis denotes the normalized density. A dashed line indicates the median, while a solid line marks the mean of the distribution.

Average Pairwise Distance (Figures 5 and 6). Both figures present the distributions of Average Pairwise Distances for the Shuffled Dataset (Prompt 1).
The *Base Models (Figure 5)* exhibit high variance and distinct instability patterns. *Mistral-7B-v0.3* displays a remarkably diffuse distribution, with density spread almost evenly across the range of 0 to 1.25, resulting in a high mean APD of ≈ 0.57. *Gemma-2-9B* proves to be the most inconsistent; its distribution is not only shifted to the right (Median ≈ 0.63) but the density also spikes at a high distance of ≈ 0.8. *Llama-3.1-8B* and *Qwen2.5-7B* show slightly better clustering around the lower end ($0.2-0.5$), though they still maintain significant tails extending toward higher distances.
The *Instruction-Tuned Models (Figure 6)* demonstrate a nearly uniform improvement in consistency. Across three of the four families, the distributions shift noticeably to the left compared to their base variants, concentrating density in the low-distance region ($0-0.5$) and tapering off sharply thereafter. The central tendencies reflect this stabilization: *Gemma-2-9B-IT*, which was the worst-performing Base model, achieves the lowest mean APD (≈ 0.23) alongside *Llama-3.1-8B-IT* (≈ 0.27). *Qwen2.5-7B-IT* stands as the ex-

ception, showing minimal improvement over its base variant, with the median remaining unchanged and the mean decreasing only slightly. This general convergence suggests that, for most architectures, instruction tuning effectively reduces sensitivity to option ordering, pulling the average disagreement down to a much narrower range.

Maximum Pairwise Distance (Figures 7 and 8). The Maximum Pairwise Distance analysis highlights the stability issues inherent in the Base models (Figure 7). Unlike the average distance, which smooths over outliers, the MPD reveals the extent of the "worst-case" divergence for each question.
The *Base Models* display distributions that span the entire possible range from 0 to 2. Instead of clustering near zero (which would indicate stability), the density is dispersed almost evenly across the axis for many models. *Gemma-2-9B* illustrates this instability most clearly: while it shows a density spike at 0 (indicating perfect stability for some questions), it also exhibits a distinct cluster at the maximum value of 2.0. A distance of 2.0 represents a complete polarity inversion (e.g., flipping from "Strongly Agree" to "Completely Disagree") triggered solely by rewording the answer options. The other models follow a similar pattern of high dispersion. *Llama-3.1-8B* and *Qwen2.5-7B* exhibit high central tendencies (Mean ≈ 0.70), while *Mistral-7B-v0.3* averages ≈ 0.58. The fact that the median MPD for most models is ≥ 0.5 indicates that for the majority of questions, at least one prompt variation causes a substantial shift in the interpreted sentiment.
The *Instruction-Tuned Models (Figure 8)* demonstrate a mostly clear improvement in stabilizing these worst-case divergences. Overall, the density shifts notably toward the lower end of the spectrum (0 to 1), with reduced mass in the high-volatility region (> 1.5). *Gemma-2-9B-IT* exhibits the most profound recovery: the chaos of the base model is replaced by a distribution that effectively zeroes out above $x = 1.0$. With a median MPD of 0.33 and a mean of 0.38, it completely eliminates polarity inversions. *Llama-3.1-8B-IT* similarly succeeds in truncating the tail, showing no density above $x \approx 1.3$, though its central tendency (Mean ≈ 0.53) remains higher than Gemma's. However, instruction tuning is not a universal solution for all architectures. Both *Mistral-7B-v0.3-IT* and *Qwen2.5-7B-IT*, despite showing lower overall means (≈ 0.52 and 0.56 respectively), still retain non-zero density at $x = 2.0$. Notably, *Mistral-7B-v0.3-IT* shows the lowest relative improvement compared to its base variant, indicating that while these models are generally more stable, they remain susceptible to occasional complete polarity inversions.

Comparison with Complementary Datasets. Finally, it is important to note the structural differences in the distributions between the two datasets. As detailed in Appendix A.5, results from the Shuffled Dataset typically exhibit distinct spikes and gaps due to the fixed number of answer options, whereas the Reworded Dataset produces more dispersed, smooth distributions due to variable scale lengths. Figures 14 and 15 in the Appendix show the complementary analysis on the opposing datasets, highlighting the distinct impact of option ordering on the distance metrics compared to variations in wording and scale.

5.2.2 Response Volatility

Complementing the pairwise distance metrics, this section evaluates the *Mean Standard Deviation* to quantify the overall dispersion of model responses. While APD measures

the average disagreement between pairs, the MSD captures how tightly the entire set of responses for a given question clusters around a central value.

Volatility in Reworded Scenarios. Table 3 summarizes the aggregated volatility scores for the Reworded Dataset. Lower values indicate higher stability and a stronger resistance to changes in answer wording.

Table 3: Mean Standard Deviation on the Reworded Dataset. The table compares Base and Instruction-Tuned models across the four prompt variations. Lower values indicate that the model consistently assigns the same numerical "opinion" despite changes in answer wording.

	Mean Standard Deviation			
Model	**P1 (Numeric)**	**P2 (Alpha)**	**P3 (Unlabeled)**	**P4 (Ext.)**
Base Models				
Llama-3.1-8B	0.357	0.315	0.411	0.331
Mistral-7B-v0.3	0.299	0.259	0.454	0.275
Qwen2.5-7B	0.254	0.304	0.416	0.221
Gemma-2-9B	0.344	0.325	0.421	0.229
Instruction-Tuned Models				
Llama-3.1-8B-Instruct	0.207	0.230	0.248	0.203
Mistral-7B-Instruct-v0.3	0.221	0.227	0.287	0.199
Qwen2.5-7B-Instruct	0.232	0.249	0.244	0.235
Gemma-2-9B-IT	0.174	0.171	0.179	0.148

Observations from Table 3. A stark contrast between model classes is demonstrated. Base models exhibit an overall volatility roughly twice that of their instruction-tuned counterparts (e.g., Llama Base ≈ 0.35 vs. IT ≈ 0.21). The Base models are particularly prone to high variance in Prompt 3 (Unlabeled), where values spike to > 0.40 (e.g., Mistral-Base at 0.454). This suggests that without the "anchor" of explicit labels (A, B, C), Base models struggle to map semantic content to a consistent numerical score.
Among the tuned models, *Gemma-2-9B-IT* emerges as the most consistent performer, achieving the lowest MSD across all prompts (≈ 0.17). This finding is particularly notable given that its Base variant was among the most unstable. This confirms that Gemma's stability is not an inherent architectural feature but a direct result of effective instruction tuning and RLHF.

Volatility in Shuffled Scenarios. Figure 9 expands this analysis to the Shuffled Dataset, visualizing how option reordering impacts internal consistency.

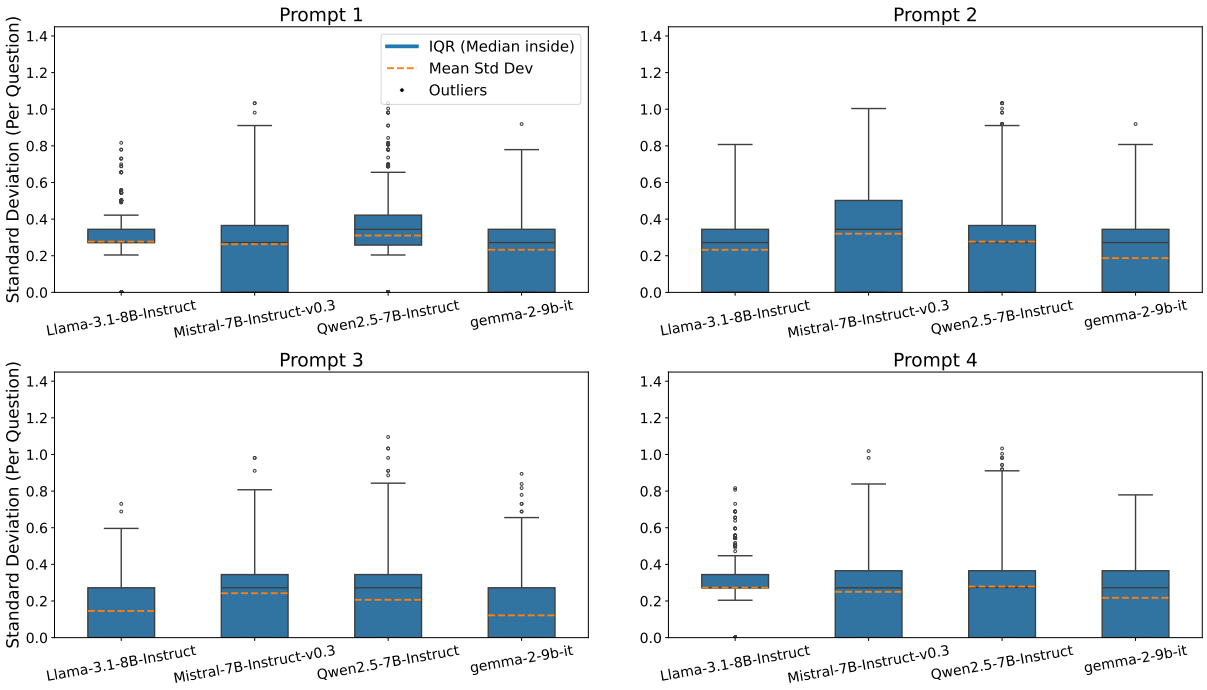


Figure 9: Distribution of response volatility (Standard Deviation) for Instruction-Tuned models on the Shuffled Dataset.

Observations from Figure 9. Figure 9 is divided into four subplots, one for each prompt. In each subplot, the box represents the Interquartile Range (IQR) containing the middle 50 % of the data, with the median marked by a solid line inside the box. The orange dashed line indicates the mean standard deviation, while points beyond the whiskers represent outliers (questions where the model exhibited unusually high volatility).
While the mean values broadly align with the ranges observed in Table 3, the boxplots reveal the consistency of the volatility itself. A narrow box indicates that the model's instability is uniform across questions, implying it answers with a consistent level of variation regardless of the specific question. Conversely, a wider box suggests that the model's stability is highly dependent on the question content; this means that some questions trigger high volatility while others are answered robustly.
Llama-3.1-8B-Instruct exhibits very thin boxes for Prompts 1 and 4 (IQR range ≈ 0.25 to 0.35). This suggests a highly consistent behavior across the dataset: the model answers with slight variation consistently across all questions. Prompts 2 and 3 show wider distributions but lower means ($\approx$ 0.23 and 0.15 respectively), indicating less uniform behavior.
Mistral-7B-Instruct-v0.3 displays similar distributions for Prompts 1, 3, and 4 (IQR 0 to 0.4, mean ≈ 0.3), but shows a higher volatility in Prompt 2, where the distribution stretches to 0.5 with a mean of ≈ 0.37. This suggests that the Alphabetic labels in Prompt 2 introduce a specific sensitivity, causing the model to react more inconsistently than it does to numeric or unlabeled options.
Qwen2.5-7B-Instruct generally maintains stable distributions (mean ≈ 0.3) across Prompts 2, 3, and 4, but Prompt 1 is noticeably tighter and higher, clustered between 0.3 and 0.42.

This indicates that in Prompt 1, Qwen answers consistently with higher variation than in the other prompts, where it answers with less volatility on average but exhibits greater fluctuation across different questions.
Gemma-2-9B-IT shows consistent distributions for Prompts 1, 2, and 4 (means ≈ 0.2), but stands out in Prompt 3 with a significantly compressed distribution (IQR 0 to 0.3) and the lowest mean and median of all models (≈ 0.12). Notably, *Gemma*, *Llama*, and *Qwen* all demonstrate a lower MSD in Prompt 3 compared to other prompts, whereas *Mistral* shows an increase in MSD for Prompt 3 relative to its performance on other prompts.

The "Prompt 3 Paradox". A counter-intuitive pattern emerges when comparing the Reworded results (Table 3) with the Shuffled results (Figure 9 and Appendix A.6), specifically regarding the Unlabeled Prompt 3. In the Reworded dataset, Prompt 3 caused the highest instability for Base and IT models, as they lacked the structural guidance of labels. However, for Instruction-Tuned models on the Shuffled dataset, Prompt 3 often yields the lowest volatility (e.g., Gemma-IT drops to a mean of $\approx$ 0.12). This suggests that in shuffled contexts, explicit labels (Prompts 1, 2, 4) may actually introduce a conflict, where the model is torn between a positional bias (preferring "A") and a semantic preference (preferring "Agree"). In Prompt 3, the absence of labels forces the model to rely solely on semantic content, resulting in higher consistency despite the randomization.

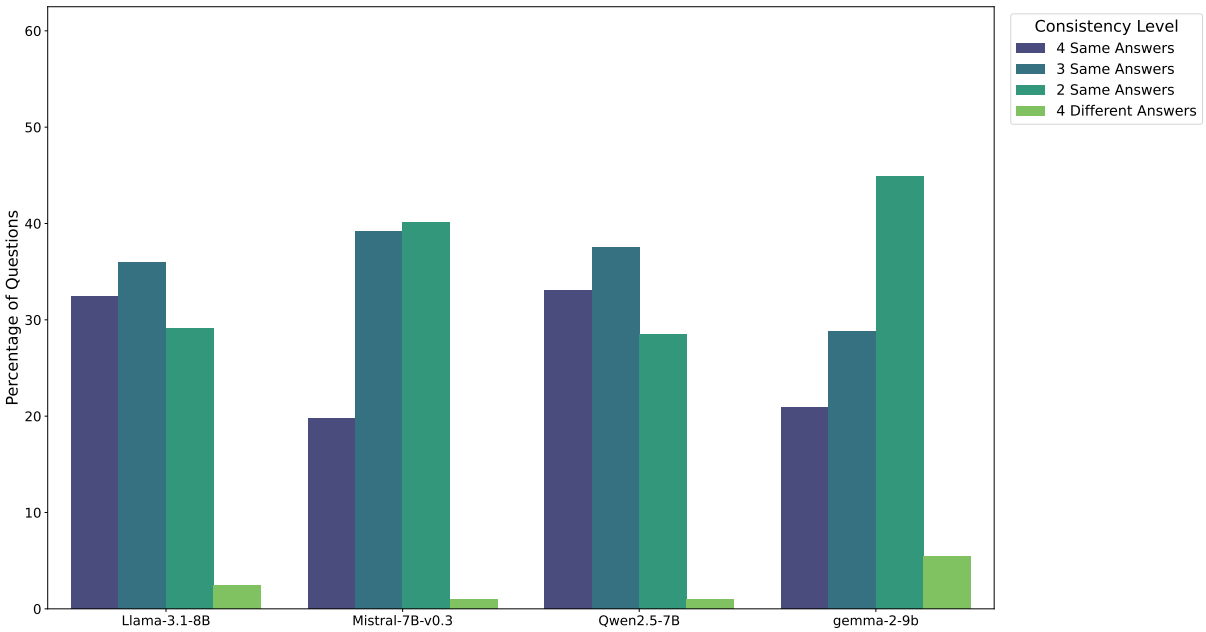
5.3 Structural Robustness across Prompt Variations

While the previous sections analyzed consistency within a single prompt structure (intra-prompt consistency), this section evaluates the inter-prompt stability. Specifically, it investigates whether models maintain a consistent logical stance when the structural format of the question changes, such as switching from numeric labels (Prompt 1) to alphabetic labels (Prompt 2), or removing labels entirely (Prompt 3).
This analysis is critical for determining whether a model's "opinion" is grounded in the semantic content of the statement or if it is an artifact of the specific multiple-choice format used.

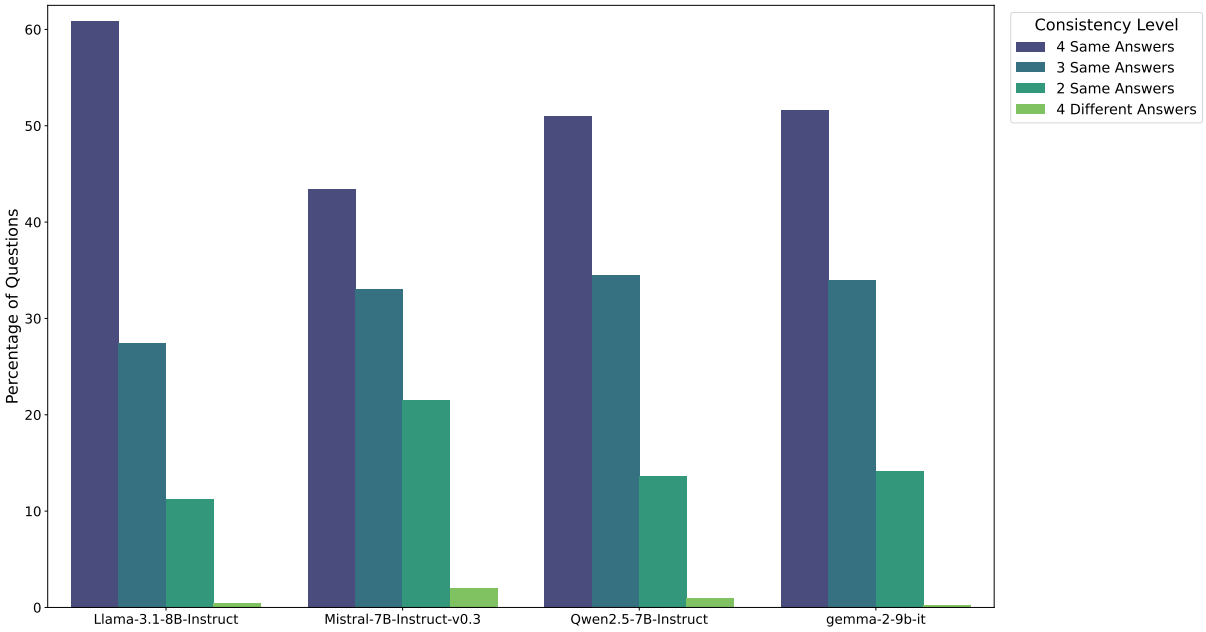
5.3.1 Categorical Consistency

First, strict categorical agreement is assessed across all four prompt variations on the Reworded Dataset. For a given question, this analysis determines whether the model maintains the exact same categorical stance (e.g., "Agree") regardless of how the question is presented.
The analysis holds the question text and the answer options constant, isolating the prompt format as the only variable. Two important methodological details apply to this analysis:

- **Data Exclusion:** Any question where the model produced an unusable response in one or more prompt variations is excluded from this comparison to ensure a valid cross-prompt baseline.
- **Extended Scale (Prompt 4):** Since Prompt 4 introduces two additional options that do not exist in Prompts 1 – 3, selecting one of these options inherently prevents a "4 Same Answers" outcome.



(a) Base Models Stability across Prompts



(b) Instruction-Tuned Models Stability across Prompts

Figure 10: Categorical consistency of model responses across four different prompt variations on the Reworded Dataset. The bars indicate the percentage of questions where the model provided the identical categorical answer (e.g., "Agree") across all prompt formats.

Observations. Figures 10a and 10b display the distribution of categorical consistency levels for Base and Instruction-Tuned models, respectively. The x-axis represents the model families, while the y-axis denotes the percentage of questions falling into each consistency category. The bars are color-coded to indicate the level of stability across all

Prompts.

Base Models (Figure 10a). The Base models exhibit a pattern of logical fragmentation, where perfect consistency is never the dominant outcome. *Llama-3.1-8B* and *Qwen2.5-7B* show similar profiles: the most frequent outcome is "3 Same Answers" ($\approx 35\,\%$), followed closely by "4 Same" ($\approx 32\,\%$) and "2 Same" ($\approx 27\,\%$). This indicates that while these models often maintain their stance across the majority of prompts, the change in format frequently causes at least one deviation.
Mistral-7B-v0.3 struggles significantly with cross-prompt stability, with "4 Same Answers" remaining below $20\,\%$. The bulk of its responses are split between "2 Same" and "3 Same" ($\approx 40\,\%$ each). It is important to note, however, that Mistral Base produced a high rate of unusable answers in Prompt 4 ($\approx 35\,\%$, see Section 5.1.1), reducing the effective sample size for this specific comparison.
Gemma-2-9B displays the lowest stability of the group. Its dominant category is "2 Same Answers" ($\approx 45\,\%$), and it exhibits the highest rate of total disagreement ("4 Different") at $\approx 5\,\%$. This confirms that without instruction tuning, Gemma is highly sensitive to the structural presentation of the question.

Instruction-Tuned Models (Figure 10b). Instruction tuning induces a major shift toward structural robustness. Across all models, the dominant category becomes "4 Same Answers," ranging from $43\,\%$ to $62\,\%$. *Llama-3.1-8B-Instruct* achieves the highest stability, with $\approx 62\,\%$ of questions receiving the identical categorical response across all four prompts. The "4 Different" category is negligible ($< 1\,\%$).
Qwen2.5-7B-Instruct and *Gemma-2-9B-IT* perform similarly, with perfect consistency rates of $\approx 51\,\%$. *Gemma-2-9B-IT* represents the most dramatic improvement relative to its base variant, shifting from a "2 Same" dominant profile to a "4 Same" dominant one, with "4 Different" dropping to $< 1\,\%$.
Mistral-7B-Instruct-v0.3 is the least stable of the tuned models, with "4 Same" at $\approx 43\,\%$ and a comparatively higher rate of "2 Same" ($\approx 23\,\%$). Overall, the high prevalence of "3 Same Answers" (typically $\approx 27-34\,\%$) across all IT models likely reflects the structural difference of Prompt 4; since Prompt 4 offers two options not available in Prompts $1-3$, a model selecting "Don't know" or "Refused" in Prompt 4 would inherently fall into the "3 Same Answers" category despite maintaining a logically consistent stance.

5.3.2 Correlational Stability

To capture more subtle relationships between model outputs, a Pearson correlation analysis was conducted. Unlike the binary consistency check (which asks if the answer remained identical), Pearson's r measures the linear relationship between the numeric scores of different prompts, quantifying how strongly a model's response in one format predicts its response in another.
For each specific question and answer-option permutation, the numeric value of the answer given in Prompt A is paired with the value given in Prompt B. To ensure valid comparisons, pairwise deletion is used; any pair where at least one response is invalid or represents a non-answer (e.g., "Don't Know" (-98), "Refused" (-99) or "unusable") is excluded. A high correlation ($r \approx 1.0$) indicates that the model preserves the relative logical ranking of questions across formats, even if the absolute values shift. Conversely, a low

correlation implies that the change in prompt format fundamentally alters the model's perception of the question, scrambling the logical ordering of answers.
Table 4 presents the correlation matrix for the Shuffled Dataset. It is important to note that while this dataset introduces task complexity via randomized answer options, the correlation metric strictly compares the same question and answer-option combination across different prompts. Thus, it tests the model's ability to maintain consistent logic across prompt formats (e.g., Numeric vs. Alphabetic) within a randomized context. The corresponding results for the Reworded Dataset can be found in Appendix A.8.

Table 4: Pearson Correlation Coefficients (r) on the Shuffled Dataset. Comparisons involving Prompt 3 generally show lower correlations, while the pair $1-4$ is the most stable.

	Prompt Pair Correlations					
Model	**1 – 2**	**1 – 3**	**1 – 4**	**2 – 3**	**2 – 4**	**3 – 4**
Base Models						
Llama-3.1-8B	0.545	0.354	0.594	0.441	0.511	0.413
Mistral-7B-v0.3	0.309	0.183	0.496	0.191	0.269	0.244
Qwen2.5-7B	0.608	0.428	0.849	0.431	0.572	0.400
Gemma-2-9B	0.468	0.155	0.499	0.312	0.557	0.378
Instruction-Tuned Models						
Llama-3.1-8B-Instruct	0.797	0.636	0.796	0.686	0.778	0.644
Mistral-7B-Instruct-v0.3	0.730	0.709	0.901	0.627	0.728	0.718
Qwen2.5-7B-Instruct	0.886	0.783	0.921	0.777	0.882	0.780
Gemma-2-9B-IT	0.849	0.733	0.856	0.784	0.824	0.717

Observations. The correlation analysis reveals a fundamental difference in how models process prompt variations. While Instruction-Tuned models maintain high stability ($r \approx 0.70 - 0.90$) regardless of format, Base models exhibit significantly lower correlations ($r \approx 0.30 - 0.60$). This suggests that for Base models, changes in option labels or the addition of new options disrupts the model's reasoning, leading to inconsistent outputs.
Structural Anchors and Outliers. Two consistent structural trends emerge in Table 4. First, the pair $1-4$ consistently yields the highest correlations (often > 0.90 for IT models and up to 0.85 for Qwen-Base). This confirms that models recognize the structural similarity between Prompt 1 (Standard Numeric) and Prompt 4 (Extended Numeric), treating them nearly identically despite the difference in scale. Second, comparisons involving Prompt 3 ($1-3$, $2-3$, $3-4$) consistently show the lowest correlations. This indicates that Prompt 3, which lacks explicit option labels, is the structural outlier. Without labels acting as "anchors", models are more likely to select a different option compared to when option labels are present, especially in the Shuffled context where the option order is not fixed.
Instruction-Tuned Models: Format Robustness. Qwen2.5-7B-Instruct demonstrates the highest Robustness, achieving the highest correlations across almost all prompt pairs. It is particularly robust in the $1-4$ comparison ($r = 0.921$), indicating that its internal

logic is almost perfectly preserved even when new options are added. While *Gemma-2-9B-IT* showed the highest stability in the Reworded dataset (see Appendix A.8), it falls slightly behind Qwen in the Shuffled context. This suggests that while its linguistic understanding is superior, its logical robustness to format changes is slightly weaker when the options' positions are randomized. *Llama-3.1-8B-Instruct* generally performs well but shows noticeable drops in pairs involving Prompt 3, reinforcing the finding that the label-free format is the most challenging for maintaining cross-prompt consistency.
Base Models: Fragility and Exceptions. Qwen2.5-7B stands as the distinct outlier among Base models. With correlations approaching those of IT models (e.g., 0.849 in $1-4$), it is the most robust base model in this context, likely due to a substantial volume of logic or multiple-choice examples in its pre-training data that allows it to generalize across formats even without instruction tuning. In contrast, *Gemma-2-9B* illustrates extreme fragility. While its IT variant performs very well, the Base model's logic collapses when the prompt format changes in a shuffled context. The correlation for pair $1-3$ drops to a negligible 0.155. This highlights that Gemma's sophisticated language understanding is heavily dependent on instruction tuning; without it, the base model is extremely sensitive to structural variations, failing to recognize that Prompt 1 and Prompt 3 ask the same question. *Mistral-7B-v0.3* displays the lowest overall stability, effectively "guessing" with correlations frequently dropping below 0.20, indicating a very low logical transfer between different prompt formats.

Comparison with Reworded Dataset. Finally, comparing these results to the Reworded Dataset (see Appendix A.8) confirms that logical consistency is harder for the models to maintain in a shuffled context. Correlations are universally higher in the Reworded dataset, indicating that models can more easily identify semantic equivalence when the option order remains fixed. This drop in stability is most severe for Base models (with the exception of Qwen), which show distinct declines in cross-prompt consistency when moving from the Reworded to the Shuffled task, exposing a reliance on positional heuristics.

6 Discussion

The empirical results presented in Chapter 5 highlight a significant divide in the reliability of LLMs for synthetic survey research. While IT models demonstrate a capability for consistent semantic reasoning, Base models remain driven by surface-level patterns and formatting cues rather than genuine semantic stability. These findings are interpreted in this chapter through the lens of model architecture and training paradigms, before critically examining the limitations of the current study and proposing avenues for future research.

6.1 Interpretation of Results

6.1.1 The RLHF Hypothesis: Alignment as a Reliability Filter

The superior stability of IT models across all metrics most likely stems from a fundamental shift in training objectives. Base models optimize for next-token prediction, treating surveys as pattern completion tasks where arbitrary features, like option order or label format, alter the output probabilities. This aligns with findings by Zhao et al. (2021), who demonstrated that raw language models exhibit severe biases toward specific answers based on their frequency in pre-training data and their position in the prompt (recency bias).

In contrast, the full alignment pipeline (see Section 2.1) aligns the model to maximize helpfulness and faithfulness (Ouyang et al., 2022). This process acts as a reliability filter, training the model to abstract away from surface-level formatting and focus on semantic intent. Consequently, IT models converge on a logically "best" answer, rather than just the most likely continuation. However, this alignment is not without artifacts. Sharma et al. (2024) warn that RLHF can introduce "sycophancy," causing models to bias responses toward perceived user expectations rather than objective truth. Thus, while IT models are more mechanically consistent, their stability may partially reflect a form of "learned compliance" alongside robust reasoning.

6.1.2 The Supporting Role of Option Labels

The consistency collapse of Base models in Prompt 3 (Unlabeled) underscores the mechanical function of option labels. In the other prompts, labels likely serve as attention "anchors", supporting the mapping of semantic probabilities to discrete outputs. When removed, the model must rely solely on the semantic text of the answers, a task where Base models struggle to maintain a stable mapping. This observation is supported by Xue et al. (2024), who identify this specific failure mode as a lack of "Symbol Binding" capability in standard training. The stability of IT models suggests that instruction tuning reduces this structural dependency, strengthening the model's ability to attend directly to semantic content.

6.2 Limitations and Future Work

While this study provides a robust baseline for measuring LLM consistency, several limitations in scope and methodology highlight important directions for future research.

6.2.1 Critique of Consistency Metrics

Measuring consistency in LLMs inherently depends on the reliability of the answer extraction pipeline. As Molfese et al. (2025) demonstrate, standard evaluation strategies often fail to distinguish between reasoning failures and formatting failures, leading to "Right Answer, Wrong Score" scenarios. In this study, the reliance on strict output formatting imposed definitions of consistency that may have introduced specific biases:

The Additional Options. In the analysis of structural robustness, selecting "Don't Know" or "Refused" in Prompt 4 was categorized as an inconsistency. A more nuanced analysis could exclude these non-substantive responses to calculate a "pure" consistency score. As noted by Kadavath et al. (2022), the ability of a model to identify its own knowledge boundaries is a desirable alignment feature; thus, shifting to "Don't Know" may represent a valid expression of uncertainty rather than a failure of stability.

Handling of Unusable Responses. The current intra-prompt stability metrics excluded unusable answers. This is particularly relevant for *Mistral-7B-v0.3*, which exhibited a high rate of unusable responses. By excluding these failures, the analysis focused only on the "surviving" well-formed answers, potentially artificially inflating the perceived stability of the model. Future studies should investigate how stability counts shift if unusable outputs are treated as distinct, valid mismatch categories (e.g., "Agree", "Unusable", "Unusable", "Disagree" = "4 Different answers").

6.2.2 Scope of Variations

The perturbations in this study were limited to answer options and prompt structure. Future work should expand the scope of variation to verify robustness more thoroughly:

Question Rewording. In this study, the question text was held constant. Given the observed sensitivity to changes in answer option wording, it is probable that models are also sensitive to the phrasing of the question itself. Investigating whether LLMs maintain their stance when the query is rephrased (e.g., "Do you support X?" vs. "Are you in favor of X?") is a critical next step, as semantic inconsistencies under paraphrasing remain a known issue even for large models (Elazar et al., 2021).

Scale Nuances. The study utilized 4-, 5-, and 6-point scales. Psychometric literature suggests that scale properties (e.g., the presence of a neutral point, unipolar vs. bipolar phrasing) significantly affect human respondents (Menold, 2021). While this study focused on mechanical consistency of LLMs, future work should isolate these factors to determine if LLMs exhibit similar biases regarding scale design. Initial work by Rupprecht et al. (2025) has already begun to explore this, finding that while prompt perturbations trigger consistent recency bias, the effects on central tendency (e.g., preference for middle options) remain mixed.

Temperature and Decoding. All responses were generated using greedy decoding (temperature = 0). While this isolates the model's most probable path, it does not reflect the diversity of outputs generated in higher-temperature applications. Investigating how consistency degrades as temperature increases would provide bounds for creative applications of survey simulation.

6.2.3 Granularity of Analysis

Due to time constraints, this analysis aggregated results across all questions. However, the dataset includes manual classifications of question topics (e.g., "Politics", "Economy", "Race and Ethnicity"). A stratified analysis could reveal whether models are more robust on certain topics, while being more fragile on others.

6.2.4 The Lack of Human Baselines

Finally, this study evaluated LLMs against a theoretical standard of perfect logical consistency. However, human respondents are historically inconsistent due to fatigue, acquiescence bias, or genuine ambivalence (Krosnick, 1991). Without a comparative human baseline answering the exact same set of reworded and shuffled questions, it is difficult to determine where current models stand. Aher et al. (2023) suggest that LLM samples often exhibit "hyper-accuracy" and lower variance than human samples; thus, models might be "super-human" (lacking natural variance) or, conversely, exhibit volatility driven by mechanical artifacts that has no parallel in human psychology.

7 Conclusion

This thesis investigated the reliability and consistency of LLMs when deployed as simulated respondents in survey research. By subjecting four major model families (Llama, Mistral, Qwen, and Gemma) to rigorous changes in answer options and their presentation, including answer rewording, option shuffling, and prompt reformatting, this study isolated the mechanical artifacts of text generation from genuine semantic stability. The empirical findings presented in Chapter 5 allow for definitive conclusions regarding the feasibility of using LLMs for synthetic survey research.

7.1 Summary of Findings

The overarching finding of this research is that Instruction Tuning is the primary determinant of survey reliability. While model architecture and size play a role, the alignment process (SFT + RLHF, see Section 2.1) fundamentally alters how models process multiple-choice tasks, shifting them from probabilistic text completers to semantically grounded respondents.

7.1.1 The Stabilization Effect of Instruction Tuning

The comparison between Base and Instruction-Tuned models revealed a stark divergence in consistency. Base models demonstrated high volatility and susceptibility to "choice architecture" effects. For instance, *Llama-3.1-8B* exhibited a strong primacy bias, defaulting to the first option, while *Gemma-2-9B* displayed extreme polarity instability, flipping from one end of the Likert scale to the other (e.g., "Fully agree" to "Fully disagree") based solely on option order. This confirms the findings of Zheng et al. (2024), who identify selection bias as an intrinsic weakness of base models. Crucially, the direction of this bias appears unstable. While this study observed primacy effects, Rupprecht et al. (2025) demonstrate that such biases are highly sensitive to prompt syntax, further highlighting the structural fragility of unaligned models.
In contrast, Instruction-Tuned models reduced response volatility by approximately 50 % (as measured by MSD). They effectively mitigated positional biases, maintaining consistent semantic profiles even when answer options were shuffled or reworded. This confirms that the "opinions" extracted from base models are often artifacts of next-token prediction probabilities, whereas IT models are capable of maintaining a more coherent logical stance.

7.1.2 Robustness to Variation: Language, Order, and Structure

The analysis disaggregated consistency into three distinct dimensions:

Language Stability (Rewording). While generally less disruptive than shuffling, reworded variations exposed a distinct capability gap. Base models exhibited higher volatility when the formulation changed. However, Instruction-Tuned models remained highly stable, with *Gemma-2-9B-IT* demonstrating that instruction tuning effectively grounds the model in semantic meaning, allowing it to treat synonymous labels (e.g., "Fully agree" vs. "Strongly agree") as logically equivalent.

Order Stability (Shuffling). This dimension proved more challenging for the models than rewording. Base models frequently collapsed, with *Gemma-2-9B* yielding values that indicate random guessing. Instruction-Tuned models were far more robust, often reducing volatility by half compared to their base counterparts. Notably, a divergence emerged regarding the Unlabeled Prompt 3. In the Reworded dataset, the absence of labels generally hurt performance. However, in the Shuffled dataset, a "paradox" occurred for IT models: Prompt 3 yielded the highest consistency. This phenomenon is best explained by the theory of "Multiple Choice Symbol Binding" (MCSB) proposed by Xue et al. (2024). Models often fail not at selecting the content, but at binding that content to an abstract symbol (e.g., mapping "Agree" to "A"). By removing the labels, Prompt 3 allows IT models to bypass this error-prone binding step and choose based purely on semantic content.

Structural Stability (Prompt Formats). The study identified noticeable sensitivity to the presence of option labels. Prompt 3 (Unlabeled) acted as a structural outlier; without the labels, cross-prompt consistency dropped for all models. Conversely, the high correlation between Prompt 1 (Numeric labels) and Prompt 4 (Numeric labels & extended scale) indicates that models stuck to their choice in this context, rarely selecting the newly introduced "Don't Know" and "Refused" options.

7.2 Implications for Synthetic Survey Research

These findings have direct practical implications for researchers utilizing LLMs to simulate human populations or generate synthetic data:

Base models are unsuitable for direct surveying. Despite their capabilities in other domains, Base models lack the requisite stability for multiple-choice tasks. Their responses are too heavily influenced by arbitrary formatting choices to be treated as valid proxies for opinion. Future research should rely on Instruction-Tuned variants to ensure that variability in the data reflects simulated human diversity rather than mechanical noise (Santurkar et al., 2023).

Prompt design acts as a reliability filter. The results suggest that "minimalist" prompting (e.g., removing labels as in Prompt 3) generally degrades reliability. Numeric or alphabetic labels appear to help the model map its internal probability distribution to a discrete selection, resulting in more consistent answers. The notable exception is for IT models in randomized option ordering; in this specific context, the absence of labels actually increases consistency. This suggests that while labels generally stabilize responses, they introduce a specific vulnerability to "symbol binding" errors (Xue et al., 2024) when the option order is perturbed, creating a trade-off between structural guidance and binding accuracy.

Verification requires perturbation. A single query is insufficient to gauge an LLM's true "opinion." The volatility observed predominantly in Base models suggests that researchers should employ ensemble prompting (querying the model multiple times with shuffled options or reworded answers) to average out stochastic noise. As Khan et al. (2025) argue, without such robustness checks, what appears to be the model's opinion

can easily change with some variations in presentation, making it indistinguishable from random noise.

Relevance to Open-Ended Tasks. While this study focused on rigid multiple-choice formats to rigorously quantify consistency, LLMs are increasingly utilized for answering open-ended questions. The instability quantified here serves as a critical baseline: if a model cannot consistently maintain a stance in a structured environment, its open-ended justifications may be similarly fragile or "unfaithful" to the model's true internal state (Turpin et al., 2023). Thus, robustness checks are equally vital for free-text generation tasks.

7.3 Final Remarks

In conclusion, while Large Language Models exhibit promising capabilities for social science simulation, their reliability is not inherent. It is a product of specific fine-tuning processes and careful prompt engineering. By understanding the mechanical sensitivities of these models (to order, wording, and structure) researchers can better separate the signal of synthetic opinion from the noise of stochastic generation.

A Appendix

A.1 Use of AI Tools

In accordance with academic integrity guidelines, this section outlines the use of Generative AI tools in the preparation of this thesis. Large Language Models (specifically GPT-4o and GPT-5) were utilized for the following purposes:

- **Code Generation and Documentation:** Assistance in writing, debugging, and documenting Python scripts. This encompassed the development of the experimental pipeline used to prompt the models, as well as scripts for downstream data processing (pandas) and visualization (matplotlib/seaborn).
- **Text Refinement:** Improving the clarity, grammar, and flow of written sections. This included converting draft notes into academic prose and generating LaTeX code for tables and figures.

All experimental design choices, data interpretations, theoretical arguments, and final conclusions were determined solely by the author. All content generated or improved by AI tools was critically reviewed and independently verified. Final responsibility for all content lies with the author.

A.2 Distribution of Scale Types

To better understand the structural properties of the test sets, Table 5 details the distribution of "Scale Types", defined here as the combination of the answer scale's cardinality (number of options) and its polarity (Unipolar vs. Bipolar), across both datasets.
The Reworded Dataset (Table 5a) contains a wider variety of scale lengths (including 6-point scales) by experimental design; the generation of five distinct reworded variations per question was explicitly used to test model performance across different cardinalities and polarities. In contrast, the Shuffled Dataset (Table 5b) strictly adheres to the original answer options provided in the source opinion surveys, resulting in a higher concentration of 4-point scales and the absence of 6-point variations.

Table 5: Distribution of Answer Scale Types across the two experimental datasets.

(a) Reworded Dataset ($N = 6,175$)

Scale Type	Count	Pct.
4-point Unipolar	1,858	30.1 %
4-point Bipolar	1,760	28.5 %
5-point Bipolar	1,375	22.3 %
5-point Unipolar	635	10.3 %
6-point Bipolar	512	8.3 %
6-point Unipolar	35	0.6 %

(b) Shuffled Dataset ($N = 7,410$)

Scale Type	Count	Pct.
4-point Unipolar	3,408	46.0 %
4-point Bipolar	2,424	32.7 %
5-point Bipolar	1,350	18.2 %
5-point Unipolar	228	3.1 %

Note: These counts represent the baseline Likert options present in Prompts 1, 2, and 3. Prompt 4 (Extended) adds two additional options ("Refused" and "Don't Know") to every question, effectively increasing the cardinality of all scales shown above by +2 for that specific prompt variation.

A.3 Additional Answer Distribution Plots

This section presents the answer percentage distributions for the Instruction-Tuned models. These figures complement the analysis of Base models presented in Section 5.1.1 (Figure 3).
Figures 11 and 12 illustrate the response patterns for IT models across the Reworded and Shuffled datasets, respectively. In contrast to the Base models, which exhibited high volatility and mechanical biases (e.g., primacy bias or polarity flipping), the IT models demonstrate remarkable stability.

Semantic Stability. The most notable observation is the much higher similarity of distributions across the two datasets. While Base models frequently shifted their preference based on the presentation method, IT models maintained consistent response profiles. For the Shuffled Dataset (Figure 12), where answers are mapped back to their original semantic index (e.g., "Option 1" always represents "Fully Agree"), the distributions closely mirror those of the Reworded Dataset (Figure 11). This indicates that IT models successfully attend to the semantic meaning of the options rather than their position or phrasing.

Response Validity. Furthermore, the rate of "Unusable" responses is negligible for IT models. Unlike *Mistral-7B-v0.3* (Base), which failed in Prompts 1, 2, and 4, the IT variants produced valid outputs across all prompt types. The only exception was *Mistral-7B-Instruct-v0.3*, which produced a negligible number of unusable responses ($< 0.1\,\%$) in Prompt 3 for both datasets.

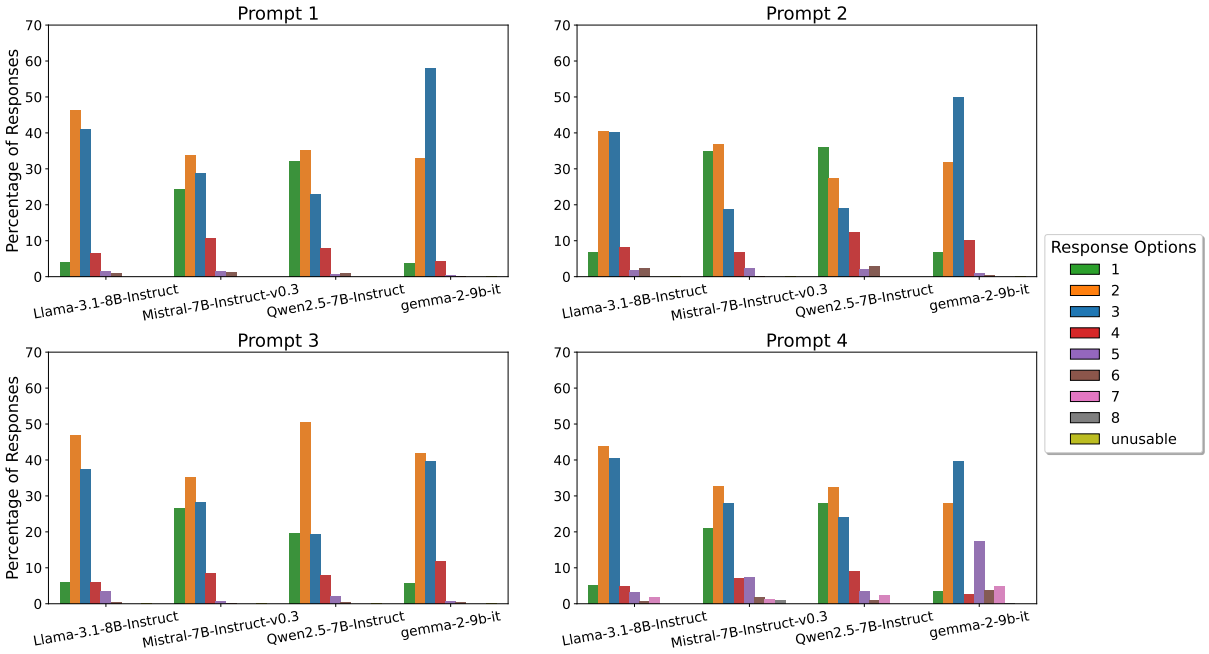


Figure 11: Distribution of response categories for Instruction-Tuned Models on the Reworded Dataset

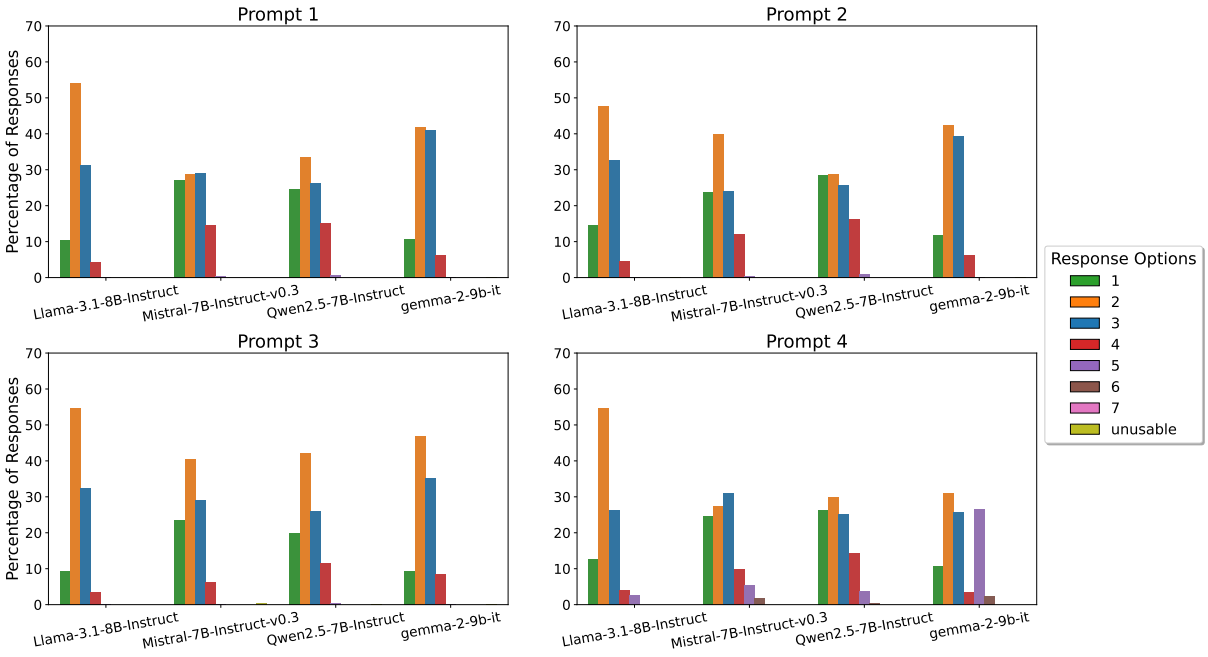


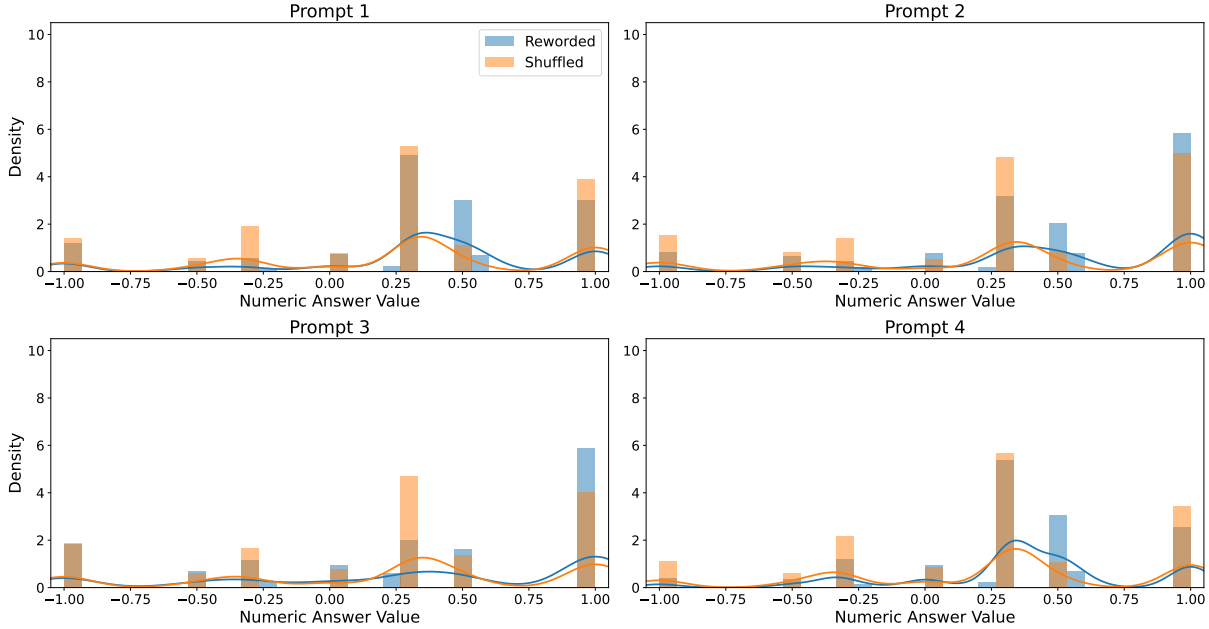
Figure 12: Distribution of response categories for Instruction-Tuned Models on the Shuffled Dataset. As in Figure 3, the indexing uses the original semantic order (Option 1 = "Fully Agree"), allowing for direct comparison with Figure 11.

A.4 Aggregate Distributional Bias (Qwen)

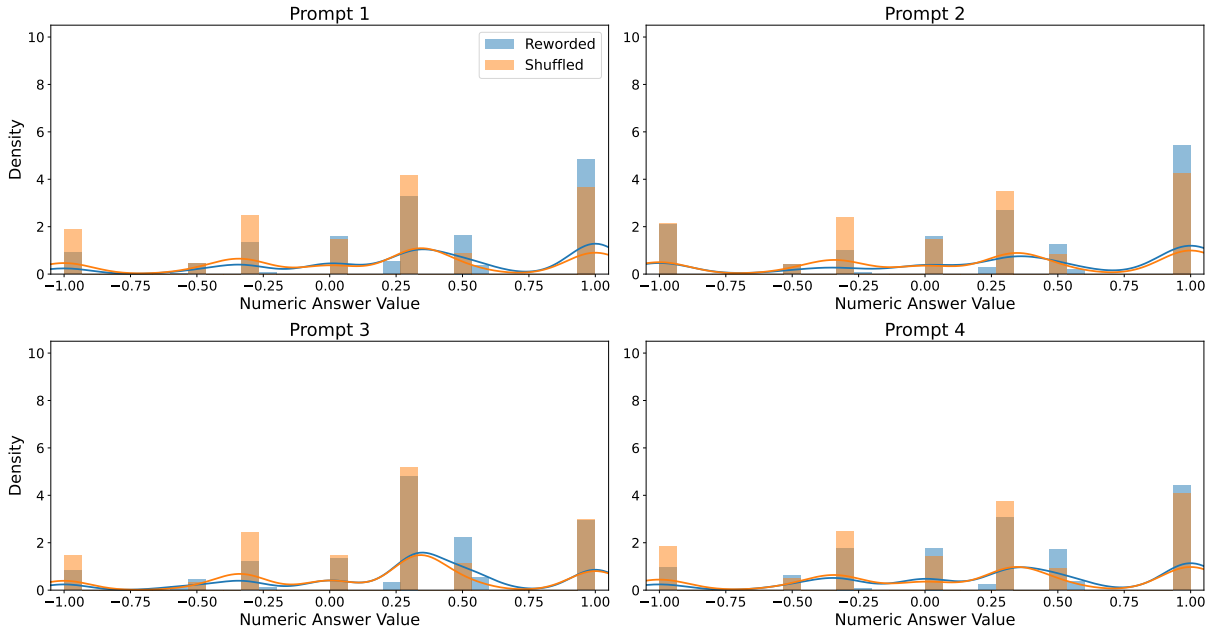
This section extends the aggregate distributional analysis presented in Section 5.1.2. While the main text focused on *Gemma*, which exhibited the most extreme divergence between Base and IT variants, Figure 13 illustrates the behavior of the *Qwen* family. In contrast to the extreme polarity instability observed in Gemma, the Qwen family demonstrates a much higher degree of intrinsic stability, even in its Base variant.

Base vs. IT Similarity: Unlike Gemma, where the Base model favored extremes (± 1) and the IT model favored neutrality, Qwen shows a remarkably consistent distributional profile across both versions. Both *Qwen2.5-7B* (Figure 13a) and *Qwen2.5-7B-Instruct* (Figure 13b) distribute density relatively evenly across the scale, with a slight preference for the positive spectrum.

Comparison to Other Families: Qualitative analysis of the remaining models (for visualizations see the GitHub repository linked in Section B) reveals that *Mistral* follows a pattern similar to Qwen, exhibiting relative stability in both Base and IT forms, though the IT variant shows a slightly higher density in the neutral region. *Llama*, on the other hand, shows similarities to the Gemma pattern. The Base model exhibits high density at the extreme answers (though exclusively at the positive end $+1.0$, unlike Gemma's bipolarity), while the IT variant successfully redistributes this density toward the center of the scale.



(a) Base version of Qwen (Qwen2.5-7B)



(b) Instruction Tuned version of Qwen (Qwen2.5-7B-Instruct)

Figure 13: Comparative histograms of projected numeric answers for the Qwen family. Note the high similarity between the Base (a) and Instruction-Tuned (b) distributions, contrasting sharply with the divergence observed in the Gemma family (Figure 4).

A.5 Complementary Pairwise Distance Analysis

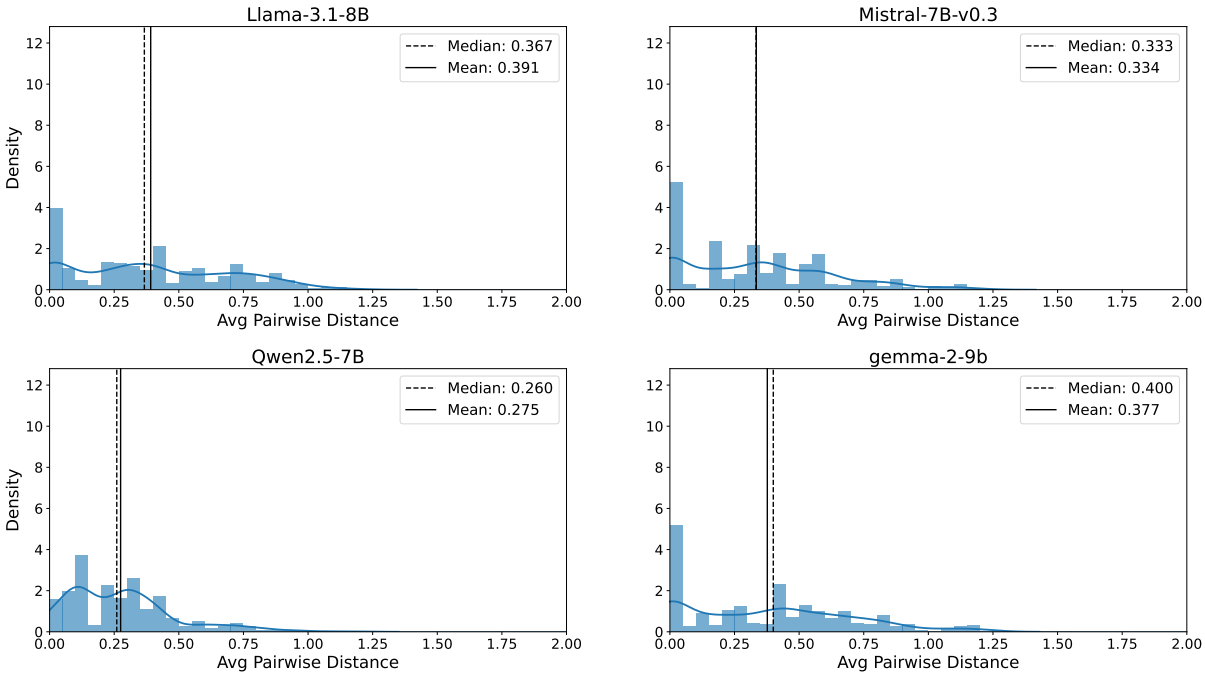
This section complements the pairwise distance analysis in Section 5.2.1. While the main text focused on APD for the Shuffled Dataset and MPD for the Reworded Dataset, the

inverse configurations are presented here.

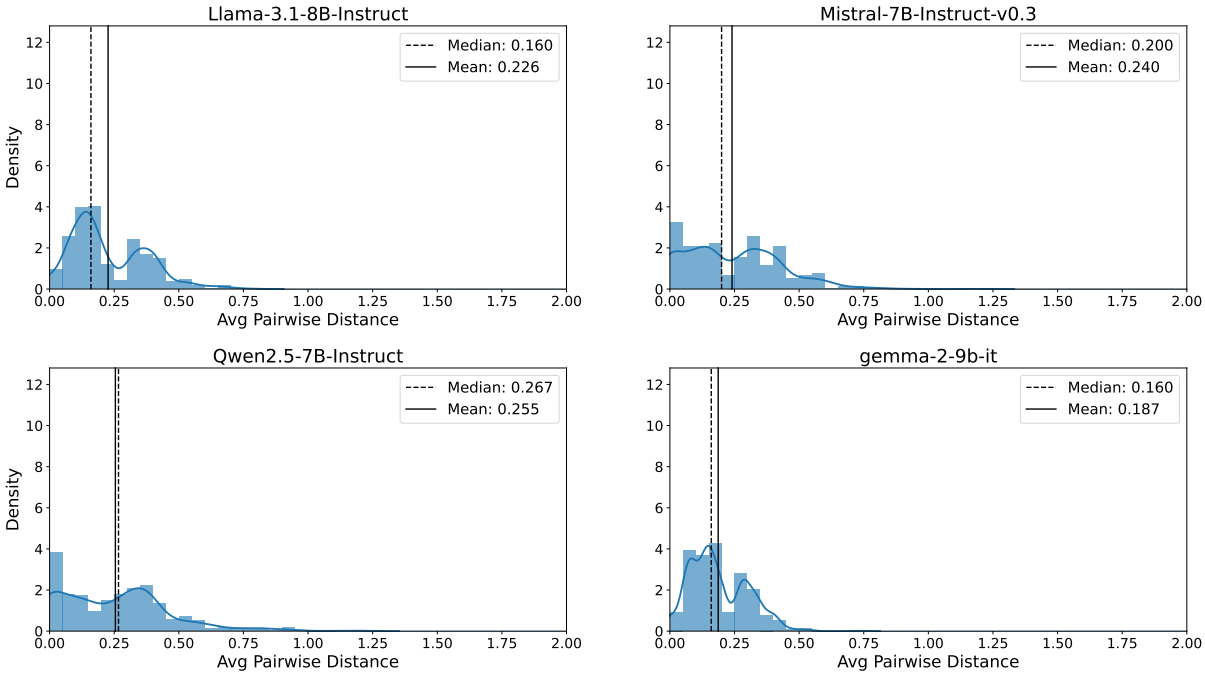
Structural Artifacts. A distinct visual difference is observable between the two datasets. The distributions for the Shuffled Dataset (Figures 15a and 15b) exhibit distinct gaps and high peaks ("spikes"). This occurs because the Shuffled dataset always uses the same number of answer options, limiting the possible distances between answers. In contrast, the Reworded Dataset (Figures 14a and 14b) produces smoother, more spread-out distributions, as the varying scale lengths (4, 5, or 6 options) generate a wider array of possible intermediate distance values.

APD on Reworded Dataset (Figure 14). Comparing these results to the Shuffled analysis in the main text reveals that rewording is generally less disruptive than shuffling for Base models. In Prompt 1, all Base models exhibit improved consistency (lower APD) on the Reworded dataset compared to the Shuffled dataset. *Mistral-7B-v0.3* and *Gemma-2-9B* show the most significant recovery, with mean APDs decreasing by ≈ 0.2, while *Llama* and *Qwen* show modest improvements. Qualitative analysis of the remaining prompts (available in the electronic appendix B) confirms this trend for Prompts 1, 2, and 4. However, Prompt 3 (Unlabeled) acts as an outlier: for Base models, it yields the worst consistency on the Reworded dataset, exhibiting higher means and distinct density shifts toward higher distances.
For IT models, the trend is similar but subtler. They exhibit slightly lower APDs on the Reworded dataset (≈ 0.05 decrease in mean), confirming high stability. The only exception is the "Prompt 3 Paradox" noted in the main text: IT models achieve their highest consistency on the Shuffled dataset for Prompt 3, outperforming the Reworded results shown here.

MPD on Shuffled Dataset (Figure 15). This analysis captures the "worst-case" impact of option ordering. For Base models, shuffling triggers more extreme polarity flips than rewording. The distributions shift toward higher distances, indicating that a mere change in order causes models to flip from one end of the scale to the other more frequently. IT models effectively mitigate these extremes, shifting density back toward the lower end. However, consistent with the base trend, the MPD values are generally higher here than in the Reworded dataset, confirming that ordering effects remain the more challenging semantic variation to overcome. Again, Prompt 3 remains the exception, where IT models (particularly Llama and Gemma) show remarkably low MPD values on the Shuffled dataset compared to the Reworded one.

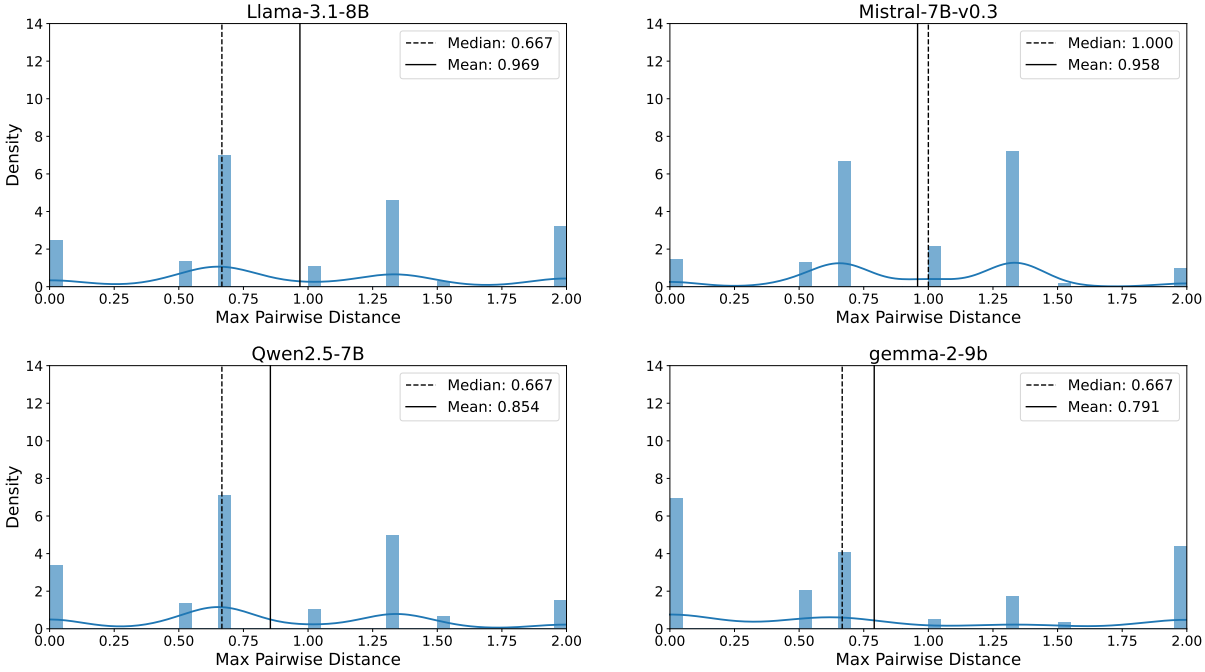


(a) Base Models (APD on Reworded Dataset, Prompt 1)

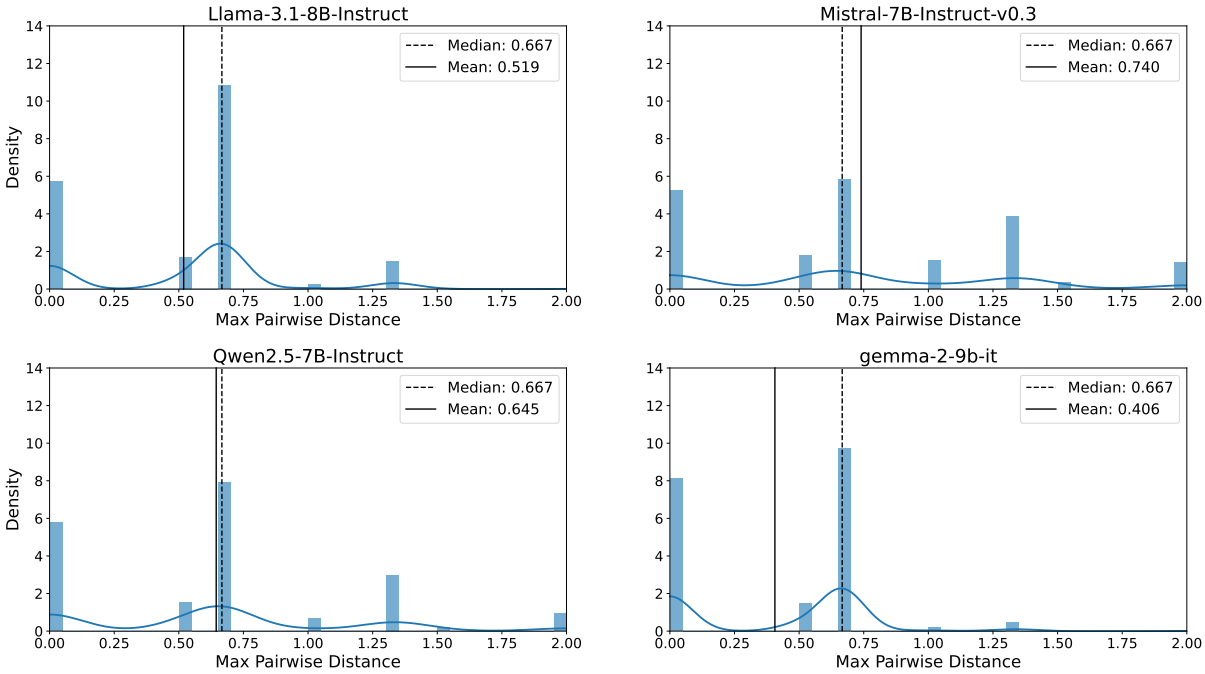


(b) Instruction-Tuned Models (APD on Reworded Dataset, Prompt 1)

Figure 14: Distribution of Average Pairwise Distances on the **Reworded Dataset**. Note the wider dispersion compared to the Shuffled Dataset due to the variable scale lengths.



(a) Base Models (MPD on Shuffled Dataset, Prompt 2)



(b) Instruction-Tuned Models (MPD on Shuffled Dataset, Prompt 2)

Figure 15: Distribution of Maximum Pairwise Distances on the **Shuffled Dataset**. Note the distinct "spikes" (gaps) caused by the fixed number of answer options.

A.6 Additional Volatility Analysis

This section complements the volatility analysis in Section 5.2.2. While the main text presented the Mean Standard Deviation for the Reworded Dataset, Table 6 presents the corresponding values for the Shuffled Dataset.

Table 6: MSD on the **Shuffled Dataset**. This table aggregates the volatility scores for both Base and Instruction-Tuned models. Note that for IT models, Column P3 (Unlabeled) consistently yields the lowest values, confirming the "Prompt 3 Paradox" discussed in the main text.

	Mean Standard Deviation			
Model	**P1 (Numeric)**	**P2 (Alpha)**	**P3 (Unlabeled)**	**P4 (Ext.)**
Base Models				
Llama-3.1-8B	0.367	0.403	0.445	0.328
Mistral-7B-v0.3	0.503	0.400	0.365	0.482
Qwen2.5-7B	0.323	0.361	0.363	0.330
Gemma-2-9B	0.553	0.346	0.415	0.378
Instruction-Tuned Models				
Llama-3.1-8B-Instruct	0.277	0.232	0.146	0.273
Mistral-7B-Instruct-v0.3	0.263	0.320	0.242	0.250
Qwen2.5-7B-Instruct	0.310	0.277	0.207	0.279
Gemma-2-9B-IT	0.233	0.187	0.122	0.217

Observations and Comparison. Comparing these values to the Reworded Dataset (Table 3) reveals a general increase in volatility, suggesting that option shuffling is typically more disruptive than rewording. This instability is most pronounced in Base models, where MSD values rise considerably across most configurations. A notable exception is *Qwen2.5-7B*, which maintains the lowest volatility among Base models, reinforcing its status as the most robust base model. Instruction-Tuned models display a more nuanced pattern: while their volatility generally rises compared to the reworded baseline (though to a lesser extent than Base models), they exhibit a unique anomaly in Prompt 3. The Unlabeled prompt consistently yields the lowest MSD for every IT model, quantitatively confirming the "Prompt 3 Paradox" where the removal of labels appears to enhance semantic consistency in randomized contexts.

A.7 Categorical Consistency on Shuffled Dataset

The categorical consistency analysis (Section 5.2) was extended to the Shuffled Dataset to verify if the trends observed in the Reworded Dataset hold when option order is randomized.

The overall patterns remain largely consistent with the main results: Instruction-Tuned models outperform Base models, with *Llama-3.1-8B-Instruct* maintaining the highest stability. The only notable deviation was observed in the *Qwen* family. Unlike other

models that show static or slightly degraded consistency under shuffling, both *Qwen2.5-7B* (Base) and *Qwen2.5-7B-Instruct* exhibit slightly higher rates of perfect cross-prompt agreement ("4 Same Answers") on the Shuffled Dataset compared to the Reworded one. This further supports the observation that the Qwen architecture possesses a distinct robustness to format variations, even when combined with semantic shuffling.
Detailed visualizations for these comparisons are available in the electronic appendix (Section B).

A.8 Correlation Stability (Reworded Dataset)

Complementing the analysis of Correlational Stability in Section 5.3.2, this section presents the Pearson correlation analysis for the Reworded Dataset. While the main text focused on the Shuffled Dataset, this dataset isolates the impact of changing the answer formulations and scale length.
Table 7 displays the correlation coefficients for all prompt pairs.

Table 7: Pearson Correlation Coefficients (r) on the **Reworded Dataset**. Note that correlations are generally higher here than in the Shuffled Dataset (Table 4), indicating that reworded answer options facilitate better cross-prompt consistency.

	Prompt Pair Correlations					
Model	**1 – 2**	**1 – 3**	**1 – 4**	**2 – 3**	**2 – 4**	**3 – 4**
Base Models						
Llama-3.1-8B	0.580	0.398	0.644	0.464	0.536	0.461
Mistral-7B-v0.3	0.485	0.408	0.454	0.354	0.336	0.278
Qwen2.5-7B	0.608	0.575	0.750	0.467	0.560	0.481
Gemma-2-9B	0.534	0.374	0.564	0.473	0.627	0.504
Instruction-Tuned Models						
Llama-3.1-8B-Instruct	0.820	0.701	0.894	0.712	0.797	0.683
Mistral-7B-Instruct-v0.3	0.707	0.722	0.913	0.646	0.709	0.703
Qwen2.5-7B-Instruct	0.818	0.743	0.883	0.737	0.799	0.726
Gemma-2-9B-IT	0.860	0.778	0.897	0.820	0.841	0.749

Observations and Comparison. Comparing these results to the Shuffled Dataset (Table 4) reveals that correlations are universally higher in the Reworded context, confirming that preserving logical ranking is easier for models when the option order is fixed. Among Base models, *Qwen2.5-7B* emerges as the clear outlier; it achieves correlations that approach the stability of Instruction-Tuned models, reinforcing the observation that its pre-training induced a stronger logical robustness than its competitors. Finally, structural trends remain consistent across datasets: the pair $1-4$ (Standard Numeric vs. Extended Numeric) yields the highest stability across all models (often > 0.90 for IT variants), confirming that models successfully recognize the structural similarity of these prompts despite the difference in scale length.

B Electronic Appendix

The complete code, datasets, and figures are available in the associated GitHub repository:

`https://github.com/martin-javier/llm-consistency_bachelors-thesis`

The repository contains the following supplementary materials referenced throughout this thesis:

- **Source Code:** Python scripts used for model prompting, data processing, and statistical analysis.
- **Datasets:** The full Reworded and Shuffled datasets, including raw and cleaned model outputs.
- **Visualizations:** All plots shown in this work along with additional plots not detailed in the main text or Appendix.

References

Aher, G. V., Arriaga, R. I. and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies, *in* A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett (eds), *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, PMLR, pp. 337–371.

Balepur, N., Rudinger, R. and Boyd-Graber, J. L. (2025). Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above, *in* W. Che, J. Nabende, E. Shutova and M. T. Pilehvar (eds), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vienna, Austria, pp. 3394–3418.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020). Language models are few-shot learners, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, pp. 1877–1901.

Couch, A. and Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable, *Journal of Abnormal and Social Psychology* **60**(2): 151–174.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**(3): 297–334.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *in* J. Burstein, C. Doran and T. Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Dykema, J., Schaeffer, N. C., Garbarski, D., Assad, N. and Blixt, S. (2022). Towards a reconsideration of the use of agree-disagree questions in measuring subjective evaluations, *Research in Social and Administrative Pharmacy* **18**(2): 2335–2344.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H. and Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models, *Transactions of the Association for Computational Linguistics* **9**: 1012–1031.

Gemma Team et al. (2024). Gemma: Open models based on gemini research and technology.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A. et al. (2024). The llama 3 herd of models.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Comput.* **9**(8): 1735–1780.

Höhne, J. K., Krebs, D. and Kühnel, S. M. (2022). Measuring income (in)equality: Comparing survey questions with unipolar and bipolar scales in a probability-based online panel, *Social Science Computer Review* **40**(5): 108–123.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S. et al. (2023). Mistral 7b.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C. and Kaplan, J. (2022). Language models (mostly) know what they know.

Khan, A., Casper, S. and Hadfield-Menell, D. (2025). Randomness, not representation: The unreliability of evaluating cultural alignment in llms, *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, New York, NY, USA, p. 2151–2165.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology* **5**(3): 213–236.

Krosnick, J. A. and Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement, *Public Opinion Quarterly* **51**(2): 201–219.

Lee, N., Hong, J. and Thorne, J. (2025). Evaluating the consistency of LLM evaluators, *in* O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio and S. Schockaert (eds), *Proceedings of the 31st International Conference on Computational Linguistics*, Association for Computational Linguistics, Abu Dhabi, UAE, pp. 10650–10659.

Likert, R. (1932). A technique for the measurement of attitudes, *Archives of Psychology* **22**(140): 1–55.

Lord, F. M., Novick, M. R. and Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley.

Lunardi, R., Mea, V. D., Mizzaro, S. and Roitero, K. (2025). On robustness and reliability of benchmark-based evaluation of llms.

Menold, N. (2021). Response bias and reliability in verbal agreement rating scales: Does polarity and verbalization of the middle category matter?, *Social Science Computer Review* **39**(1): 130–147.

Molfese, F. M., Moroni, L., Gioffré, L., Scirè, A., Conia, S. and Navigli, R. (2025). Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering, *in* W. Che, J. Nabende, E. Shutova and M. T. Pilehvar (eds), *Findings of the Association for Computational Linguistics: ACL 2025*, Association for Computational Linguistics, Vienna, Austria, pp. 18477–18494.

Mucciaccia, S. S., Meireles Paixão, T., Wall Mutz, F., Santos Badue, C., Ferreira de Souza, A. and Oliveira-Santos, T. (2025). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions, *in* O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio and S. Schockaert (eds), *Proceedings of the 31st International Conference on Computational Linguistics*, Association for Computational Linguistics, Abu Dhabi, UAE, pp. 2246–2260.

Ngweta, L., Kate, K., Tsay, J. and Rizk, Y. (2025). Towards LLMs robustness to changes in prompt format styles, *in* A. Ebrahimi, S. Haider, E. Liu, S. Haider, M. Leonor Pacheco and S. Wein (eds), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, Association for Computational Linguistics, Albuquerque, USA, pp. 529–537.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022). Training language models to follow instructions with human feedback, *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, pp. 27730–27744.

Pew Research Center (2024). The American Trends Panel (ATP). Dataset. Accessed via OpinionQA.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners, *OpenAI blog* **1**(8): 9.

Rupprecht, J., Ahnert, G. and Strohmaier, M. (2025). Prompt perturbations reveal human-like biases in large language model survey responses, *ArXiv* **abs/2507.07188**.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. and Hashimoto, T. (2023). Whose opinions do language models reflect?, *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, JMLR.org.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., DURMUS, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M. and Perez, E.

(2024). Towards understanding sycophancy in language models, *The Twelfth International Conference on Learning Representations*.

Turpin, M., Michael, J., Perez, E. and Bowman, S. R. (2023). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Curran Associates Inc., Red Hook, NY, USA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I. (2017). Attention is all you need, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T. and Sui, Z. (2024). Large language models are not fair evaluators, *in* L.-W. Ku, A. Martins and V. Srikumar (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 9440–9450.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. and Le, Q. V. (2022). Finetuned language models are zero-shot learners, *International Conference on Learning Representations*.

Xue, M., Hu, Z., Liu, L., Liao, K., Li, S., Han, H., Zhao, M. and Yin, C. (2024). Strengthened symbol binding makes large language models reliable multiple-choice selectors, *in* L.-W. Ku, A. Martins and V. Srikumar (eds), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, pp. 4331–4344.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B. et al. (2024). Qwen2 technical report.

Zhao, Z., Wallace, E., Feng, S., Klein, D. and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models, *in* M. Meila and T. Zhang (eds), *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 12697–12706.

Zheng, C., Zhou, H., Meng, F., Zhou, J. and Huang, M. (2024). Large language models are not robust multiple choice selectors, *The Twelfth International Conference on Learning Representations*.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Location, date

Name