Response Stability of Large Language Models Under Meaning-Preserving Prompt Variation

Idrees*, Liu Yongzhi

College of Foreign Languages and Cultures, Chengdu University of Technology, Erqianqiao #1, Chengdu, Sichuan 610059, China

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.100800423

Received: 19 August 2026; Accepted: 24 August 2026; Published: 07 September 2026

ABSTRACT

Single-prompt assessments provide limited evidence about whether task-relevant response properties remain stable when the same communicative intention is reformulated. This study examined response stability under controlled, meaning-preserving English prompt variation using 648 responses from 12 researcher-selected argumentative policy and education topics, six prompt variants, three accessed consumer systems, and three repeated runs. The analysis combined response-length screening, exact-duplication and cross-topic reuse detection, latent semantic analysis (LSA), structural instruction checks, bootstrap confidence intervals, and matched non-parametric comparisons. Under a fixed 100-dimensional LSA configuration, aggregate semantic stability was .975 for ChatGPT (95% bootstrap CI [.965, .985]), .968 for Claude [.948, .985], and .860 for Gemini [.791, .924]. ChatGPT and Claude showed comparably high semantic stability, whereas the accessed Gemini configuration was lower and more variable. All three run-specific Friedman tests yielded $\chi^2(2) = 24.000$, $p = 6.14 \times 10^{-6}$, Kendall's $W = 1.000$; this value indicates consistent within-run rank ordering across the 12 topics, not model superiority, and includes ties produced by exact repetition. ChatGPT and Claude met the experimental 180–250-word instruction in 100% of responses, compared with 33.3% for Gemini. Repetition complicated interpretation: Claude returned byte-identical six-variant blocks in 66.7% of run-topic blocks and Gemini in 33.3%, while Gemini Run 2 produced only 25 distinct texts across 72 responses without any complete block. LSA estimates also varied with representation settings, so model ordering based on small semantic-score differences was not robust. Only three of nine runs retained version and tier information, those documented configurations used unequal access tiers, and other session and inference conditions could not be reconstructed. The findings therefore describe the accessed consumer configurations under this task design, not controlled vendor-level or model-family differences. The study supports a multidimensional operational profile that reports semantic preservation, instruction following, topical relevance, structural adaptation, and repetition as related but distinct dimensions.

Keywords: large language models, response stability, prompt variation, semantic similarity, instruction following

INTRODUCTION

Large language models (LLMs) increasingly mediate access to information through natural-language instructions. Their outputs now appear in education, health communication, public administration, and other settings in which users may act on generated explanations or recommendations. Broad evaluation programmes such as HELM and BIG-bench have shown why model assessment should extend beyond a single accuracy score to robustness, calibration, bias, safety, and related dimensions (Bommasani et al., 2023; Chang et al., 2024; Srivastava et al., 2023). In public-service contexts the requirement is sharper still, because differences in how citizens phrase a question should not produce arbitrary changes in the substance of the guidance they receive, while explicit constraints attached to that guidance should still be honoured. Evidence from medical LLM evaluation points the same way: benchmark performance alone is insufficient, and long-form outputs require

human assessment of factuality, reasoning, potential harm, and usefulness (Singhal et al., 2023, 2025; Thirunavukarasu et al., 2023).

Prompt formulation is a plausible source of such variation. Demonstration order, template wording, formatting, paraphrase type, and other apparently minor choices can change model performance or even reverse model rankings (Lu et al., 2022; Sclar et al., 2024; Wahle et al., 2024; Webson & Pavlick, 2022). Multi-prompt studies have therefore challenged the common practice of characterising a model from a single instruction template (Mizrahi et al., 2024). Recent robustness frameworks treat prompt sensitivity as a property that requires repeated, controlled evaluation rather than as an incidental prompt-engineering nuisance (Chatterjee et al., 2024; Hua et al., 2025; Lior et al., 2025; Nalbandyan et al., 2025; Zhuo et al., 2024).

Open-ended public-service responses create an additional measurement problem. Two answers may differ lexically while retaining the same position, reasoning, risk assessment, and safeguard. Conversely, two nearly identical answers may both be inappropriate if they ignore a changed instruction or are reused for a different topic. Automatic generation metrics are useful but must be matched carefully to the construct being measured, and human judgement remains important for dimensions such as relevance and harmfulness (Gehrmann et al., 2021; Liu et al., 2023; Schmidtova et al., 2024; Stureborg et al., 2024; Zheng et al., 2023). A stability study should therefore distinguish semantic preservation from instruction compliance, functional component retention, topical relevance, structural adaptation, and exact repetition.

This distinction matters because several forms of repetition have different meanings. Identical responses to equivalent variants may represent legitimate invariance, whereas failure to change for Variant D indicates non-adaptation to a genuinely changed instruction, and cross-topic reuse indicates irrelevance. A similarity coefficient alone cannot distinguish these cases.

The present study evaluates these dimensions in a controlled design involving accessed ChatGPT, Claude, and Gemini consumer configurations. Twelve researcher-selected policy- and education-related tasks were expressed through six English prompt variants and collected across three recorded runs, yielding 648 responses. Every prompt requested 180–250 words, a position, two reasons, one risk, and one safeguard; Variant D additionally changed the requested order of these elements. The contribution is a multidimensional operational profile for asking whether systems preserve the properties that should remain stable while adapting to properties that the instruction genuinely changes. These measures represent related evaluation dimensions rather than indicators of a single latent stability construct. The study does not treat any system as universally superior, and it does not infer internal causes from observed differences in output.

LITERATURE REVIEW

Prompt sensitivity and meaning-preserving variation

Robustness research established early that language models can be sensitive to perturbations that do not alter the intended task (Moradi & Samwald, 2021). Prompt-based systems extend this problem because the instruction itself becomes part of the input. Lu et al. (2022) demonstrated strong few-shot order sensitivity, while Webson and Pavlick (2022) showed that successful prompt-based prediction does not guarantee human-like interpretation of prompt meaning. Linguistically focused work has found that grammatical, lexical, syntactic, discourse, and formatting choices can change outcomes even when intended meaning is preserved (Leidinger et al., 2023; Sclar et al., 2024; Wahle et al., 2024). Qiang et al. (2024) reported performance deterioration under synonym, paraphrase, and related perturbations, and proposed perturbation-consistency training as a mitigation; Gan and Mori (2023) reported comparable template sensitivity in Japanese text classification. Together these findings support explicit robustness testing rather than reliance on a single canonical prompt.

Multi-prompt and repeated evaluation

Recent work has moved from isolated examples toward systematic prompt-sensitivity measurement. POSIX quantifies changes under intent-preserving prompts and reports substantial sensitivity in open-ended generation (Chatterjee et al., 2024), and ProSA models sensitivity at the instance level (Zhuo et al., 2024), while Mizrahi et

al. (2024) show that instruction paraphrases can change both absolute scores and model rankings. SCORE argues for repeated testing under varied non-adversarial setups (Nalbandyan et al., 2025), and ReliableEval formalises evaluation over distributions of meaning-preserving perturbations (Lior et al., 2025). Hua et al. (2025) add an important caution: a substantial share of apparent prompt sensitivity is produced by rigid evaluation methods that fail to credit semantically valid alternative responses. Adversarial paraphrase work likewise treats semantically equivalent wording as a robustness stress test (Fu & Barez, 2025). These studies jointly motivate repeated, multi-prompt designs and metrics suited to open-ended language.

Automated metrics, semantic similarity, and human evaluation

Evaluation of generated text has long required a distinction between lexical overlap and semantic quality. LSA represents documents in a reduced semantic space derived from term–document structure (Deerwester et al., 1990). Later contextual methods such as Sentence-BERT and BERTScore provide alternative semantic representations (Reimers & Gurevych, 2019; Zhang et al., 2020). The present analysis uses LSA because it is fully transparent and reproducible from the archived corpus; contextual metrics are reserved for future validation. Surveys of natural language generation evaluation warn that automatic metrics are often adopted without sufficient justification or human validation (Schmidtova et al., 2024), and GEM treats evaluation as a combination of datasets, metrics, and human judgements rather than a single score (Gehrmann et al., 2021); human evaluation in turn requires explicit criteria and transparent procedures (Hämäläinen & Alnajjar, 2021). LLM-based evaluators offer scale but introduce their own biases and uncertainty (Liu et al., 2023; Stureborg et al., 2024; Vu et al., 2024; Watts et al., 2024; Zheng et al., 2023), and the choice of similarity coefficient and of evaluation dimensions itself affects conclusions about text quality (Fu et al., 2024; Wang et al., 2024; Zhelezniak et al., 2019).

Instruction following, public information, and oversight

Instruction following is a distinct property from semantic similarity. Verifiable constraints such as a required length or content element permit objective checks and have become the focus of dedicated instruction-following evaluation (Zhou et al., 2023). Medical research offers a useful analogue for high-stakes settings: MultiMedQA and Med-PaLM evaluations combine benchmark scores with human judgements of factuality, reasoning, bias, and potential harm, and later work continues to emphasise specialist review (Singhal et al., 2023, 2025), while reviews of LLMs in medicine stress verification and governance before deployment (Thirunavukarasu et al., 2023). These concerns transfer to public-service communication, where inconsistent risk statements or safeguards could alter how information is understood or acted upon.

Research gap

The literature establishes prompt sensitivity, the value of multi-prompt evaluation, and the limitations of single automatic metrics. Less attention has been given to extended argumentative responses evaluated simultaneously for semantic stability, functional components, topical relevance, explicit length compliance, discourse-order adaptation, exact within-topic repetition, cross-topic reuse, and repeated-run variation. The present study addresses this narrower gap with a common six-variant design applied across three proprietary consumer systems. Its component coding is AI-assisted rather than human double coding, so those results are treated as descriptive and are reported without an intercoder-reliability coefficient (Hayes & Krippendorff, 2007).

Research Questions And Objectives

The study addressed five research questions:

RQ1. To what extent do responses remain semantically and functionally stable across meaning-preserving English prompt variants?

RQ2. How do stability profiles differ across ChatGPT, Claude, and Gemini?

RQ3. How do models differ in structural responsiveness, that is, in whether they adapt a response's wording as against its requested organisation?

RQ4. How do models respond to the changed discourse-order instruction in Variant D?

RQ5. How much do repeated runs alter the observed stability profile?

The corresponding objectives were to quantify semantic and surface stability, to assess instruction compliance and structural responsiveness, to compare model and run effects, and to evaluate whether a profile-based framework is more informative than a single similarity score. RQ3 is addressed through the two structural measures that are independently reproducible from the response text; the retention of individual argumentative components is treated as exploratory for the reasons set out in Section 4.5.

METHODOLOGY

Research design and corpus

A controlled comparative repeated-measures design was used. The corpus contained three accessed product labels, three recorded runs per label, 12 underlying topics, and six prompt variants per topic: $3 \times 3 \times 12 \times 6 = 648$ responses. The run–topic block, comprising the six variant responses produced by one model for one topic in one run, was the principal matched unit for semantic inference. The topics concerned public-service and education propositions that require an argumentative response rather than a factual short answer, so that surface wording could vary substantially while the underlying position remained constant. Table 1 summarises the design, and Figure 1 shows the analysis workflow.

Table 1. Experimental design and analytical dimensions

Dimension	**Specification**
Accessed product labels	ChatGPT, Claude, Gemini
Recorded runs	3 per model
Underlying topics	12
Prompt variants	6 per topic (A–F)
Responses per run	72
Responses per model	216
Total responses	**648**
Required output	180–250 words; a position; two reasons; one risk; one safeguard
Special manipulation	Variant D requested the risk and safeguard before the position and reasons

Note. Product-family labels are used throughout. Where provenance was recorded, accessed configurations differed by tier: ChatGPT Run 1 was ChatGPT 5.6 on a paid plan, Claude Run 2 was Claude Opus 5 on a paid plan, and Gemini Run 1 was Gemini 3.5 Flash Lite on a free plan. The remaining six runs carry no recorded version or plan. Results therefore pertain to accessed consumer configurations rather than to model families as such.

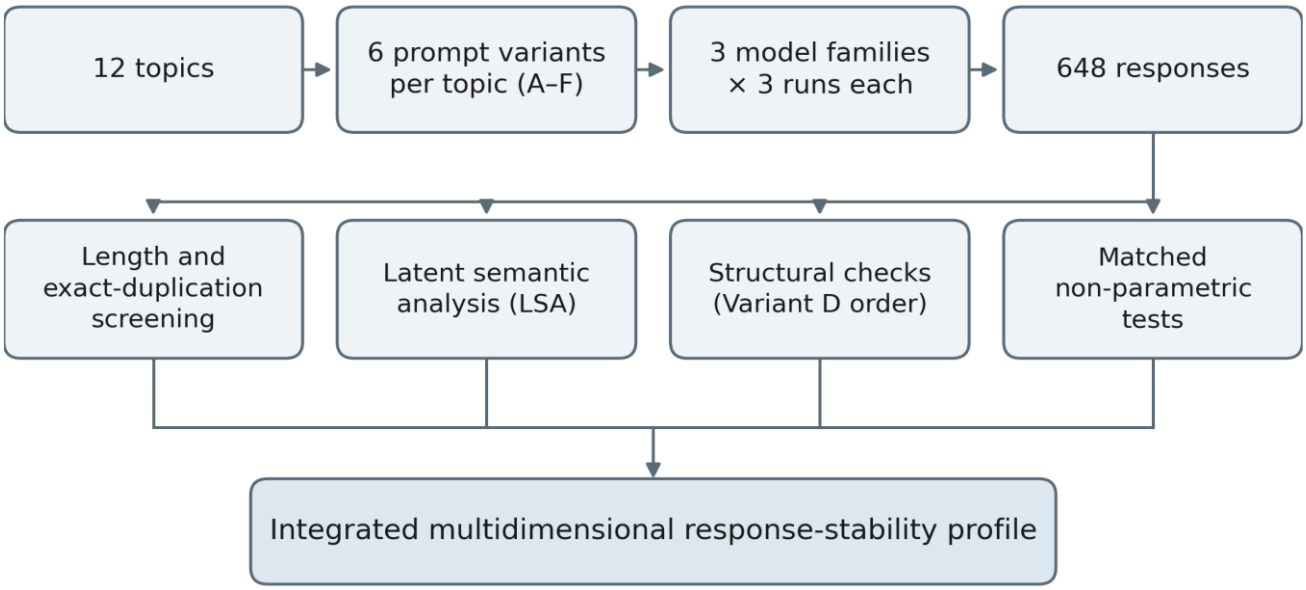

Figure 1. Experimental design and analysis workflow. The corpus is constructed by crossing 12 topics with 6 prompt variants, 3 accessed product labels, and 3 recorded runs, giving 648 responses; four analytical strands are then applied to that corpus and combined into the integrated profile reported in Table 6 and Figure 7.

Prompt design and provenance

Within each topic, variants A–F preserved the same communicative task while varying wording, syntax, register, or information order. Every prompt consisted of a question stem followed by a fixed instruction block. For variants A, B, C, E, and F that block read: "Respond in 180 to 250 words. State your position, give two reasons, identify one risk, and propose one safeguard." For Variant D it read: "Identify one risk and propose one safeguard. Then state your position and give two reasons. Keep your response between 180 and 250 words." The instruction block was therefore held constant across five of the six variants, so that variation was confined to the stem.

The stems instantiate a consistent typology. Variant A is the baseline formulation (mean 10.3 words). Variant B is a plain-language paraphrase substituting everyday vocabulary for institutional terms (13.2 words). Variant C reframes the proposition as an interrogative addressed to the model (16.1 words). Variant E adopts a formal, nominalised register (15.3 words). Variant F topicalises the object of the proposition, typically through passivisation (12.8 words). Variant D is the order manipulation: its stem is character-identical to Variant A in all twelve topics, and only the instruction block differs. Any difference between the A and D responses for a given topic and run is therefore attributable to the change in requested discourse order alone, with the propositional content held exactly constant. The complete prompt set is reproduced in Appendix A.

Responses were retained verbatim, without editing or truncation, so that every derived variable could be traced back to its source text. Each response was stored against its model, run, topic, and variant identifiers, and the completeness of the 648-cell design was confirmed by checking that every identifier combination occurred exactly once.

Provenance metadata were incomplete and, where available, showed unequal access tiers: ChatGPT Run 1 was recorded as ChatGPT 5.6 on a paid plan, Claude Run 2 as Claude Opus 5 on a paid plan, and Gemini Run 1 as Gemini 3.5 Flash Lite on a free plan. The other six runs did not preserve equivalent version or plan information. Exact collection dates and timestamps were not retained. It also cannot be reconstructed whether each prompt used an independent conversation, whether variants were always administered in A–F order, whether memory or personalisation features were enabled, or which system prompts and inference parameters applied. These missing conditions create possible prompt-order, session-context, temporal, and tier confounds. All numerical

comparisons are therefore descriptive of the accessed configurations and do not constitute controlled vendor-level comparisons.

Length, duplication, and topical screening

Word counts were computed under one reproducible tokenisation rule applied identically to all 648 responses. Length compliance was scored as 1 for responses of 180–250 words inclusive and 0 otherwise. The interval was an artificial but objectively verifiable instruction-following constraint in this experiment; it is not presented as an inherent dimension of semantic stability or as a general reliability criterion. Exact normalised response strings were compared within each model–run–topic block to identify six-variant repetition, that is, blocks in which all six variants returned byte-identical text after whitespace normalisation. A separate cross-topic screen identified verbatim reuse across different topics. Topical relevance in the confirmatory profile was derived only from this objective screen: each response was scored relevant unless its text was demonstrably reused from another topic in a way that made it off-topic. This corpus-derived measure is distinct from the broader AI-assisted topical coding. The within-topic and cross-topic screens measure different phenomena and are reported separately throughout.

Latent semantic analysis

All 648 responses were transformed into TF-IDF-weighted unigram and bigram features with sublinear term-frequency scaling and English stop-word removal. Truncated singular value decomposition reduced the pooled response matrix to 100 dimensions, after which vectors were length-normalised. Cosine similarity was calculated for the 15 unique pairs among the six variants within each model–run–topic block, and the arithmetic mean of those 15 values defined block semantic stability. The 100-dimensional setting was fixed in advance for the main analysis.

Formally, for a model–run–topic block containing the six reduced, length-normalised variant vectors v_a … v_f, the similarity of any pair is the cosine

$$cos(v_i , v_j) = (v_i \cdot v_j) / (\|v_i\| \|v_j\|) \quad (1)$$

and block semantic stability S is the mean of the 15 unique pairs,

$$S = (1/15) \Sigma_{i<j} cos(v_i , v_j) \quad (2)$$

so that S = 1 indicates that all six variant responses occupy the same direction in the reduced space, and S = 0 indicates orthogonality. Because vectors are length-normalised and TF-IDF weights are non-negative, S is bounded on [0, 1].

Because the decomposition is fitted across the pooled corpus, block-level similarities depend on corpus composition. A robustness grid therefore examined TF-IDF without SVD, SVD at 50, 100, and 150 dimensions, and 100-dimensional SVD fitted separately within each model–run. This sensitivity analysis is reported alongside the main estimates and cautions against treating small differences in LSA scores as invariant model properties.

Structural and AI-assisted coding

A structured AI-assisted workflow was used to code position, reasons, risk, safeguard, topical relevance, and Variant D order. The broader AI-assisted coding materials are incomplete and those results are not used in the confirmatory profile. Two structural measures are independently reproducible from the response text and are reported in the main results. The first records whether the Variant D response differs textually from the Variant A response for the same topic and run. The second is an opening-risk proxy for Variant D sequence adaptation: a response satisfies the proxy when its opening sentence does not state a position on the proposition and instead begins with a statement of risk. Position statements were identified by first-person stance expressions, by “in my view” or “in my opinion,” by a sentence-initial “yes” or “no,” and by deontic modals. The rule was applied

identically to all 108 Variant D responses. This proxy does not verify the complete risk → safeguard → position → reasons sequence and is not described as full order compliance.

The archived coding materials do not include the coding prompt, the complete coded matrix, or the script required to reproduce the position, reason, risk, and safeguard retention percentages independently. Those measures are therefore treated as exploratory, are excluded from the confirmatory integrated profile, and are not used for model ranking. No independent human double-coding dataset exists, so no intercoder-reliability statistic is reported (Hayes & Krippendorff, 2007).

Statistical analysis and reproducibility

For semantic stability, Friedman tests compared the three matched accessed-configuration scores separately within each recorded run across the 12 topics. Exact equality was imposed for byte-identical response blocks before ranking, which prevents floating-point residuals from creating artificial rank differences, and the tie-corrected form of the statistic was used. Because run labels did not represent fully matched occasions across providers and provenance differed, these tests describe within-run output differences rather than controlled model-family effects. Kendall's W summarised agreement in rank ordering across topics, obtained from the Friedman statistic for k conditions and n blocks as

$$W = \chi^2 / [n\,(k - 1)] \quad (3)$$

so that W = 1 denotes complete agreement of the rank ordering across blocks and W = 0 denotes none. Response length was compared across the 216 matched prompt cells using Friedman and Wilcoxon signed-rank procedures, with the matched-pairs rank-biserial correlation as the pairwise effect size (Kerby, 2014; Lakens, 2013). Non-parametric bootstrap confidence intervals summarised uncertainty around aggregate semantic means, using 10,000 percentile-method resamples of the 36 block-level scores per model with the random seed fixed at 20260820 (Davison & Hinkley, 1997). Semantic robustness was assessed across the alternative representations described in Section 4.4 rather than by relying on four-decimal point estimates from a single configuration. All statistics reported in this article were computed from the archived corpus in a single scripted pass, so that every value in the tables and figures is traceable to the response text from which it derives.

Ethics and data handling

The analysis used model-generated text and did not involve human participants or personal data. Raw responses were preserved to maintain an auditable link between source text and derived variables. Any collection, archiving, or redistribution of outputs from proprietary consumer interfaces remains subject to the applicable platform terms and journal policies. The study treats the reported measures as an operational reliability profile rather than evidence of factual correctness, so human verification remains necessary before generated public information is relied upon in consequential settings.

RESULTS

Dataset integrity and response length

All 648 expected response cells were present, and no cell required imputation. ChatGPT and Claude met the 180–250-word requirement in all 216 responses each. Gemini complied in 72 of 216 responses (33.3%); compliance was 0% in Runs 1 and 2 and 100% in Run 3. The matched length analysis across the 216 prompt cells yielded $\chi^2(2) = 45.940$, $p = 1.06 \times 10^{-10}$, with Kendall's W = .106, indicating a reliable but small omnibus difference in length ranks. Matched-pairs rank-biserial correlations were −.472 for ChatGPT versus Claude, +.706 for ChatGPT versus Gemini, and +.776 for Claude versus Gemini, with positive values indicating that the first-named model produced the longer responses. Table 2 and Figure 2 report the descriptive picture.

All values reported in this article were computed from the archived 648-response corpus using the pipeline specified in Section 4.6.

Table 2. Response-length characteristics by model and run

Model	Run	Mean words	SD	Range	Length compliance
ChatGPT	1	219.07	7.50	205–236	**100.0%**
	2	188.42	4.94	180–197	**100.0%**
	3	193.49	7.91	180–210	**100.0%**
Claude	1	200.33	3.89	193–207	**100.0%**
	2	220.72	8.05	192–235	**100.0%**
	3	210.25	5.39	202–220	**100.0%**
Gemini	1	106.12	6.69	95–124	0.0%
	2	138.57	10.82	123–161	0.0%
	3	227.08	1.67	224–229	**100.0%**

Note. Compliance denotes the proportion of responses falling within the requested 180–250-word interval. Each row summarises 72 responses. Shaded rows mark the model–run cells that fell outside the requested interval. Word counts use a single tokenisation rule in which a token is a maximal run of alphanumeric characters, optionally containing an internal apostrophe or hyphen. All rows were computed from the archived corpus under this rule.

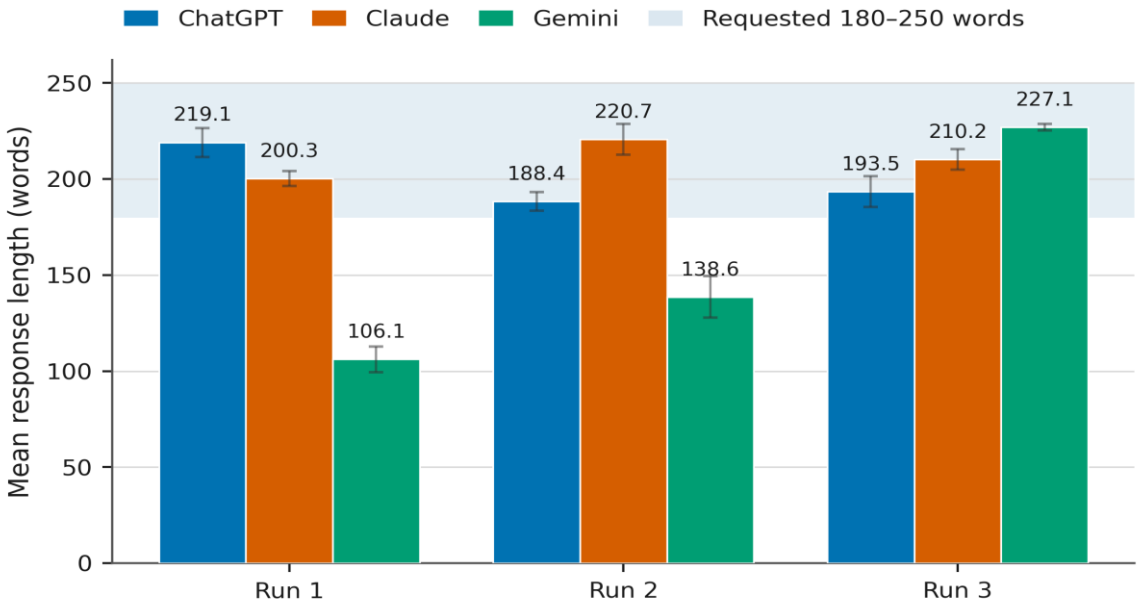

Figure 2. Mean response length by model and run. Each bar summarises n = 72 responses (12 topics × 6 variants). Error bars show ±1 standard deviation. The shaded band marks the requested 180–250-word interval; a bar lying wholly outside it indicates zero length compliance for that model–run cell.

Exact repetition and cross-topic reuse

Exact repetition differed markedly across models. ChatGPT produced no run–topic block in which all six variants were byte-identical. Claude produced 24 identical blocks out of 36 (66.7%), all in Runs 1 and 3. Gemini produced 12 identical blocks (33.3%), all in Run 3. Gemini Run 2 contained only 25 distinct response texts

across its 72 responses, so duplication was extensive without ever extending to a complete six-variant block, which is why the block count for that cell is zero. In Gemini Run 3, one response was reused verbatim across topics 1, 10, 11, and 12. This resulted in three affected topic blocks, corresponding to 18 off-topic responses, and reduced Gemini's overall topical relevance to 91.7%. Within-topic repetition is therefore reported as surface invariance, whereas cross-topic reuse is treated as reduced task responsiveness; the two are distinct problems and are not combined into a single score. Table 3 and Figure 3 summarise the duplication results.

Table 3. Exact-response duplication by model and run

Model	Run 1texts	Run 1 blocks	Run 2 texts	Run 2 blocks	Run 3 texts	Run 3 blocks
ChatGPT	72	0	72	0	72	0
Claude	12	**12**	72	0	12	**12**
Gemini	72	0	25	0	9	**12**

Note. "Texts" denotes the number of distinct response strings within a run, out of 72. "Blocks" denotes the number of topics, out of 12, in which all six A–F responses were byte-identical. These are different units and are not comparable to one another. Cross-topic reuse is evaluated separately and is not represented in this table.

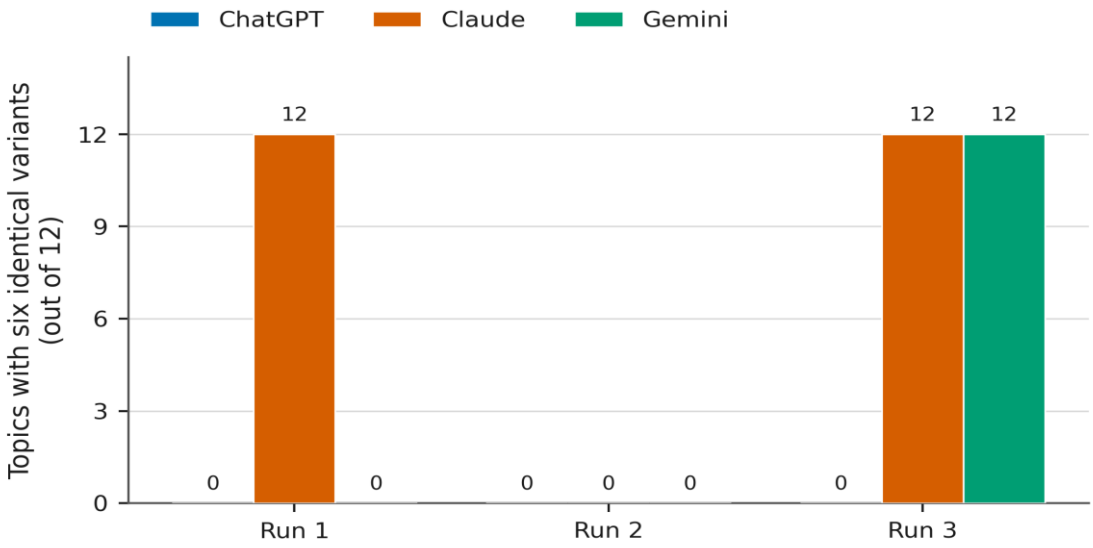

Figure 3. Exact within-topic repetition by model and run. Bars show the number of topics, out of 12, in which all six prompt variants returned byte-identical text after whitespace normalisation. A value of 0 indicates that no topic in that model–run cell was fully duplicated; it does not imply the absence of partial duplication, which is reported separately in Table 3.

Semantic stability

Under the fixed 100-dimensional LSA configuration, aggregate semantic stability was .975 for ChatGPT (95% bootstrap CI [.965, .985]), .968 for Claude [.948, .985], and .860 for Gemini [.791, .924]. The bootstrap intervals for ChatGPT and Claude overlap substantially, so the two should not be ranked against each other on this measure. Run-specific means were .991, 1.000, and .579 in Run 1; .997, .905, and 1.000 in Run 2; and .937, 1.000, and 1.000 in Run 3, for ChatGPT, Claude, and Gemini respectively. For each run, the tie-corrected Friedman test yielded $\chi^2(2) = 24.000$, $p = 6.14 \times 10^{-6}$, Kendall's $W = 1.000$. Here W indicates complete agreement in within-run rank ordering across the 12 topics; it is not evidence of universal superiority, especially because exact repetition creates ceiling ties. In Run 3, Claude and Gemini are tied by construction because all six variants were byte-identical within every topic for both accessed configurations, and no Claude–Gemini pairwise difference is claimed. Table 4 and Figure 4 report these values.

The Gemini Run 2 value reaches the ceiling without any fully identical block, and the duplication screen explains why: that run returned only 25 distinct texts across its 72 responses, so most variants within a topic repeat one of two or three answers. Partial duplication of this kind raises within-topic similarity almost as effectively as complete duplication, which is why the block count in Table 3 and the similarity value in Table 4 must be read together rather than separately.

Table 4. Latent-semantic stability by model and run

Model	Overall mean	95% bootstrap CI	Run 1	Run 2	Run 3
ChatGPT	**0.975**	[0.965, 0.985]	0.991	0.997	0.937
Claude	**0.968**	[0.948, 0.985]	**1.000**[a]	0.905	**1.000**[a]
Gemini	**0.860**	[0.791, 0.924]	0.579	1.000[b]	**1.000**[a]

Note. Values are reported to three decimals because the fourth decimal is not stable across LSA configurations. [a] *Cells marked with this superscript reflect byte-identical six-variant responses and therefore represent exact repetition rather than independent evidence of semantic preservation.* [b] *Gemini Run 2 contains no fully identical six-variant block, but only 25 distinct texts across its 72 responses (Table 3); the ceiling value reflects that heavy partial duplication rather than adaptive preservation of meaning. ChatGPT and Claude have overlapping bootstrap intervals and are not ranked against each other.*

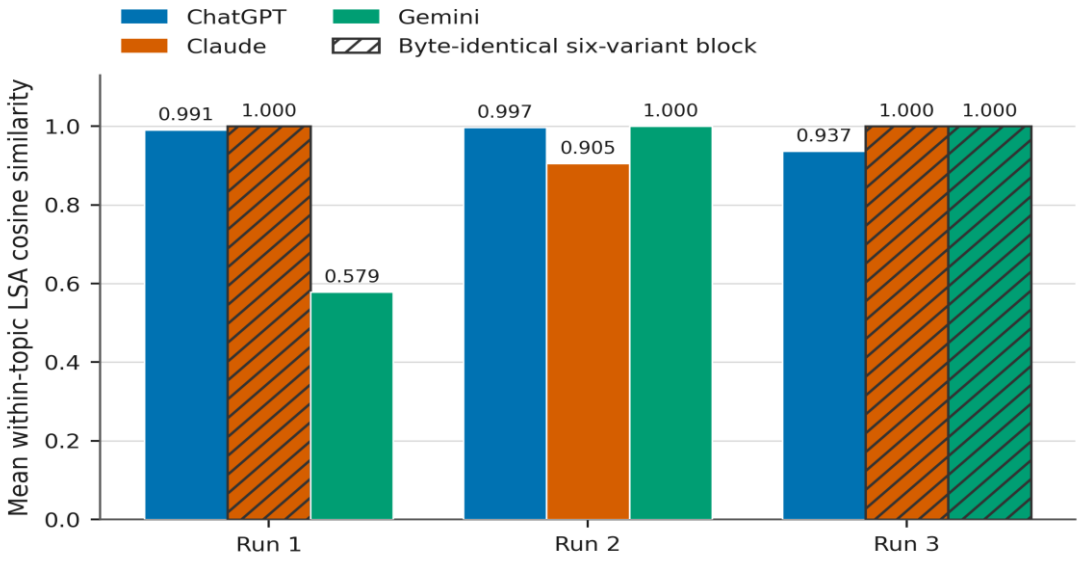

Figure 4. Mean within-topic LSA cosine similarity by model and run, computed under the fixed 100-dimensional pooled configuration (Equations 1–2). Each bar is the mean of 12 topic blocks, each itself the mean of 15 pairwise cosines, so each bar summarises 180 pairwise comparisons. Hatched bars mark cells in which all six variants were byte-identical, so the value of 1.000 reflects exact repetition rather than adaptive preservation of meaning. The unhatched Gemini Run 2 bar also reaches the ceiling, but through heavy partial duplication rather than complete six-variant repetition (Section 5.3). LSA = latent semantic analysis.

Aggregating across models, runs, and topics, similarity between individual prompt-variant pairs was uniformly high but not uniform. Figure 5 shows that the A–D pair was the most similar (.970), consistent with Variant D frequently reproducing the Variant A text rather than reordering it, whereas pairs involving Variant E were the least similar, with the E–F pair lowest at .899. The spread across pairs is narrow, which indicates that no single variant behaved as an outlier in the design.

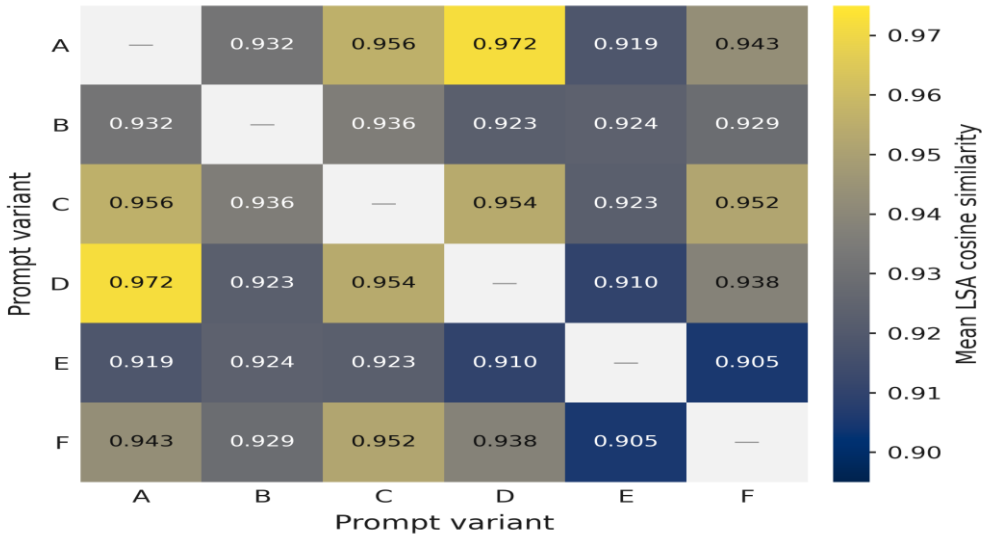

Figure 5. Mean LSA cosine similarity between prompt-variant pairs, aggregated across models, runs, and topics. Each off-diagonal cell is the mean of 108 comparisons (3 models × 3 runs × 12 topics) and the matrix is symmetric. Diagonal cells are omitted because self-similarity is 1.000 by definition. Variant labels A–F follow Appendix A. Colour encodes similarity on the scale shown; values are printed in each cell so the figure can be read without reference to the colour bar.

The representation-sensitivity analysis showed that aggregate scores shift substantially with the LSA configuration, and that they do so unevenly across cells. Figure 6 plots all nine model–run cells over the five configurations. The three cells containing byte-identical responses, Claude Runs 1 and 3 and Gemini Run 3, sit at 1.000 everywhere, as they must, because identical inputs produce identical vectors regardless of the representation. The three ChatGPT cells move within a narrow band, from about .85 under raw TF-IDF to about .99 at SVD-50 and back to about .87 under within-run fitting, so their ordering is stable even though their absolute values are not. The two cells with the least duplication move furthest: Gemini Run 1 ranges from .072 to .900 and Claude Run 2 from .282 to .958. Gemini Run 2 stays near the ceiling throughout, which is consistent with its heavy partial duplication. The direction of Gemini's lower aggregate stability is therefore robust across configurations, but the fine-grained ordering of ChatGPT relative to Claude is not, and no conclusion in this article rests on a difference of a few thousandths.

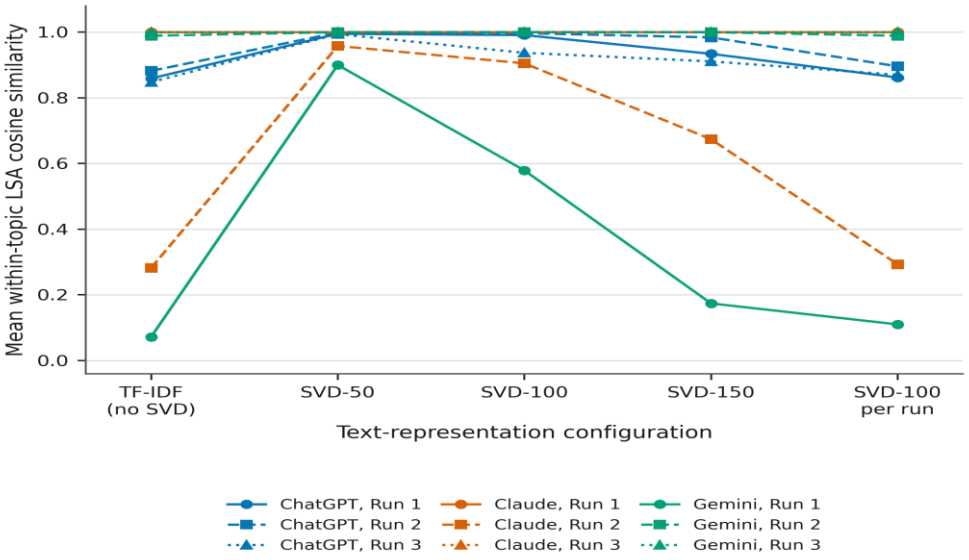

Figure 6. Sensitivity of within-topic LSA stability estimates to the text-representation configuration. Each plotted point is a run-level mean over 12 topic blocks, each itself the mean of 15 pairwise cosines. All nine model–run cells are plotted; colour identifies the model and line style the run. SVD-k denotes truncated singular value

decomposition to k dimensions fitted on the pooled corpus; "per run" denotes the same reduction fitted within each model–run separately. SVD = singular value decomposition.

Structural responsiveness to the Variant D manipulation

The Variant D response differed textually from the Variant A response in 100.0% of ChatGPT blocks, 33.3% of Claude blocks, and 66.7% of Gemini blocks. Opening-risk proxy satisfaction under the rule stated in Section 4.5 was 100.0% for ChatGPT, 33.3% for Claude, and 66.7% for Gemini. The two measures coincided in all 108 blocks: wherever an accessed configuration produced a distinct Variant D response, the response began with risk; wherever it did not, the response reproduced Variant A. This result demonstrates adaptation of the opening structure, not verified compliance with the complete four-component sequence. The Claude and Gemini figures follow from the duplication results, since Variant D cannot differ from Variant A in a byte-identical block: Claude had only its 12 Run 2 blocks available for adaptation and used all of them, and Gemini had its 24 Run 1 and Run 2 blocks and used all of them. ChatGPT, which never produced an identical block, satisfied the proxy throughout. Table 5 combines these structural measures with the relevance and repetition diagnostics.

Table 5. Structural responsiveness, relevance, and repetition diagnostics

Model	Variant D opening-risk proxy	Variant D differs from A	Topical relevance	Identical blocks	Interpretation
ChatGPT	**100.0%**	**100.0%**	**100.0%**	0.0%	Rewrites and reorders in every block
Claude	33.3%	33.3%	**100.0%**	**66.7%**	Adapts in every block that repetition leaves available
Gemini	66.7%	66.7%	91.7%	**33.3%**	Adapts where repetition allows, but cross-topic reuse is present

Note. Percentages for the first two columns are computed over the 36 run–topic blocks per accessed configuration; topical relevance is computed over the 216 responses per accessed configuration and is derived solely from objectively detected cross-topic verbatim reuse. The first two columns are numerically identical because textual adaptation and opening-risk proxy satisfaction coincided in all 108 Variant D blocks; both are retained because they measure conceptually distinct properties. The proxy does not verify the complete requested sequence. Position, reason, risk, safeguard, and broader AI-coded topical-relevance results are not reported because the corresponding coding artefacts are incomplete. No independent human intercoder reliability was calculated.

Integrated profile

Table 6 and Figure 7 assemble the reproducible dimensions into a single operational profile. Across these dimensions, no accessed configuration dominates: ChatGPT and Claude lead on semantic stability, length compliance, and topical relevance, while ChatGPT alone adapts and satisfies the opening-risk proxy in every Variant D block, and Claude, whose semantic score is the second highest, adapts in only a third of them. The identical-block rate is reported as a diagnostic rather than as a performance dimension because repetition has different meanings: invariance under equivalent wording may be legitimate, failure to adapt to Variant D is undesirable, and cross-topic reuse indicates irrelevance. No composite index is computed because these dimensions are not assumed to measure one latent construct and point in different directions.

Table 6. Integrated multidimensional response-stability profile

Measure	ChatGPT	Claude	Gemini
Semantic stability (%)	**97.5**	**96.8**[a]	86.0
Length compliance (%)	**100.0**	**100.0**	33.3
Topical relevance (%)	**100.0**	**100.0**	91.7
Variant D adapted and opening-risk proxy met (%)	**100.0**	33.3	66.7
Identical six-variant blocks (%)	0.0	66.7	33.3

Note. Semantic stability is the mean LSA cosine similarity multiplied by 100 for display; all other values are percentages. Textual adaptation and opening-risk proxy satisfaction are reported as a single row because they coincided in every Variant D block (Table 5); the proxy does not verify the complete four-component sequence. [a] The Claude value is inflated by exact repetition and is not comparable to the ChatGPT value. The shaded final row is a diagnostic and is not scored as desirable stability. Component-level retention measures are omitted pending archiving of the AI-assisted coding materials. No composite ranking is calculated.

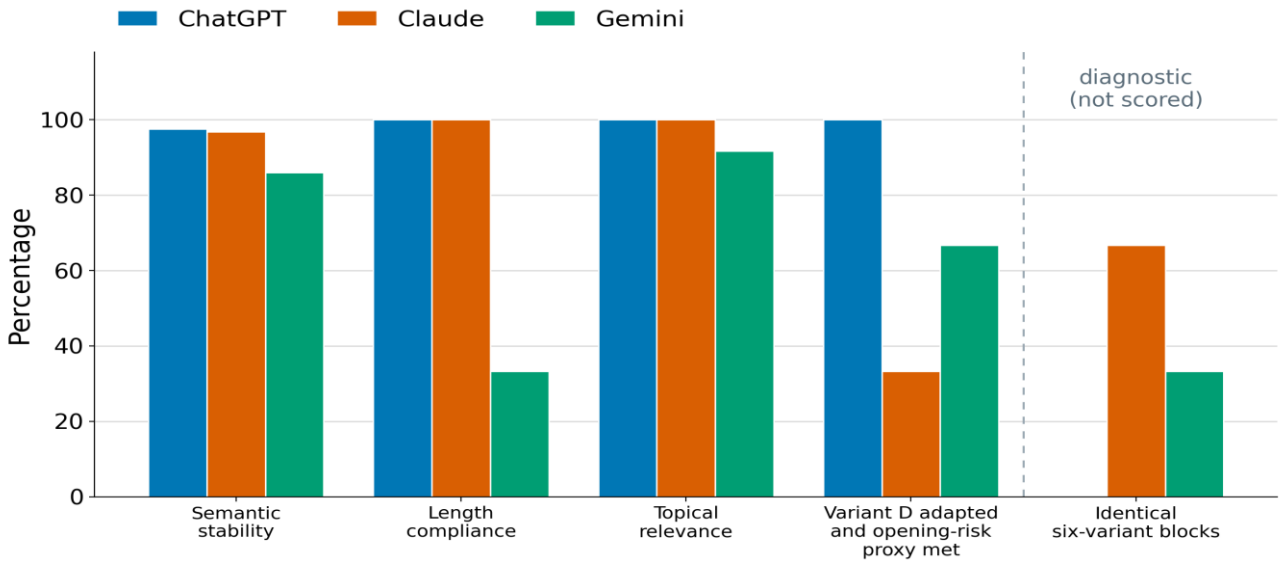

Figure 7. Integrated response-stability profile, plotting the five dimensions of Table 6. Semantic stability is the mean LSA cosine similarity multiplied by 100 for display; all other dimensions are percentages. Denominators differ by dimension: length compliance and topical relevance are computed over 216 responses per model, and the Variant D and identical-block measures over 36 run–topic blocks per model. The final group, to the right of the dashed rule, is a repetition diagnostic and is not scored as desirable stability.

DISCUSSION

The strongest finding of this study is that semantic similarity alone cannot distinguish several response patterns. Identical answers to the equivalent wording variants A, B, C, E, and F may represent legitimate invariance when the underlying task is unchanged. By contrast, reproducing the same answer for Variant D indicates failure to adapt to a genuinely changed discourse-order instruction, while reuse across unrelated topics indicates reduced relevance. In Claude Runs 1 and 3, every topic produced the same response across all six variants, including Variant D. Those blocks receive an LSA score of 1.000 while showing no adaptation to the changed instruction.

In Gemini Run 3, exact repetition also produced perfect within-topic similarity, and one response was reused across four different topics. Stability should therefore be evaluated by preserving properties that ought to remain constant while separately checking adaptation where the instruction changes and relevance where the topic changes.

Partial duplication produces the same effect by a weaker route. Gemini Run 2 contains no fully identical block, yet returns only 25 distinct texts across 72 responses, and its similarity sits at the ceiling. A block-level duplication screen alone would have recorded that cell as clean. Reporting the count of distinct texts alongside the count of identical blocks is therefore necessary, not merely descriptive. A high similarity score is in any case a joint property of the responses and of the representation used to compare them: the sensitivity analysis shows the least duplicated cells moving across most of the available range as the representation changes, so neither a high value nor a low one can be read as a model characteristic unless the representation is stated alongside it.

The semantic results also require methodological restraint. Under the 100-dimensional configuration, ChatGPT and Claude both show high aggregate stability with overlapping bootstrap intervals, while Gemini is lower. Yet the sensitivity grid shows that absolute values depend strongly on representation choices, with the least duplicated cells moving across most of the available range. Differences of a few thousandths between ChatGPT and Claude are therefore not robust enough to support a stable ordering, and reporting them to four decimal places would overstate their precision. This observation complements prior work showing that prompt-sensitivity conclusions depend partly on evaluation design and metric choice (Hua et al., 2025; Mizrahi et al., 2024).

Repeated-run differences must be interpreted alongside provenance. The three runs with preserved metadata were not collected on comparable tiers: ChatGPT Run 1 and Claude Run 2 were recorded on paid configurations, whereas Gemini Run 1 was recorded as Gemini 3.5 Flash Lite on a free plan, and the remaining six runs lack equivalent provenance. Because Gemini's weakest length and semantic results occur in the documented free-tier run, those outcomes cannot be attributed cleanly to the Gemini family. The analysis therefore treats model labels as accessed consumer configurations and avoids claims of universal superiority.

Instruction following provides a complementary signal. ChatGPT and Claude met the experimental 180–250-word requirement in every response, whereas Gemini met it only in Run 3. This length measure is one verifiable example of compliance with an imposed instruction, not a general criterion of semantic stability. The Variant D opening-risk proxy was highest for ChatGPT, the only system that never returned an identical block, and lowest for Claude, whose repetition left only a third of its blocks available for adaptation. Because proxy satisfaction and textual adaptation coincided in this corpus, the systems were effectively all-or-nothing on the observed opening structure: they either rewrote the response and began with risk as requested, or reproduced the Variant A text unchanged. The proxy does not establish compliance with the entire four-component sequence. This limitation reinforces the need to report structural responsiveness separately from semantic similarity.

The AI-assisted component layer is retained only with clear evidential boundaries. The Variant D opening-risk proxy is reproducible from structural markers, but it does not verify the complete requested sequence. The archived materials also do not support independent reproduction of the position, reason, risk, and safeguard retention percentages. Those values do not drive confirmatory conclusions and should remain exploratory until the coding prompt, coded matrix, and script are archived or the corpus is independently recoded. Separating reproducible corpus-derived findings from exploratory automated coding makes the evidential status of each claim explicit.

For public-service communication, the practical implication is methodological rather than a claim that one vendor is preferable. Evaluators should test multiple meaning-preserving formulations, repeat generation across sessions, verify explicit constraints, screen for reuse, and inspect whether high similarity arises from substantive preservation or from repetition. Because this study evaluates response stability rather than factual truth, human verification remains necessary wherever consequential public guidance is involved.

Limitations And Future Research

The study covers twelve researcher-selected English argumentative topics concentrated in policy and education, including six topics on AI governance or AI in higher education. The 648 responses are therefore not 648 statistically independent observations; the principal matched semantic unit is the run–topic block, with only 12 topics per run. The findings do not generalise automatically to a defined population of prompts, multilingual interaction, factual question answering, summarisation, creative generation, technical reasoning, or other task types. Provenance is incomplete for six of the nine runs. Where metadata exist, the accessed tiers were unequal: ChatGPT Run 1 and Claude Run 2 were recorded on paid plans, whereas Gemini Run 1 was recorded as Gemini 3.5 Flash Lite on a free plan. Exact dates, timestamps, prompt order, conversation resets, memory and personalisation settings, system prompts, and inference parameters were not retained. These omissions limit temporal reproducibility and create possible tier, session, and order confounds, especially because Gemini's weakest results occur in the documented free-tier run. Future work should use pinned API model versions, matched access tiers, controlled independent sessions, fixed inference settings, randomised prompt order, exact timestamps, and a larger, more diverse topic sample.

LSA is transparent and reproducible, but the sensitivity analysis shows that absolute estimates depend strongly on dimensionality and fitting strategy. The reported values should therefore be read as measurement-specific estimates rather than precise, invariant system properties. Future replication should preregister representation settings and triangulate LSA with contextual methods such as Sentence-BERT or BERTScore and human semantic-equivalence judgments on a representative subset (Reimers & Gurevych, 2019; Zhang et al., 2020). The opening-risk rule used for Variant D is only a proxy and does not verify the full risk → safeguard → position → reasons sequence. The AI-assisted component layer lacks a complete archived coding prompt, coded matrix, script, and independent human double coding, so its component-level results remain exploratory and no intercoder-reliability statistic is reported. Future work should verify the complete Variant D sequence, add at least two human raters for a stratified subset, report intercoder agreement, preserve all coding artefacts, extend the design to multilingual and varied task types, and evaluate factual accuracy alongside stability.

CONCLUSION

Across 648 responses, this analysis supports a multidimensional rather than a rank-based account of LLM response stability. ChatGPT and Claude showed comparably high LSA stability under the main configuration and met the requested length constraint throughout, whereas the accessed Gemini configuration was lower and more variable and complied on length in only one of three runs. More importantly, perfect semantic similarity sometimes resulted from byte-identical repetition rather than from successful adaptation, and one Gemini response was reused across four unrelated topics. LSA scores also changed with representation settings, and unequal documented access tiers limit attribution to model families. The strongest conclusion is therefore methodological: response stability should be assessed by combining semantic preservation with instruction following, topical relevance, structural adaptation, repetition diagnostics, and repeated-run testing. Such a profile gives a more defensible account of reliability than a single similarity score or a vendor ranking.

Data And Analysis Availability Statement

The analytical corpus contains all 648 responses, parsed from the nine recorded model-run files, and every statistic reported here was recomputed from it in a single scripted pass. Reproducible materials include the parsed master dataset, block-level stability metrics, duplication checks, LSA sensitivity outputs, structural Variant D checks, editable table and figure source data, and high-resolution figures. The AI-assisted coding artefacts for position, reason, risk, and safeguard retention are incomplete and are not represented as fully reproducible materials. Public repository deposition remains to be completed by the authors if required by the journal.

Declarations

Ethical approval

The study analysed AI-generated textual responses and did not involve human participants, identifiable personal data, animals, or clinical interventions. Formal ethical approval was therefore not required.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research was self-funded by the authors. It received no grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

Idrees: conceptualisation, methodology, investigation, data curation, formal analysis, writing – original draft, writing – review and editing. Liu Yongzhi: conceptualisation, methodology, validation, supervision, writing – review and editing. Both authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors thank colleagues at the College of Foreign Languages and Cultures, Chengdu University of Technology, for discussion of the evaluation design.

Use of AI tools

The responses analysed in this study are the object of investigation rather than a writing aid. A structured AI-assisted workflow was used to code argumentative components, as described in Section 4.5, and its evidential limits are stated there. The authors take full responsibility for the content of the manuscript.

Data availability

The study data and analysis materials are held in the authors' working archive. They have not yet been deposited in a public repository. Repository details and any associated DOI will be added once a deposit has been made.

REFERENCES

1. Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. Annals of the New York Academy of Sciences, 1525(1), 140–146. https://doi.org/10.1111/nyas.15007
2. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 39, 1–45. https://doi.org/10.1145/3641289
3. Chatterjee, A., Renduchintala, H. S. V. N. S. K., Bhatia, S., & Chakraborty, T. (2024). POSIX: A prompt sensitivity index for large language models. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 14550–14565). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.852
4. Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843
5. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

6. Fu, T., & Barez, F. (2025). Same question, different words: A latent adversarial framework for prompt robustness. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 31464–31481). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1595
7. Fu, W., Wei, B., Hu, J., Cai, Z., & Liu, J. (2024). QGEval: Benchmarking multi-dimensional evaluation for question generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 11783–11803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.658
8. Gan, C., & Mori, T. (2023). Sensitivity and robustness of large language models to prompt template in Japanese text classification tasks. In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (pp. 1–11). Association for Computational Linguistics. https://aclanthology.org/2023.paclic-1.1/
9. Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K. D., Du, W., Durmus, E., Dušek, O., Emezue, C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., … Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. In Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (pp. 96–120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.10
10. Hämäläinen, M., & Alnajjar, K. (2021). Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (pp. 84–95). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.9
11. Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1(1), 77–89. https://doi.org/10.1080/19312450709336664
12. Hua, A., Tang, K., Gu, C., Gu, J., Wong, E., & Qin, Y. (2025). Flaw or artifact? Rethinking prompt sensitivity in evaluating LLMs. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 19889–19899). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1006
13. Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. Comprehensive Psychology, 3, Article 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1
14. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, Article 863. https://doi.org/10.3389/fpsyg.2013.00863
15. Leidinger, A., van Rooij, R., & Shutova, E. (2023). The language of prompting: What linguistic properties make a prompt successful? In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 9210–9232). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.618
16. Lior, G., Habba, E., Levy, S., Caciularu, A., & Stanovsky, G. (2025). ReliableEval: A recipe for stochastic LLM evaluation via method of moments. In Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics. https://aclanthology.org/2025.findings-emnlp.594/
17. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 2511–2522). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.153
18. Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8086–8098). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.556
19. Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? A call for multi-prompt LLM evaluation. Transactions of the Association for Computational Linguistics, 12, 933–949. https://doi.org/10.1162/tacl_a_00681
20. Moradi, M., & Samwald, M. (2021). Evaluating the robustness of neural language models to input perturbations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language

Processing (pp. 1558–1570). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.117

21. Nalbandyan, G., Shahbazyan, R., & Bakhturina, E. (2025). SCORE: Systematic consistency and robustness evaluation for large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Industry Track (pp. 470–484). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-industry.39
22. Qiang, Y., Nandi, S., Mehrabi, N., Ver Steeg, G., Kumar, A., Rumshisky, A., & Galstyan, A. (2024). Prompt perturbation consistency learning for robust language models. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 1357–1370). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-eacl.91
23. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410
24. Schmidtova, P., Mahamood, S., Balloccu, S., Dušek, O., Gatt, A., Gkatzia, D., Howcroft, D. M., Platek, O., & Sivaprasad, A. (2024). Automatic metrics in natural language generation: A survey of current evaluation practices. In Proceedings of the 17th International Natural Language Generation Conference (pp. 557–583). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.inlg-main.44
25. Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting. In International Conference on Learning Representations. https://openreview.net/forum?id=RIu5lyNXjT
26. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
27. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaekermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., … Natarajan, V. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 31(3), 943–950. https://doi.org/10.1038/s41591-024-03423-7
28. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. https://openreview.net/forum?id=uyTL5Bvosj
29. Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Characterizing the confidence of large language model-based automatic evaluation metrics. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 76–89). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-short.9
30. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29, 1930–1940. https://doi.org/10.1038/s41591-023-02448-8
31. Vu, T., Krishna, K., Alzubi, S., Tar, C., Faruqui, M., & Sung, Y.-H. (2024). Foundational autoraters: Taming large language models for better automatic evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 17086–17105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.949
32. Wahle, J. P., Ruas, T., Xu, Y., & Gipp, B. (2024). Paraphrase types elicit prompt engineering capabilities. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 11004–11033). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.617
33. Wang, Y., Chen, L., Cai, S., Xu, Z., & Zhao, Y. (2024). Revisiting automated evaluation for long-form table question answering. In Proceedings of the 2024 Conference on Empirical Methods in Natural

Language Processing (pp. 14663–14677). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.815

34. Watts, I., Gumma, V., Yadavalli, A., Seshadri, V., Swaminathan, M., & Sitaram, S. (2024). PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 7900–7932). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.451
35. Webson, A., & Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2300–2344). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.167
36. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations. https://openreview.net/forum?id=SkeHuCVFDr
37. Zhelezniak, V., Savkov, A., Shen, A., & Hammerla, N. (2019). Correlation coefficients and semantic textual similarity. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 951–962). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1100
38. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track. https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html
39. Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-following evaluation for large language models. arXiv. https://arxiv.org/abs/2311.07911
40. Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). ProSA: Assessing and understanding the prompt sensitivity of LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 1950–1976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.108

Appendix A. Complete Prompt Set

All 648 responses were elicited with the 72 prompts reproduced below: twelve topics, each expressed through six variants. Every prompt consists of a question stem followed by a fixed instruction block. To avoid repeating the block 72 times, it is stated once here and only the stems are listed.

Instruction block for variants A, B, C, E, and F: "Respond in 180 to 250 words. State your position, give two reasons, identify one risk, and propose one safeguard."

Instruction block for Variant D: "Identify one risk and propose one safeguard. Then state your position and give two reasons. Keep your response between 180 and 250 words."

Each full prompt is the concatenation of the stem and the applicable instruction block. The Variant D stem is character-identical to the Variant A stem in every topic, so the two differ only in the instruction block.

Table A1. Topic index

No.	Topic	Domain
1	**Public-Facing Policy Summaries (LLMs)**	Public-sector AI governance
2	**Mandatory Human Review of AI Public Information**	Public-sector AI governance
3	**Conversational AI for Citizen Enquiries**	Public-sector AI governance

4	**Student Disclosure of Generative AI in Academic Work**	AI in higher education
5	**Institution-Wide Access to Generative AI in Universities**	AI in higher education
6	**Compulsory AI Literacy in Undergraduate Education**	AI in higher education
7	**Restricting Private Car Use in Air-Polluted Zones**	Environmental policy
8	**Government Subsidies for Household Solar Panels**	Environmental policy
9	**Taxation on Single-Use Plastics**	Environmental policy
10	**Machine-Readable Local Government Budget Data**	Transparency and public administration
11	**Identity Verification for Political Advertising**	Transparency and public administration
12	**Automated Systems for Public Service Eligibility**	Transparency and public administration

Note. Topics 1–6 concern the governance and use of AI systems; topics 7–9 concern environmental policy; topics 10–12 concern transparency and administrative automation. The set was chosen so that every prompt invites an argumentative response rather than a factual short answer.

Topic 1. Public-Facing Policy Summaries (LLMs)

A — baseline. Should government agencies use large language models to draft public-facing policy summaries?

B — plain-language paraphrase. Should public authorities use large language models to prepare policy summaries for citizens?

C — interrogative reframing. Do you think large language models should be used by government agencies when drafting policy summaries for the public?

D — order manipulation (stem as A). Should government agencies use large language models to draft public-facing policy summaries?

E — formal register. Is it appropriate for public-sector institutions to employ large language models in the preparation of policy summaries intended for public audiences?

F — topicalised / passive. Should the drafting of public-facing policy summaries by government agencies involve the use of large language models?

Topic 2. Mandatory Human Review of AI Public Information

A — baseline. Should every piece of AI-generated public-service information receive human review before publication?

B — plain-language paraphrase. Should all public-service information produced by AI be checked by a person before it is released?

C — interrogative reframing. Do you think a human should review every item of AI-generated public-service information before the information is published?

D — order manipulation (stem as A). Should every piece of AI-generated public-service information receive human review before publication?

E — formal register. Should public institutions mandate human verification of all AI-generated public-service information prior to its publication?

F — topicalised / passive. Should human review be required before any AI-generated public-service information is published?

Topic 3. Conversational AI for Citizen Enquiries

A — baseline. Should public agencies use conversational AI for routine citizen enquiries?

B — plain-language paraphrase. Should government offices use AI chat systems to answer common questions from citizens?

C — interrogative reframing. Do you think routine enquiries from citizens should be handled by conversational AI used by public agencies?

D — order manipulation (stem as A). Should public agencies use conversational AI for routine citizen enquiries?

E — formal register. Should public-sector bodies deploy conversational artificial intelligence to manage routine enquiries from members of the public?

F — topicalised / passive. Should the handling of routine citizen enquiries by public agencies be assigned to conversational AI?

Topic 4. Student Disclosure of Generative AI in Academic Work

A — baseline. Should universities require students to disclose the use of generative AI in assessed work?

B — plain-language paraphrase. Should higher education institutions ask students to report their use of generative AI in graded assignments?

C — interrogative reframing. Do you think students should be required by universities to disclose when they use generative AI in assessed work?

D — order manipulation (stem as A). Should universities require students to disclose the use of generative AI in assessed work?

E — formal register. Should higher education institutions establish a formal requirement for students to declare any use of generative AI in assessed academic work?

F — topicalised / passive. Should disclosure of generative AI use in assessed work be required from university students?

Topic 5. Institution-Wide Access to Generative AI in Universities

A — baseline. Should universities provide institution-wide access to generative AI tools?

B — plain-language paraphrase. Should higher education institutions give all students and staff access to generative AI systems?

C — interrogative reframing. Do you think access to generative AI tools should be provided across the whole university?

D — order manipulation (stem as A). Should universities provide institution-wide access to generative AI tools?

E — formal register. Should universities implement institution-wide provision of generative artificial intelligence tools for their academic communities?

F — topicalised / passive. Should access to generative AI tools be made available throughout universities?

Topic 6. Compulsory AI Literacy in Undergraduate Education

A — baseline. Should AI literacy become a compulsory part of undergraduate education?

B — plain-language paraphrase. Should all undergraduate students be required to learn how to understand and use AI responsibly?

C — interrogative reframing. Do you think undergraduate education should include AI literacy as a compulsory component?

D — order manipulation (stem as A). Should AI literacy become a compulsory part of undergraduate education?

E — formal register. Should higher education institutions incorporate mandatory artificial intelligence literacy instruction into undergraduate curricula?

F — topicalised / passive. Should the inclusion of AI literacy in undergraduate education be made compulsory?

Topic 7. Restricting Private Car Use in Air-Polluted Zones

A — baseline. Should cities restrict private car use in areas with persistent air pollution?

B — plain-language paraphrase. Should urban authorities limit the use of personal vehicles in places with continuing air-quality problems?

C — interrogative reframing. Do you think private car use should be restricted by cities in areas where air pollution remains persistent?

D — order manipulation (stem as A). Should cities restrict private car use in areas with persistent air pollution?

E — formal register. Should municipal governments impose restrictions on private vehicle use in districts experiencing persistent air pollution?

F — topicalised / passive. Should restrictions on private car use be introduced in urban areas affected by persistent air pollution?

Topic 8. Government Subsidies for Household Solar Panels

A — baseline. Should governments subsidize household solar-panel installation?

B — plain-language paraphrase. Should public funds help households pay for installing solar panels?

C — interrogative reframing. Do you think the installation of household solar panels should receive financial support from governments?

D — order manipulation (stem as A). Should governments subsidize household solar-panel installation?

E — formal register. Should governments provide financial subsidies to support the installation of residential solar-energy systems?

F — topicalised / passive. Should household solar-panel installation be supported through government subsidies?

Topic 9. Taxation on Single-Use Plastics

A — baseline. Should governments impose higher taxes on single-use plastics?

B — plain-language paraphrase. Should public authorities charge higher taxes on disposable plastic products?

C — interrogative reframing. Do you think single-use plastics should be subject to higher taxation by governments?

D — order manipulation (stem as A). Should governments impose higher taxes on single-use plastics?

E — formal register. Should governments introduce increased taxation on single-use plastic products as an environmental policy measure?

F — topicalised / passive. Should higher taxes be imposed on single-use plastics by governments?

Topic 10. Machine-Readable Local Government Budget Data

A — baseline. Should local governments publish budget data in machine-readable formats?

B — plain-language paraphrase. Should municipal authorities release financial budget information in formats that computers can process?

C — interrogative reframing. Do you think budget data should be published by local governments in machine-readable formats?

D — order manipulation (stem as A). Should local governments publish budget data in machine-readable formats?

E — formal register. Should local government bodies be required to release public budget data in standardized, machine-readable formats?

F — topicalised / passive. Should the publication of local-government budget data in machine-readable formats be required?

Topic 11. Identity Verification for Political Advertising

A — baseline. Should social-media platforms verify the identity of people who purchase political advertising?

B — plain-language paraphrase. Should social networks confirm who is paying for political advertisements?

C — interrogative reframing. Do you think the identity of anyone buying political advertising should be verified by social-media platforms?

D — order manipulation (stem as A). Should social-media platforms verify the identity of people who purchase political advertising?

E — formal register. Should social-media companies implement mandatory identity verification for individuals or organizations purchasing political advertisements?

F — topicalised / passive. Should identity verification be required for purchasers of political advertising on social-media platforms?

Topic 12. Automated Systems for Public Service Eligibility

A — baseline. Should automated systems handle routine eligibility enquiries for public services?

B — plain-language paraphrase. Should automated tools answer common questions about whether people qualify for public services?

C — interrogative reframing. Do you think routine questions about eligibility for public services should be handled by automated systems?

D — order manipulation (stem as A). Should automated systems handle routine eligibility enquiries for public services?

E — formal register. Should public-service institutions deploy automated systems to manage routine enquiries concerning service eligibility?

F — topicalised / passive. Should the handling of routine public-service eligibility enquiries be assigned to automated systems?

Abbreviations

BERT, bidirectional encoder representations from transformers; CI, confidence interval; LLM, large language model; LSA, latent semantic analysis; NLG, natural language generation; SD, standard deviation; SVD, singular value decomposition; TF-IDF, term frequency–inverse document frequency.