Observer-Side Diagnosis of Prompt-Induced Interference in Large Language Models: A Macro-Group Vocabulary and Targeted Cross-Lingual Stress Tests

Anonymous authors
Paper under double-blind review

Abstract

Small prompt fragments can change not only an intended surface property of a response (e.g., style), but also secondary reliability-relevant behaviors such as epistemic commitment, scope of alternatives, and reasoning presentation. Such interaction-level behavioral shifts are difficult to characterize with conventional task-level prompt evaluation. This study proposes an *observer-side* diagnostic vocabulary for describing such prompt-induced interference in large language models (LLMs). The vocabulary organizes prompt effects into four macro-groups—framing (role/task/audience/objective), reasoning (process/scope), expression (style/format/length), and epistemic control (stance/constraints)—and is instantiated here as the Z-model, an auditable 11-axis reference basis for reporting and comparison under black-box access. This reference basis is a pragmatic descriptive choice for interpretability and reporting, rather than an ontological, minimality, or model-internal claim. Empirically, we do *not* attempt to validate all four macro-groups at once. Instead, we run a targeted Japanese/English stress test of one high-leverage pathway: an expression-oriented politeness cue and its secondary effects on epistemic- and scope-related proxies. Under a matched interaction protocol (five benign topics; 250 samples per language-condition), the same politeness cue reliably changes expression while redistributing uncertainty and alternative/conditional markers in language-dependent ways. These are interpreted as protocol-level effects, and potential confounds from model training and alignment differences across languages are explicitly discussed. As a prediction-to-observation check, we additionally run a small factorial 2×2 probe and observe localized non-additivity consistent with structured latent interference. Key directional patterns are also reproduced on a pinned open-weight model checkpoint. Overall, the contribution is a scoped diagnostic framework plus evidence that one targeted cross-group pathway can be made auditable with lightweight black-box probes; the inverse direction is presented as a post-hoc diagnostic workflow rather than a validated latent estimator.

1 Introduction

As large language models (LLMs) become integrated into high-stakes communication and decision-support settings, seemingly minor prompt variations can induce reliability-relevant shifts in model behavior. These shifts often extend beyond stylistic variation, affecting epistemic stance, conceptual scope, reasoning dynamics, and the interactional posture of responses. Such sensitivity is particularly salient in Japanese, where sentence-final morphology and stance marking compactly encode pragmatic signals without altering propositional content (Gan & Mori, 2023; Mikami et al., 2025; Yin et al., 2024). Small changes in epistemic framing or politeness morphology can systematically move responses from tentative, conditionally reasoned outputs to assertive or prematurely closed conclusions.

Scope and terminology. This study investigates prompt sensitivity as an interaction-level phenomenon under black-box access. We therefore treat prompt–response interactions as observable behavioral traces of a black-box LLM system, and introduce a lightweight diagnostic vocabulary for systematically probing these interaction-level effects. The term *latent configuration* is used in an observer-side sense: it denotes unobserved degrees of freedom that summarize how a prompt steers model behavior, without making claims about model-internal representations or causal mechanisms. Empirical evidence is therefore stated as within-protocol, distribution-level deltas in observable output proxy features under controlled A/B perturbations. Although lightweight observable proxies are employed, robustness analyses under indicator-set reduction (Appendix G) show that the core interference signatures remain directionally stable under coarsening.

For reference, the compact observer-side latent block used throughout the paper is

$$p \in \mathcal{P}, \qquad \boldsymbol{z}(p) = f_{\text{enc}}(p) \in \mathbb{R}^{11}, \qquad y \sim q(\cdot \mid \boldsymbol{z}(p)), \qquad r(y) \in \mathbb{R}^m, \tag{1}$$

where p is a prompt, $\boldsymbol{z}(p)$ is the latent configuration, y is a sampled output, and $r(y)$ denotes the observable indicator vector used for diagnosis. This notation is descriptive and observer-side: it does not assume that $\boldsymbol{z}$ is an identifiable model-internal state.

Much of the existing prompt-sensitivity literature evaluates how prompt variation affects task-level performance, correctness, or output quality (Liu et al., 2023; White et al., 2023; Sclar et al., 2024; Razavi et al., 2025; Zhuo et al., 2024). This paper instead studies interaction-level behavioral redistribution under prompt variation: how a local prompt fragment can induce coupled shifts across multiple response dimensions even when the intended task and propositional content remain fixed. We refer to this phenomenon as *latent interference*. The Z-model is introduced not as a static taxonomy of prompt attributes, but as an observer-side diagnostic vocabulary for describing and auditing such coupled shifts under black-box access.

Japanese provides a revealing stress-test setting for this question. Subject omission, stance marking, and politeness bundle pragmatic signals into compact surface forms, making cross-dimensional coupling easier to observe but harder to interpret systematically. We therefore use Japanese not as a language-specific curiosity, but as a diagnostic lens for interaction-level latent interference. Cross-lingual differences are interpreted as *protocol-level* signals rather than clean typological effects, since training and alignment differences across languages may also contribute.

To analyze such coupling under black-box access, we introduce a *macro-group* diagnostic vocabulary for prompt-induced instruction conflicts: **Framing**, **Reasoning**, **Expression**, and **Epistemic control**. This view is operationalized by the **Z-model**, an observer-side latent coordinate system for interaction-level behavior configuration, instantiated here as an auditable *11-axis reference basis*. The choice of 11 axes is pragmatic rather than ontological: axes may be merged, refined, or extended without changing the central notion of latent interference. The vocabulary supports two uses: a *forward* view, which anticipates directional shifts in observables from a prompt fragment, and an *inverse* diagnostic workflow, which organizes post-hoc explanations from outputs back to plausible latent dimensions. We do not claim that the inverse direction uniquely recovers or validates latent states. The empirical stress test below deliberately targets one high-leverage Expression→Epistemic/Scope pathway rather than surveying the full Z-space.

To make this interaction-level configuration concrete, consider two minimal Japanese prompts that differ by a single fragment:

(A)**「慎重に検討してください。」**
(B)**「慎重に検討して説明してください。結論は正しいですよね？」**

Although (B) is linguistically redundant and uncommon as a natural prompt, it makes explicit an expectation of correctness and justification that is often left implicit in (A). The point is not that (B) is a better prompt, but that adding a single fragment can redistribute multiple observable behaviors—e.g., uncertainty markers, alternatives, and conditional framing. Appendix A.5 reports a minimal A/B case study of this pattern, together with a lightweight LLM-as-a-Judge directional check aligned to the proxy definitions (Table A.4).

This study does not propose a new prompting algorithm, training method, alignment technique, or performance benchmark. Instead, it provides an interpretable diagnostic vocabulary and a minimal probing protocol for reasoning about interaction-level reliability without assuming access to model internals. The contributions

are intentionally split into *conceptual* and *empirical* parts so that the evidence-bearing claims remain clearly scoped.

Conceptual contributions.

(C1) **Latent interference + diagnostic workflow.** Systematic secondary shifts outside an intended macro-group are formalized as *latent interference*, together with a forward prediction view and an inverse *diagnostic* workflow under black-box access.
(C2) **Macro-group vocabulary.** Framing / Reasoning / Expression / Epistemic control is proposed as a compact scaffold for describing typical instruction conflicts and their cross-group coupling in LLM prompting, oriented toward diagnosing interaction-level behavioral structure rather than statically classifying prompts.
(C3) **Z-model instantiation.** An auditable instantiation of this scaffold is provided as an 11-axis observer-side reference basis, intended as a reporting and diagnosis vocabulary rather than an ontology, minimality claim, or model-internal theory.

Empirical contributions.

(E1) **Targeted cross-lingual stress test.** Under a matched Japanese/English interaction protocol, we test one high-leverage pathway in which an Expression-oriented politeness cue is expected to induce secondary shifts in epistemic- and scope-related proxies. The resulting redistribution signatures are interpreted as *protocol-level* effects.
(E2) **Prediction-to-observation check.** A factorial non-additivity probe is included as a focused check that interaction effects localize to macro-group-coupled proxies, consistent with structured interference on the targeted pathway.
(E3) **Reporting-level robustness.** Key directional patterns are reproduced on a pinned open-weight model checkpoint, and basis-reduction/coarsening plus indicator-ablation analyses justify the retained reporting resolution.

Roadmap. Section 2 reviews prompt sensitivity and instruction-conflict work. Section 3 introduces the macro-group scaffold and the Z-model instantiation. Section 3.4-3.5 presents the minimal cross-lingual probing protocol and results. Section 4 presents the analysis, implications, and limitations.

An open-weight reproduction on a pinned checkpoint is additionally reported in Appendix F.

2 Background and Related Work

This paper connects four adjacent literatures: prompt engineering, instruction-conflict benchmarking, calibration and hallucination, and language-specific prompting. The review below focuses on results that help explain how prompt variation, alignment constraints, and pragmatic form shape observable response behavior. Rather than re-surveying prompt optimization broadly, the goal is to situate the present interaction-level diagnostic perspective within these neighboring lines of work.

2.1 Prompt Sensitivity and Behavioral Evaluation

Recent studies have shown that LLM outputs can be highly sensitive to seemingly minor prompt variations, including wording, formatting, role framing, and reasoning instructions. Prompting work such as Chain-of-Thought prompting demonstrates that small structural changes can substantially alter reasoning behavior and task outcomes (Wei et al., 2022; Kojima et al., 2022). More broadly, behavioral evaluation in NLP has emphasized that models should be assessed not only by aggregate end-task accuracy but also through controlled perturbation-based probes. For example, CheckList introduced capability-oriented behavioral testing templates for systematically surfacing model weaknesses under targeted perturbations (Ribeiro et al., 2020).

Taken together, prior studies show that prompt variation can materially affect model behavior, typically assessed through task outcomes, correctness, or output quality (Sclar et al., 2024; Zhuo et al., 2024; Razavi et al., 2025). The present work builds on this literature but shifts the unit of analysis from task-level

outcomes to interaction-level behavioral redistribution under controlled A/B perturbations. The coupled cross-dimensional effects of interest are later formalized as *latent interference*.

2.2 Instruction Tuning, Alignment, and Instruction Conflicts

Instruction tuning and human-feedback-based alignment fundamentally shape how LLMs respond to prompts. InstructGPT demonstrated that fine-tuning with human demonstrations and reinforcement learning from human feedback (RLHF) yields models that follow instructions more reliably and are preferred by human evaluators (Ouyang et al., 2022). The Flan collection further systematized instruction tuning data and prompt-format mixtures, improving generalization and reducing the need for downstream fine-tuning (Chung et al., 2022).

A practical consequence of alignment is that prompts often encode *multiple constraints* (style, safety, scope, hedging, and format) that can become partially conflicting in realistic workflows. Recent evaluation work has explicitly focused on this regime: He et al. (2025) proposed ConInstruct to benchmark conflict detection and resolution when instructions contain incompatible constraints, highlighting that models may detect conflicts yet fail to surface them to users. This setting is closely related to the notion of latent interference in the Z-model: a single prompt fragment (e.g., politeness or safety cues) can unintentionally reweight other behavioral objectives, changing the resulting epistemic presentation, scope, or reasoning depth.

While related, *latent interference* is distinguished from classic instruction-conflict settings along three axes. First, instruction-conflict benchmarks typically define conflict as two or more explicitly stated constraints that may be mutually incompatible; latent interference is defined at the fragment level via systematic secondary shifts (Eq. (4)) even in the absence of an explicit conflict. Second, conflict resolution asks whether a model detects, surfaces, or prioritizes constraints, whereas our diagnostic goal is to characterize cross-dimensional *redistribution* in observable indicators under matched A/B perturbations. Third, the Z-model makes a composition-level prediction: fragments can combine non-additively in structured ways, operationalized by nonzero interaction effects ($d_I \neq 0$) in the factorial probe (Section 3.4).

2.3 Calibration, Hallucination, and Epistemic Presentation

A growing body of work examines calibration and uncertainty properties of modern LLMs, analyzing how pretraining and alignment stages affect confidence reliability and proposing output-level methods such as post-hoc calibration, uncertainty-aware scoring, or multi-sample aggregation (Geng et al., 2024; Desai & Durrett, 2020; Kadavath et al., 2022). In parallel, hallucination, fluent but factually incorrect or ungrounded output, remains a major barrier to deployment. Surveys have distinguished factuality from faithfulness hallucinations and reviewed detection and mitigation approaches, including retrieval-augmented generation, self-consistency, and verifier models (Ji et al., 2023; Huang et al., 2025; Tonmoy et al., 2024).

The Z-model is complementary in that it does not propose a new calibration or hallucination-mitigation algorithm. Instead, it introduces a latent-variable vocabulary, particularly via $z_{\text{epistemic}}$ and z_{scope}, to reason about how prompts configure epistemic stance and coverage *before* a post-hoc method is applied. This distinction matters for interaction-level reliability: even when factual accuracy is unchanged, changes in *epistemic presentation* (e.g., hedging versus assertiveness, responsibility framing, and refusal versus compliance) can substantially alter downstream user trust and risk.

2.4 Pragmatics, Language-Specific Prompting, and Japanese as a Stress Test

Most prompt engineering and alignment research implicitly assumes English as the primary language; however, prompting behavior is strongly language- and pragmatics-dependent. A recent survey of pragmatic evaluation resources highlighted the breadth of pragmatic phenomena relevant to modern LLMs (e.g., implicature, reference, modality, and social meaning) and the challenges of measuring them reliably (Ma et al., 2025). Cross-lingual studies on politeness have further shown that the effect of prompt politeness on LLM performance differs across languages, including English, Chinese, and Japanese (Yin et al., 2024).

For Japanese specifically, Gan & Mori (2023) analyze sensitivity to prompt templates in Japanese text classification and report large performance swings under small template perturbations. Mikami et al. (2025) constructed a Japanese natural language interface (NLI) dataset for comparatives and observed that model performance can be sensitive to prompt format in zero-shot settings and to example properties in few-shot settings, with difficulties on Japanese-specific phenomena.

Collectively, these studies demonstrated that Japanese prompts can be unusually fragile and that politeness, sentence structure, and other pragmatic-linguistic properties strongly modulate model behavior. However, many previous studies have reported empirical deltas without an interaction-level latent account of why certain Japanese fragments (e.g., sentence-final epistemic control or politeness markers) can simultaneously affect style, epistemic stance, reasoning depth, and scope. The Z-model directly addresses this gap. Japanese is treated as a typological stress test that makes latent coupling visible and exposes interference effects that are often obscured in English-centric analyses.

Linguistic grounding (Japanese). The Japanese-specific mechanisms discussed in this study are not introduced as ad-hoc properties of LLM outputs, but are grounded in well-established descriptions of Japanese grammar and pragmatics. Japanese is discourse- and topic-oriented, with pervasive argument omission and heavy reliance on contextual inference, as well as a rich system of politeness, honorifics, and sentence-final modality that jointly shape social stance and epistemic commitment. These linguistic properties motivate the treatment of Japanese prompt fragments as a stress test for prompt-induced coupling, helping to explain why minimal surface perturbations can systematically cascade into multiple behavioral dimensions.

Positioning. Prior prompt-engineering studies provided catalogs of patterns and best practices, often evaluated by task success. The focus of this study is complementary: prompts are characterized as interaction-level interventions that redistribute multiple behavioral degrees of freedom (e.g., epistemic stance and scope), and cross-dimensional side effects ("latent interference") measurable via paired A/B probing are emphasized. This study also distinguishes its setting from factuality or calibration benchmarks: it does not claim to measure truthfulness directly; instead, *epistemic presentation* and *responsibility framing* are measured as observable correlates that can influence downstream reliability judgments.

3 Z-model: Latent Dimensions for Prompt-Induced Behavior

As illustrated by the introductory example, prompt fragments often induce effects that cannot be attributed to a single control axis. Instead, small surface variations can trigger coupled shifts across stylistic, epistemic, and scope-related behaviors. The Z-model is introduced to capture and reason about such interactions in a systematic and operational manner. LLMs do not merely respond to the propositional content of prompts. During inference, prompts implicitly configure a range of behavioral properties, such as persona, reasoning mode, epistemic stance, and conceptual breadth, that are systematically expressed in generated outputs but are not directly observable.

This section introduces the Z-model, an interpretable finite-dimensional latent coordinate vocabulary (instantiated here with 11 reference dimensions) for reasoning about such prompt-induced behavior. Here, "latent" denotes an analytical abstraction for describing prompt-induced behavioral tendencies at the interaction level, without claiming identifiable internal variables or modules. The 11 dimensions of $\mathbf{Z}$ are not proposed as an exhaustive ontology, but as a compact coordinate system sufficient to (i) factor commonly co-occurring instructional intents, and (ii) make cross-dimensional interference empirically observable at the output level. Concretely, the Z-model organizes prompt effects along a set of behavioral dimensions, including role, task interpretation, audience assumption, stylistic register, output structure, verbosity, reasoning process, scope of coverage, epistemic stance, constraints, and implicit objectives. These dimensions are introduced to support the analysis and interpretation, rather than for estimation or control. The dimensions are not assumed to be independent, nor directly measurable or optimizable.

Throughout this study, the term *latent* is used in an analytical rather than ontological sense. This study does not claim that the dimensions of $\mathbf{Z}$ correspond to identifiable or disentangled internal variables of the model.

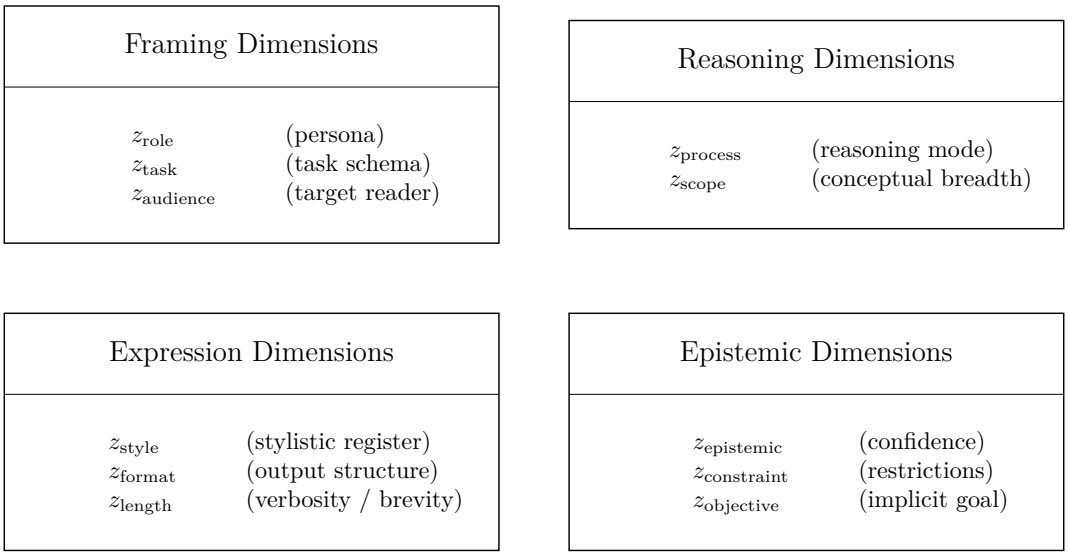

Figure 1: Z-model latent dimensions and their organization into four macro-groups (framing, reasoning, expression, and epistemic control). Each dimension represents an observer-side behavioral degree of freedom implicitly configured by a prompt during interaction, rather than a model-internal variable. The finite-dimensional reference basis (instantiated here with 11 dimensions) is introduced for interpretability, auditability, and operational comparison under black-box access, rather than as a claim of completeness or ontological structure. The dimensions are not assumed to be independent or orthogonal; apparent overlap reflects empirically recurrent sources of cross-dimensional coupling, which are central to the notion of latent interference. This figure represents an analytical coordinate system for reasoning about prompt-induced behavioral shifts, rather than a taxonomy of prompt types or a decomposition of internal model representations.

Instead, **Z** serves as an interpretive coordinate system for describing regularities and interference patterns observable in model behavior.

Notation. When directional notation (↑ / ↓) is used for $z_{\text{epistemic}}$, $z_{\text{epistemic}}$ ↑ denotes higher epistemic commitment (more categorical or definitive statements), whereas $z_{\text{epistemic}}$ ↓ denotes lower commitment (more explicit uncertainty and hedging).

Linguistic cues rarely affect a single dimension in isolation. The notion of *latent interference* is therefore introduced: a prompt fragment intended to modulate one dimension can systematically perturb others. As shown in the following text, this phenomenon is particularly salient in Japanese, where linguistic properties, such as subject omission, pragmatically overloaded modifiers, and politeness–epistemic coupling create structured coupled shifts.

The remainder of this section formalizes the Z-model and its dimensions, providing a shared vocabulary for analyzing prompt–behavior interactions before moving to forward analysis and inverse diagnosis in subsequent sections. The contribution is not the general idea of forward/inverse framing itself, but the observation that prompt control is fundamentally non-separable: small linguistic fragments induce structured, reproducible coupling across behavioral dimensions, which becomes particularly measurable in Japanese. A concrete worked example illustrating how the Z-model supports forward prediction and post-hoc inverse diagnosis is provided in Appendix D.

Figure 1 summarizes the four macro-groups and the 11-axis reference-basis instantiation used throughout the study.

3.1 Overview of the Z-model and compact latent block

An explicit finite-dimensional basis is required not for completeness, but to discuss cross-dimensional coupling, interference, and comparison operationally under black-box access. The Z-model is formalized as a latent coordinate system for describing behavioral degrees of freedom induced by prompts, intended for analysis rather than statistical estimation.

Terminological clarification. The term "coordinate system" is used here in a representational, observer-side sense. It denotes a structured descriptive scheme for organizing directional tendencies in observable output space, without presupposing metric structure, linear embedding, or latent identifiability in a geometric sense. The ambient notation $\mathbb{R}^{11}$ is adopted solely to indicate finite dimensionality and bookkeeping convenience.

Vector notation (e.g., Δz) expresses directional changes in proxy-aligned dimensions and should not be interpreted as evidence for an underlying continuous latent geometry or internal embedding space. In this sense, the Z-model functions as a structured diagnostic vocabulary instantiated as a reference-basis representation, rather than as a statistical or geometric model of internal states.

Let

$$\boldsymbol{z} = (\boldsymbol{z}_{\text{role}}, \boldsymbol{z}_{\text{task}}, \boldsymbol{z}_{\text{audience}}, \boldsymbol{z}_{\text{style}}, \boldsymbol{z}_{\text{format}}, \boldsymbol{z}_{\text{length}}, \boldsymbol{z}_{\text{process}}, \boldsymbol{z}_{\text{scope}}, \boldsymbol{z}_{\text{epistemic}}, \boldsymbol{z}_{\text{constraint}}, \boldsymbol{z}_{\text{objective}}) \in \mathcal{Z} \subset \mathbb{R}^{11}. \tag{2}$$

where each coordinate represents a distinct behavioral degree of freedom implicitly configured by a prompt. Here, $\mathbb{R}^{11}$ is used purely as a notational convenience to indicate dimensionality, rather than to imply metric structure, continuity, or statistical estimability.

Equation (1) makes explicit the compact observer-side latent block assumed throughout: prompts configure $\boldsymbol{z}$, $\boldsymbol{z}$ biases the distribution over outputs, and only observable indicators $r(y)$ are measured. All empirical claims in this paper are therefore made at the level of output distributions or proxy expectations under controlled perturbations, not at the level of single-output determinism or identifiable model-internal variables.

The choice of 11 dimensions follows the four macro-groups in Figure 1 (framing, reasoning, expression, and epistemic control), which were sufficient to describe the recurrent degrees of freedom surfaced by the Japanese case studies and the minimal probing protocol. This set is intended as a compact basis for analysis: future work may merge, refine, or extend dimensions without changing the definition of latent interference.

Japanese is used in this study as a typological stress test: because stance and politeness are often encoded morphosyntactically, a single Expression-level cue can more readily couple into Epistemic-control and scope-related shifts that are visible under lightweight, black-box observables. Analogous (though often less salient) coupling is expected in other languages.

3.2 Definition of the 11 Latent Dimensions

(1) $z_{\textbf{role}}$—**Role / Persona.** The adopted persona, social stance, and perceived epistemic authority (e.g., expert, tutor, critic, and narrator), shaping perspective-taking, expected domain knowledge, and pragmatic commitments.
(2) $z_{\textbf{task}}$—**Task Schema.** The task schema inferred from the prompt (e.g., summarization, explanation, critique, and rewriting), that determines high-level organization and normative success criteria.
(3) $z_{\textbf{audience}}$—**Target Audience.** Assumptions about the expertise and informational needs of the reader, modulating vocabulary choice and explanatory depth.
(4) $z_{\textbf{style}}$—**Stylistic Register.** Stylistic register, including politeness level, formality, affect, and conversational tone. In Japanese, this dimension is linguistically dense due to honorific morphology.
(5) $z_{\textbf{format}}$—**Output Structure.** Structural organization of the output, such as bullet points, tables, JSON, or code blocks.
(6) $z_{\textbf{length}}$—**Verbosity / Brevity.** The expected response length, which often constrains other dimensions.
(7) $z_{\textbf{process}}$—**Reasoning Process.** The mode and shape of reasoning expressed in the output, including stepwise justification and contrastive reasoning.

Table 1: 11 latent dimensions of the Z-model (core definitions).

Dimension	Description (concise)
z_{role}	Adopted persona or social identity (e.g., expert, tutor, critic), influencing perspective-taking and epistemic authority.
z_{task}	Internal task schema inferred by the model (e.g., summarization, explanation, critique, translation).
z_{audience}	Assumed target reader and expertise level, modulating terminology choice and explanatory depth.
z_{style}	Stylistic register including politeness, formality, and affect; pragmatically dense in Japanese.
z_{format}	Structural organization of the output, such as bullet points, tables, JSON, or stepwise formats.
z_{length}	Expected verbosity or brevity of the response, often constraining other latent dimensions.
z_{process}	Reasoning structure, including step-by-step, contrastive, or outline-first reasoning.
z_{scope}	Breadth of conceptual coverage, including the presence of alternatives and multiple perspectives.
$z_{\text{epistemic}}$	Epistemic stance expressed by the model, including confidence, hedging, and acknowledgment of uncertainty.
$z_{\text{constraint}}$	Explicit restrictions or prohibitions imposed by the prompt, such as forbidden content or stylistic bans.
$z_{\text{objective}}$	Implicit optimization objective pursued by the model (e.g., accuracy, creativity, persuasion, or critique).

(8) $z_{\mathbf{scope}}$—**Conceptual Breadth / Coverage.** How broadly the response explores the conceptual space, including alternatives and multiple perspectives.

(9) $z_{\mathbf{epistemic}}$—**Epistemic Stance.** Confidence, hedging, uncertainty expression, and willingness to acknowledge ignorance.

(10) $z_{\mathbf{constraint}}$—**Constraints and Prohibitions.** Explicit restrictions, such as forbidden content or stylistic bans.

(11) $z_{\mathbf{objective}}$—**Implicit Optimization Objective.** The implicit quality criterion pursued (e.g., accuracy, creativity, and persuasion).

Table 1 lists the core definitions of the 11 latent dimensions used throughout this study.

Together, these dimensions form a reference-based representation through which prompt–model interactions can be organized analytically. Importantly, they are not independent; linguistic cues frequently activate multiple dimensions simultaneously. This property is central to the analysis of Japanese prompts.

Why retain the full 11-axis reference-basis instantiation? The *conceptual* core of the diagnostic vocabulary comprises the four macro-groups in Figure 1. Nevertheless, this study presents an explicit 11-axis reference-basis instantiation as the *reference basis* for three pragmatic reasons. First, **coverage**: in real prompting, many fragments primarily target Framing or implicit objectives (e.g., role/audience constraints, persuasion vs. critique), and secondary effects often propagate into Reasoning and Epistemic control; without named axes in the reference basis, such patterns are hard to report consistently. Second, **auditability**: a fixed reference basis makes primary vs. secondary shifts explicit and comparable across studies, prompts, and model snapshots under black-box access. Third, **modularity**: the same analysis can be reported at multiple resolutions, including macro-group coarsenings or probe-dependent reductions/rotations (see Figure 2 and Appendix H).

We therefore do *not* claim that "11" is minimal or unique. It is an auditable instantiation that supports consistent diagnosis and reporting; depending on the observable proxy family, lower-dimensional subspaces may already be resolution-sufficient, and we report such probe-dependent reductions where appropriate.

Dimension justification via basis reduction and ablation (relative sufficiency). To complement the conceptual motivation above, we perform a lightweight *basis reduction* check in a controlled orthogonal design. Because the 11 axes are introduced as an observer-side reporting basis rather than as learned latent factors, dimensionality is justified here by predictive sufficiency under basis reduction, 11D coarsening, basis rotation, and indicator-set ablation—not by claiming a unique PCA/factor-analysis optimum. We treat the observer-side coordinate assignment as a fully specified design matrix: each condition is defined by explicit prompt constraints along four dimensions ($z_{\text{epistemic}}, z_{\text{scope}}, z_{\text{format}}, z_{\text{length}}$), encoded as ± 1. We use a balanced 2^4 design (16 conditions) and sample $n = 50$ independent generations per condition under fixed decoding settings. For each condition, we compute a proxy-delta vector relative to a fixed baseline and fit a simple ridge regression to predict this proxy-level "signature" from the design coordinates. We then evaluate reduced bases obtained by dropping dimensions (4D→3D→2D→1D). Figure 2 reports mean cross-validated R^2 for (i) all proxy targets and (ii) stance-focused targets only, illustrating that the required dimensionality depends on which observables one aims to explain: format/length targets benefit from explicit z_{format} and z_{length}, whereas stance proxies are largely captured by the $z_{\text{epistemic}}$–z_{scope} subspace. Target-wise stance results for this orthogonal design are provided in Appendix H.1. For completeness, we also reproduce the original 11D reference-basis coarsening check on the cross-lingual politeness probe in Appendix H.2. Complementary indicator-set ablations on the same generations (10→7→5 observables) are reported in Appendix G, asking whether the interference signature persists under coarser observable summaries. Reduced bases (especially 1D–2D) lose predictive sufficiency for the full proxy family under the same protocol, indicating that redistribution signatures can distribute across multiple latent axes rather than collapsing to a single scalar control. This is the sense in which the present probe family justifies retaining the finer-grained reference basis: not as an intrinsic optimum, but as a useful reporting resolution under the current probe/proxy family.

Furthermore, to ensure that our interference detection is not an artifact of a particular axis labeling, we confirm that predictive performance remains stable under random orthogonal rotations within each macro-group subspace (Appendix H.3; Figure H.3).

Note on coverage. While Framing variables (e.g., $z_{\text{audience}}, z_{\text{objective}}$) are structurally important in practice, the present cross-lingual stress test is designed to isolate pathways where typological properties of Japanese make epistemic–scope coupling particularly diagnosable under lightweight observables. Systematic Framing-driven probes are left to future work (Sec. 5).

3.3 Primary and Secondary Effects: A Framework for Latent Interference

Prompts rarely influence a single latent dimension in isolation. This decomposition is conceptual rather than algebraic, and does not assume linearity or numerical comparability across dimensions.

For analytical clarity, the latent shift induced by a minimal prompt fragment π is decomposed as in Eq. (3).

$$\Delta \boldsymbol{z}(\pi) = \Delta \boldsymbol{z}_{\text{primary}}(\pi) + \Delta \boldsymbol{z}_{\text{secondary}}(\pi). \tag{3}$$

where $\Delta \boldsymbol{z}_{\text{primary}}(\pi)$ denotes the shift along the dimension most directly specified by the instruction-level semantics of the fragment, and $\Delta \boldsymbol{z}_{\text{secondary}}(\pi)$ captures systematic shifts along other dimensions that co-activate under the same fragment. Fragment-level effects are the focus here because they are easier to isolate and interpret than full-prompt effects.

In this framework, the *primary* dimension of a prompt fragment is determined by its instruction-level semantics, *i.e.*, the most direct operational constraint explicitly imposed on the model. Secondary dimensions are identified empirically as consistent and reproducible changes in observable output proxies induced by the fragment. In cases such as the introductory example, where a fragment primarily encodes an epistemic stance but also induces agreement-seeking closure or scope suppression, these latter effects are treated as secondary rather than as separate primary targets.

Because the same fragment can act as a primary constraint in one dimension while inducing structured secondary shifts in others, the Z-model goes beyond a static taxonomy of prompt attributes and supports both predictive and diagnostic use.

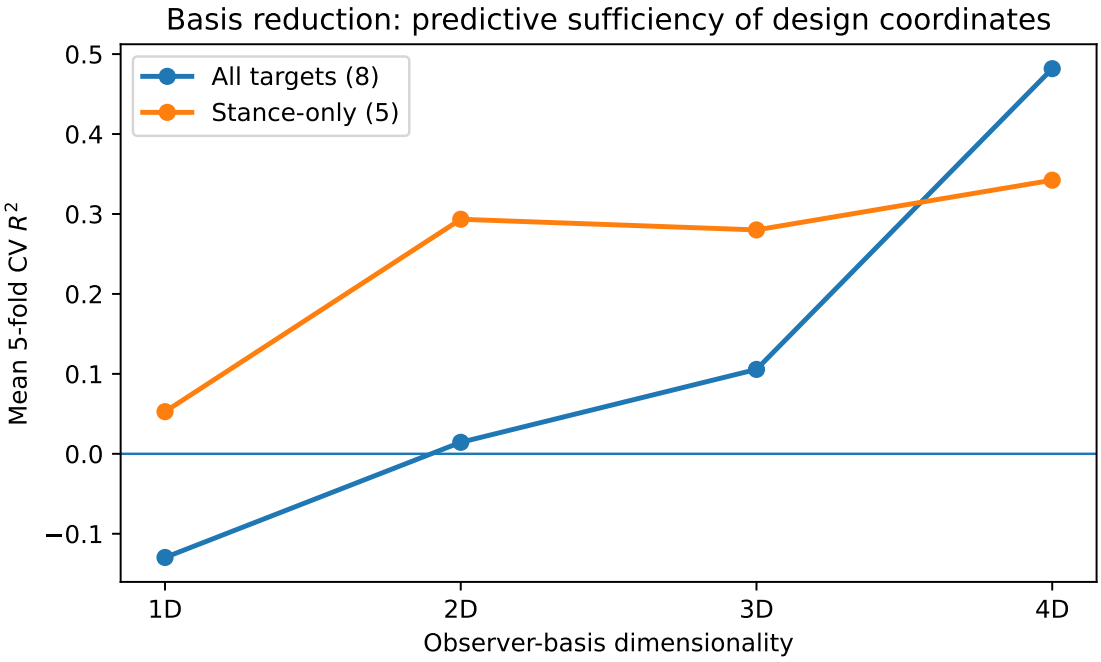

Figure 2: **Dimension justification via basis reduction under an orthogonal 4D design: predictive sufficiency of design coordinates.** Mean 5-fold cross-validated R^2 when predicting each condition's proxy-delta signature from observer-side design coordinates under reduced bases (1D–4D). "All targets" averages over the full proxy set used in the probe (stance + structural/format targets), while "Stance-only" averages over the stance-focused proxies (HEDGE, ASSERT/definite, COND, ALT, PROPENSITY). The separation between the two curves highlights that basis sufficiency is probe-dependent: capturing surface realization (format/length) requires explicit $z_{\text{format}}/z_{\text{length}}$, while stance variation is largely explained by the $z_{\text{epistemic}}$–z_{scope} subspace. This supports *relative* (protocol- and proxy-dependent) descriptive sufficiency rather than a claim of theoretical minimality.

Latent interference is defined as the existence of non-zero secondary effects on latent dimensions other than the intended target. Formally,

$$\exists i \neq j \text{ s.t. } \Delta z_j(\pi) \neq 0, \quad \text{even though the fragment explicitly specifies } z_i. \tag{4}$$

This systematic coupling between primary and secondary prompt effects is referred to as latent interference. As illustrated in Figure 3, a single prompt fragment typically induces a primary shift in its target latent dimension, while simultaneously producing secondary shifts in other dimensions. These secondary effects are not noise, but structured consequences of linguistic and pragmatic coupling.

Throughout this study, the primary dimension of a prompt fragment is defined by its instruction-level semantics (i.e., the most direct operational constraint imposed on the model), while secondary effects are identified empirically via systematic and reproducible changes in output behavior.

Primary effects are often straightforward: "summarize" directly targets z_{task}, while "in bullet points" targets z_{format}. The difficulty is that many Japanese fragments also induce systematic secondary effects. For example, a politeness marker such as「丁寧に」primarily increases z_{style} but can co-vary with $z_{\text{epistemic}}$ in either direction depending on the base framing; cautionary expressions like「慎重に」primarily soften $z_{\text{epistemic}}$ but often narrow z_{scope} and elevate z_{style}; and brevity cues such as「短く」primarily reduce z_{length} while tending to suppress alternatives and increase closure-proneness. The analytical point is not that any one fragment has a single universal signature, but that secondary effects are structured enough to be reported, compared, and probed.

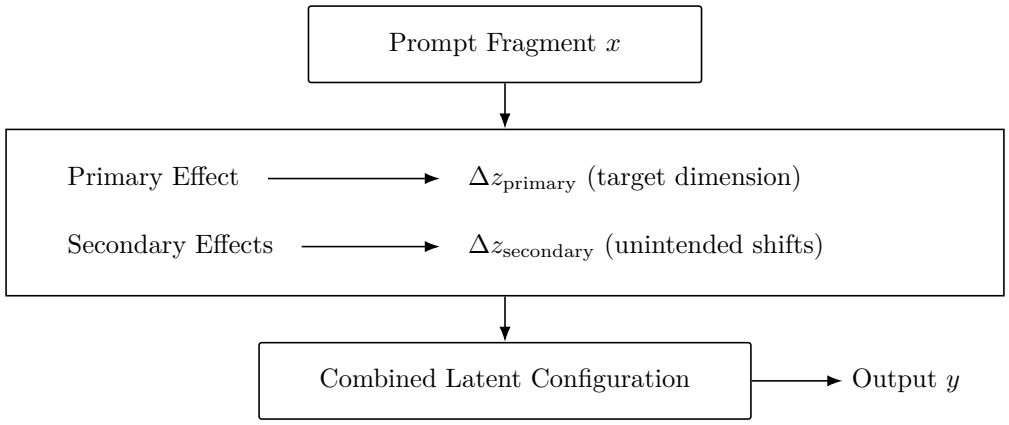

Figure 3: Latent interference caused by prompt fragments. A single prompt fragment typically induces a primary shift in its specified latent dimension ($\Delta z_{\text{primary}}$), while simultaneously producing secondary, unintended shifts in other dimensions ($\Delta z_{\text{secondary}}$).

By formalizing prompt effects in terms of primary and secondary shifts across z-dimensions, the Z-model provides an analytical foundation for diagnosing and predicting prompt-induced variability. In Section 4.2, these ideas are operationalized through forward analysis and inverse diagnosis that connect latent configurations to observable outputs. Representative prompt fragments and their associated primary and secondary effects are summarized in Table 2, which illustrates how linguistic cues systematically propagate across multiple z-dimensions. The Z-model is a diagnostic vocabulary for explaining prompt engineering as interaction-level configuration of a latent behavioral distribution, rather than as surface-level linguistic manipulation or model-internal control.

Table 2: Primary vs. secondary effects induced by Japanese prompt fragments (latent interference). **Legend:** ↑ / ↓ denote typical directional tendencies relative to a neutral baseline; (↑ / ↓) or (±) indicates context-dependent direction; ⇝ denotes qualitative restructuring; *drift* (if used) indicates unstable role/audience inference rather than a monotonic shift.

Prompt fragment (JP)	**Primary target**	**Secondary effects**	**Explanation**
「丁寧に説明して」	z_{style} ↑	$z_{\text{epistemic}}$(↑ / ↓), z_{process} ↑	Politeness cues can couple style with epistemic stance; the sign of $z_{\text{epistemic}}$ may depend on base framing (e.g., confidence inflation vs. hedging regime).
「慎重に説明して」	$z_{\text{epistemic}}$ ↓	z_{style} ↑, z_{scope} ↓	Caution encourages epistemic softening, often narrowing coverage and elevating polite style.
「専門家として」	z_{role} ↑	z_{style} ↑, $z_{\text{epistemic}}$ ↑, z_{scope} ↓	Persona framing implies authority and can narrow the assumed domain and response space.
「短く答えて」	z_{length} ↓	z_{scope} ↓, $z_{\text{epistemic}}$ ↑	Compression suppresses hedging and alternatives, increasing closure-proneness.
「箇条書きで」	z_{format}	z_{process} ⇝, z_{scope} ↓	Formatting restructures reasoning and often reduces breadth via itemization constraints.
「子供にもわかるように」	z_{audience}	z_{style} ⇝, z_{scope} ↓	Audience shift simplifies tone and typically reduces conceptual breadth.
「批判的に検討して」	$z_{\text{objective}}$	z_{style} ⇝, $z_{\text{epistemic}}$ ↑	Critical framing can induce a more assertive evaluative stance.
「A と B を比較して」	z_{process}	z_{scope} ↓, $z_{\text{objective}}$(±)	Comparison framing structures reasoning and can constrain exploration (implicit criteria may vary).
「○○に触れないで」	$z_{\text{constraint}}$	z_{scope} ↓, $z_{\text{epistemic}}$ ↑	Prohibitions reduce breadth and can promote definitive statements.
「要点だけまとめて」	z_{length} ↓	z_{scope} ↓, $z_{\text{epistemic}}$ ↑	Summarization via compression often narrows scope and increases closure.

Operationalization (observable interference). While $\mathbf{z}$ denotes a conceptual latent configuration, all empirical claims in this work are grounded in *observable* output-level measurements. Specifically, we operate on a vector of proxy features $r(y)$ computed from model outputs, whose full definitions and extraction procedures are provided in Appendix C.[1]

We say that a prompt perturbation π exhibits *operational interference* if a fragment intended to primarily target latent dimension i induces a reproducible and directionally consistent change in at least one proxy feature associated with a distinct dimension $j \neq i$. Reproducibility is assessed under matched prompts, fixed decoding parameters, and distribution-level comparisons across repeated runs (Section 3.4).

This definition is deliberately agnostic to the underlying model internals: interference is established solely through observable cross-dimensional effects in $r(y)$, without assuming direct access to latent states.

[1]Each proxy is defined independently of any single experiment and is reused across all evaluations in Section 3.4.

3.4 Minimal cross-lingual probing under matched tasks

To provide empirical grounding for a *targeted portion* of the macro-group vocabulary and the notion of latent interference, we conduct a minimal cross-lingual probing experiment under tightly matched conditions. We use Japanese/English as a *diagnostic stress test setting*: Japanese politeness and stance marking often bundle social alignment and epistemic responsibility in compact forms, which can make selected cross-group pathways more readily observable under lightweight output proxies. At the same time, we interpret any cross-lingual differences as *interaction-protocol* effects and do not attempt to fully disentangle linguistic typology from potential cross-lingual differences in training and alignment data.

While the Z-model instantiates an eleven-dimensional reference basis organized into four macro-groups (Figure 1), exhaustively probing interference across all dimensions is beyond the scope of a single empirical stress test. Instead, our minimal cross-lingual probe deliberately targets one high-leverage pathway: a stylistic perturbation that primarily raises z_{style} (Expression group) and its secondary effects on epistemic stance and response scope ($z_{\text{epistemic}}$ and z_{scope}). This focus is theoretically motivated by Japanese morphosyntax and pragmatics: politeness marking in Japanese compactly packages social alignment with epistemic responsibility, making style–epistemic coupling grammatically salient and thus diagnosable with lightweight surface indicators. The resulting protocol should therefore be read as a targeted stress test of cross-dimensional interference—not as an empirical survey of the entire Z-space or as a validation of every macro-group in the scaffold. Concretely, we do not attempt to operationalize Framing dimensions (e.g., z_{audience} or $z_{\text{objective}}$) in this minimal probe; doing so requires additional probe families and task settings.

The purpose of this experiment is not to compare task performance or factual correctness.

Cross-lingual comparability. We intentionally avoid interpreting absolute proxy rates across languages. Our claims are limited to *within-language, within-protocol* deltas under matched prompts and topics. Because Japanese and English realize stance and politeness through different lexical and morphosyntactic devices, the proxy set is designed for directional sensitivity rather than exhaustive coverage (Appendix C). Instead, it examines whether a purely stylistic cue induces different *secondary effects* across languages, consistent with the latent interference framework.

All experiments were conducted under a black-box, API-served large language model, with fixed system prompts and decoding parameters, during a specified execution window. The API-served model is treated as an instance of a contemporary instruction-tuned LLM, used here to demonstrate interaction-level prompt sensitivity under realistic deployment conditions.

Because API-served models may change across time and snapshots, we do not rely on bitwise reproducibility or exact string matching. Instead, all empirical claims are stated in terms of within-protocol, distribution-level directional deltas over repeated generations. The full API settings, execution window, and decoding parameters are documented in Appendix E.

Five benign technical topics are used that rely on general principles rather than domain-specific factual knowledge, and that naturally admit conditions, limitations, and alternative explanations. This design minimizes confounds from factual disputes while allowing epistemic and scope-related variations to surface. As a concrete example, one topic is *noise, signal-to-noise ratio (SNR), and the effect of averaging.*

For each (topic, language, style) cell, we sample $n = 50$ independent generations; aggregating over five topics yields $N = 250$ outputs per language and style.

For each language, four conditions are evaluated using the same task description, model, decoding parameters, and output constraints: (i) Japanese with an explicit politeness cue (“丁寧に”), (ii) Japanese without the politeness cue, (iii) English with an explicit politeness cue (“politely”), and (iv) English without the politeness cue. All other aspects of the prompt are held constant.

Rather than scoring correctness, surface-level output features are analyzed that operationalize epistemic stance and response scope, following the z-dimensions introduced earlier. Specifically, we measure lightweight lexical/structural proxies that operationalize observable shifts along $z_{\text{epistemic}}$ and z_{scope}: (1) the rate of hedging expressions (HEDGE; uncertainty markers), (2) the rate of definitive closure markers (ASSERT;

reported as `definite` in our implementation), (3) the rate of alternative/perspective markers (ALT), interpreted broadly as a proxy for multi-angle exposition and scope expansion (including additive markers such as *also/additionally/furthermore* and また/さらに/加えて), and (4) the rate of conditional statements (COND; conditions/exceptions). We additionally report PROPENSITY as an auxiliary indicator of response openness and simple structural indicators (CHARS, SENTENCES, LIST_ITEMS) for breadth and formatting.

To mitigate model- and snapshot-dependence inherent to API-served systems, we additionally reproduce the core probing experiments on a pinned open-weight, instruction-tuned model with an explicit revision hash. Unless otherwise stated, directional conclusions reported in this study are supported by results on both the API-served model and the pinned open-weight model (see Appendix F).

3.5 Results

Across five matched technical topics ($N = 250$ outputs per language and style), politeness cues induce systematic but language-dependent shifts in uncertainty expression and response scope. Specifically, politeness selectively increases epistemic hedging in English, whereas in Japanese it primarily suppresses generality/typicality hedging, motivating a component-wise decomposition of hedging behavior.

Table 4 summarizes the cross-lingual effects of politeness under the matched probing protocol described above. For readability, we report the ASSERT proxy as `definite` and the ALT proxy as `alt`. Rather than reporting accuracy, the table focuses on surface-level proxies that operationalize shifts along the $z_{\text{epistemic}}$ and z_{scope} dimensions. Unless otherwise noted, confidence intervals are computed using an output-level nonparametric bootstrap, treating individual generations as independent samples within each condition. We additionally report standardized effect sizes (Cohen's d) and histogram-based Jensen–Shannon divergence (JSD) with bootstrap confidence intervals (Table 5).

Open-weight reproduction (main-text snapshot). To reduce model/snapshot dependence, we reproduce the cross-lingual politeness probe (Exp. A) on a pinned open-weight instruction-tuned checkpoint with matched prompts, decoding settings, and proxy definitions (Qwen/Qwen2.5-7B-Instruct; Appendix F). Absolute magnitudes differ, but several qualitative directional patterns persist, supporting that the observed secondary effects are not purely an artifact of one API-served snapshot. Table 3 provides a compact side-by-side snapshot for three representative proxies (JP and EN shown on separate lines).

Robustness to indicator choice. A common concern for lightweight proxy-based probing is brittleness: do the conclusions hinge on a particular cue set? To address this without additional model calls, we perform an indicator-set ablation on the same generations used in Tables 4–5 (Appendix G). Under nested reductions from the full set to substantially smaller sets, the Japanese signature remains detectable: epistemic shifts (HEDGE/ASSERT) and scope shifts (COND/ALT) persist directionally even when only a minimal indicator set is retained. This supports the interpretation that the observed interference is not an artifact of any single lexical proxy. We report `PROPENSITY` only as an auxiliary exploratory marker; the core directional conclusions remain under a minimal indicator set that excludes `PROPENSITY` (Appendix G).

Table 3: Selected polite–plain deltas (95% bootstrap CIs) for the API-served model (Table 4) and a pinned open-weight, instruction-tuned checkpoint (Appendix F, Table F.1). Rates are reported per 1,000 characters; LIST_ITEMS is reported as a raw count difference. JP and EN results are shown on separate lines.

Proxy (delta)	**API-served** (JP / EN)	**Open-weight** (JP / EN)
	−0.004[−0.071, 0.065]	0.184[−0.116, 0.486]
HEDGE_EPI (/1,000 chars)	0.237[0.156, 0.320]	0.273[0.177, 0.375]
	−0.448[−0.808, −0.095]	−0.202[−0.410, 0.006]
HEDGE_GEN (/1,000 chars)	−0.003[−0.171, 0.162]	−0.103[−0.187, −0.018]
	−1.372[−1.768, −0.972]	−1.486[−2.624, −0.368]
LIST_ITEMS (count)	−0.732[−0.908, −0.568]	−2.590[−2.989, −2.196]

Table 4: Bootstrap 95% confidence intervals for politeness effects (polite−plain). Bootstrap mode: iid.

Metric	Δ_{JP} (polite−plain)	Δ_{EN} (polite−plain)
hedge_per_1k_chars	-0.452 [-0.808, -0.088]	0.233 [0.050, 0.421]
hedge_epi_per_1k_chars	-0.004 [-0.071, 0.065]	0.237 [0.156, 0.320]
hedge_gen_per_1k_chars	-0.448 [-0.808, -0.095]	-0.003 [-0.171, 0.162]
definite_per_1k_chars	0.066 [0.003, 0.131]	-0.003 [-0.014, 0.008]
cond_per_1k_chars	0.826 [0.333, 1.323]	-0.011 [-0.225, 0.206]
alt_per_1k_chars	0.274 [0.008, 0.546]	0.014 [-0.078, 0.103]
propensity_per_1k_chars	-0.163 [-0.601, 0.293]	0.148 [0.043, 0.257]
chars	43.496 [35.076, 51.760]	-24.652 [-37.873, -11.463]
list_items	-1.372 [-1.768, -0.972]	-0.732 [-0.908, -0.568]
sentences	2.460 [2.004, 2.904]	0.764 [0.480, 1.052]

Table 5: Effect sizes (Cohen's d) and histogram-based Jensen–Shannon divergence (JSD) with bootstrap 95% confidence intervals for politeness effects (polite−plain). Bootstrap mode: iid.

Metric	d_{JP}	$\mathrm{JSD}_{\mathrm{JP}}$	d_{EN}	$\mathrm{JSD}_{\mathrm{EN}}$
hedge_per_1k_chars	-0.216 [-0.390, -0.042]	0.127 [0.113, 0.220]	0.218 [0.047, 0.397]	0.065 [0.060, 0.140]
hedge_epi_per_1k_chars	-0.010 [-0.186, 0.165]	0.011 [0.007, 0.033]	0.503 [0.330, 0.696]	0.070 [0.049, 0.121]
hedge_gen_per_1k_chars	-0.216 [-0.390, -0.046]	0.138 [0.123, 0.230]	-0.003 [-0.178, 0.169]	0.060 [0.058, 0.135]
definite_per_1k_chars	0.178 [0.008, 0.350]	0.029 [0.017, 0.058]	-0.044 [-0.180, 0.127]	0.006 [0.000, 0.014]
cond_per_1k_chars	0.289 [0.117, 0.470]	0.105 [0.096, 0.188]	-0.009 [-0.183, 0.170]	0.096 [0.083, 0.168]
alt_per_1k_chars	0.172 [0.005, 0.345]	0.121 [0.100, 0.188]	0.027 [-0.157, 0.206]	0.020 [0.014, 0.054]
propensity_per_1k_chars	-0.063 [-0.231, 0.113]	0.097 [0.087, 0.179]	0.239 [0.071, 0.406]	0.073 [0.054, 0.119]
chars	0.919 [0.735, 1.114]	0.183 [0.167, 0.276]	-0.330 [-0.514, -0.154]	0.068 [0.067, 0.135]
list_items	-0.593 [-0.781, -0.418]	0.114 [0.083, 0.171]	-0.745 [-0.867, -0.630]	0.140 [0.102, 0.191]
sentences	0.978 [0.794, 1.167]	0.175 [0.142, 0.250]	0.467 [0.295, 0.647]	0.076 [0.050, 0.130]

Indicator-set ablation. To verify that the interference signal does not hinge on any single proxy choice, we also re-analyze the same generations under reduced indicator sets (10→7→5 indicators; Appendix G).

We interpret the deltas below as within-protocol redistribution signatures in epistemic and scope markers, rather than as task-accuracy differences.

Politeness-induced shifts in uncertainty expression. Across matched topics, politeness cues induce systematic but language-dependent shifts in uncertainty expression. To disentangle these effects, hedging is decomposed into *epistemic hedges* (expressing uncertainty about truth or knowledge) and *generality/typicality hedges* (expressing looseness, typicality, or non-commitment), and both components are analyzed separately.

In English, politeness selectively increases epistemic hedging (a positive Δ with a bootstrap CI excluding zero; Table 4), while leaving generality/typicality markers statistically unchanged (CI crossing zero). As a result, overall hedging (epistemic + generality) increases under polite prompts. This pattern indicates a first-order epistemic effect of politeness in English: polite framing encourages explicit uncertainty marking without broad stylistic diffusion.

In Japanese, by contrast, politeness does not reliably increase epistemic hedging ($\Delta \approx 0$), but significantly reduces generality/typicality hedging (a negative Δ with a CI excluding zero; Table 4). Consequently, overall hedging decreases under polite prompts. This pattern reflects a second-order redistribution effect, in which polite framing suppresses non-committal or vague generality expressions rather than amplifying epistemic uncertainty.

Taken together, these results show that politeness does not induce a uniform stylistic shift across languages. Instead, it operates through distinct mechanisms: a first-order increase in epistemic uncertainty marking in English, and a second-order suppression of generality-based hedging in Japanese. This asymmetry supports the claim that Japanese functions as a typological stress-test language, in which stylistic cues more readily

propagate into latent epistemic and scope-related dimensions. Table 5 provides complementary standardized effect-size summaries (Cohen's d) and a distributional divergence measure (JSD).

Prediction-to-observation check (factorial 2×2; structured interference). We now test whether such structured forward shifts are empirically non-additive under controlled perturbations. Paired A/B deltas establish *co-occurrence* of primary and secondary effects, but they do not by themselves rule out a purely additive explanation. Unlike the A/B "2×2" presentation grid in Table 9, here "2×2" denotes a true factorial design that crosses two prompt fragments. As an explicit *prediction-to-observation* check, we run a small 2×2 factorial probe as a controlled diagnostic probe on a representative topic (*noise, signal-to-noise ratio (SNR), and the effect of averaging*) at $T = 0.2$, crossing (A) a politeness cue ("politely" / 「丁寧に」) and (B) an epistemic anti-closure instruction ("avoid definitive claims; if there are conditions or exceptions, make them explicit"). We sample $n = 50$ independent generations per cell under fixed decoding settings. Let $\mu_{ab}^{(k)}$ denote the mean of proxy k in cell C_{ab}, where $a, b \in \{0, 1\}$ indicate the presence of fragments A and B. Under additivity, the predicted cell mean for the combined condition is

$$\mu_{11,\text{pred}}^{(k)} = \mu_{10}^{(k)} + \mu_{01}^{(k)} - \mu_{00}^{(k)}, \tag{5}$$

corresponding to a predicted combined shift $\Delta_k^{\text{pred}} = \mu_{11,\text{pred}}^{(k)} - \mu_{00}^{(k)}$. The actual combined shift is $\Delta_k^{\text{obs}} = \mu_{11}^{(k)} - \mu_{00}^{(k)}$, and their difference

$$I_k = \Delta_k^{\text{obs}} - \Delta_k^{\text{pred}} = \mu_{11}^{(k)} - \mu_{10}^{(k)} - \mu_{01}^{(k)} + \mu_{00}^{(k)} \tag{6}$$

is the interaction residual. Its standardized version is reported as d_I. For each proxy metric, we report the standardized interaction effect size d_I (difference-in-differences normalized by pooled standard deviation); $d_I \approx 0$ corresponds to additivity, while $d_I \neq 0$ indicates that the effect of one fragment depends on the presence of the other. Table 6 reports main-effect and interaction-effect magnitudes with bootstrap confidence intervals, and Figure 4 visualizes where interaction effects localize across proxies in each language. This non-additivity supports the view that prompt fragments do not operate as independent control knobs on separable axes, but instead compose through structured latent interference consistent with the observer-side Z-model, where activating one behavioral dimension can reweight or suppress others in a context-dependent manner. We emphasize that this factorial probe is intended as a focused diagnostic on one topic; extending it across topics and task families is left to future work. The observed non-additivity is not an isolated artifact of a single proxy set or basis parameterization: related redistribution signatures remain stable under indicator-set ablation (Appendix G) and under observer-basis coarsenings and rotations (Appendix H.2–H.3), supporting structured interference rather than accidental variance.

Cell means for representative metrics, together with the additive prediction from Eq. (5), are provided in Appendix H.4.

Table 6: **Effect magnitudes (standardized) for main and interaction effects (factorial probe, $T = 0.2$).** We report Cohen's d for the main effects of the politeness cue (A) and the anti-closure instruction (B), and the standardized interaction d_I (difference-in-differences normalized by pooled SD), each with 95% bootstrap confidence intervals.

English

Proxy	d_A (95% CI)	d_B (95% CI)	d_I (95% CI)
definite_per_1k_chars	-0.32 [-0.52,-0.11]	-0.59 [-0.74,-0.42]	+0.92 [+0.50,+1.26]
alt_per_1k_chars	-0.31 [-0.55,-0.06]	-0.02 [-0.25,+0.24]	-0.30 [-0.81,+0.21]
cond_per_1k_chars	+0.11 [-0.14,+0.35]	+0.30 [+0.06,+0.52]	+0.15 [-0.39,+0.62]
chars	-0.51 [-0.74,-0.28]	+0.35 [+0.10,+0.59]	+0.09 [-0.36,+0.59]
hedge_per_1k_chars	+0.20 [-0.05,+0.44]	+0.34 [+0.11,+0.58]	-0.05 [-0.54,+0.44]
sentences	-0.06 [-0.31,+0.21]	-0.21 [-0.46,+0.05]	+0.04 [-0.47,+0.54]
propensity_per_1k_chars†	+0.00 [+0.00,+0.00]	+0.00 [+0.00,+0.00]	+0.00 [+0.00,+0.00]
list_items	+0.03 [-0.22,+0.28]	+0.00 [-0.27,+0.25]	+0.00 [-0.50,+0.50]

Note: † In the English factorial probe, PROPENSITY had zero lexicon matches across all cells, so standardized effects collapse to 0. Core conclusions remain under a minimal indicator set that excludes PROPENSITY (Appendix H.4).

Japanese

Proxy	d_A (95% CI)	d_B (95% CI)	d_I (95% CI)
chars	+0.37 [+0.16,+0.56]	+1.02 [+0.83,+1.19]	-0.47 [-0.89,-0.06]
hedge_per_1k_chars	+0.49 [+0.26,+0.72]	+0.52 [+0.30,+0.75]	-0.36 [-0.85,+0.09]
list_items	-0.13 [-0.22,+0.00]	-0.13 [-0.22,+0.00]	+0.26 [+0.00,+0.45]
cond_per_1k_chars	+0.02 [-0.22,+0.26]	+0.28 [+0.02,+0.52]	+0.18 [-0.32,+0.67]
alt_per_1k_chars	+0.04 [-0.20,+0.30]	+0.03 [-0.22,+0.28]	-0.17 [-0.68,+0.33]
definite_per_1k_chars	-0.21 [-0.43,+0.05]	+0.00 [-0.26,+0.25]	+0.08 [-0.42,+0.61]
propensity_per_1k_chars	-0.16 [-0.41,+0.09]	+0.23 [-0.02,+0.46]	+0.06 [-0.46,+0.56]
sentences	+0.36 [+0.12,+0.59]	+0.17 [-0.07,+0.40]	+0.06 [-0.44,+0.54]

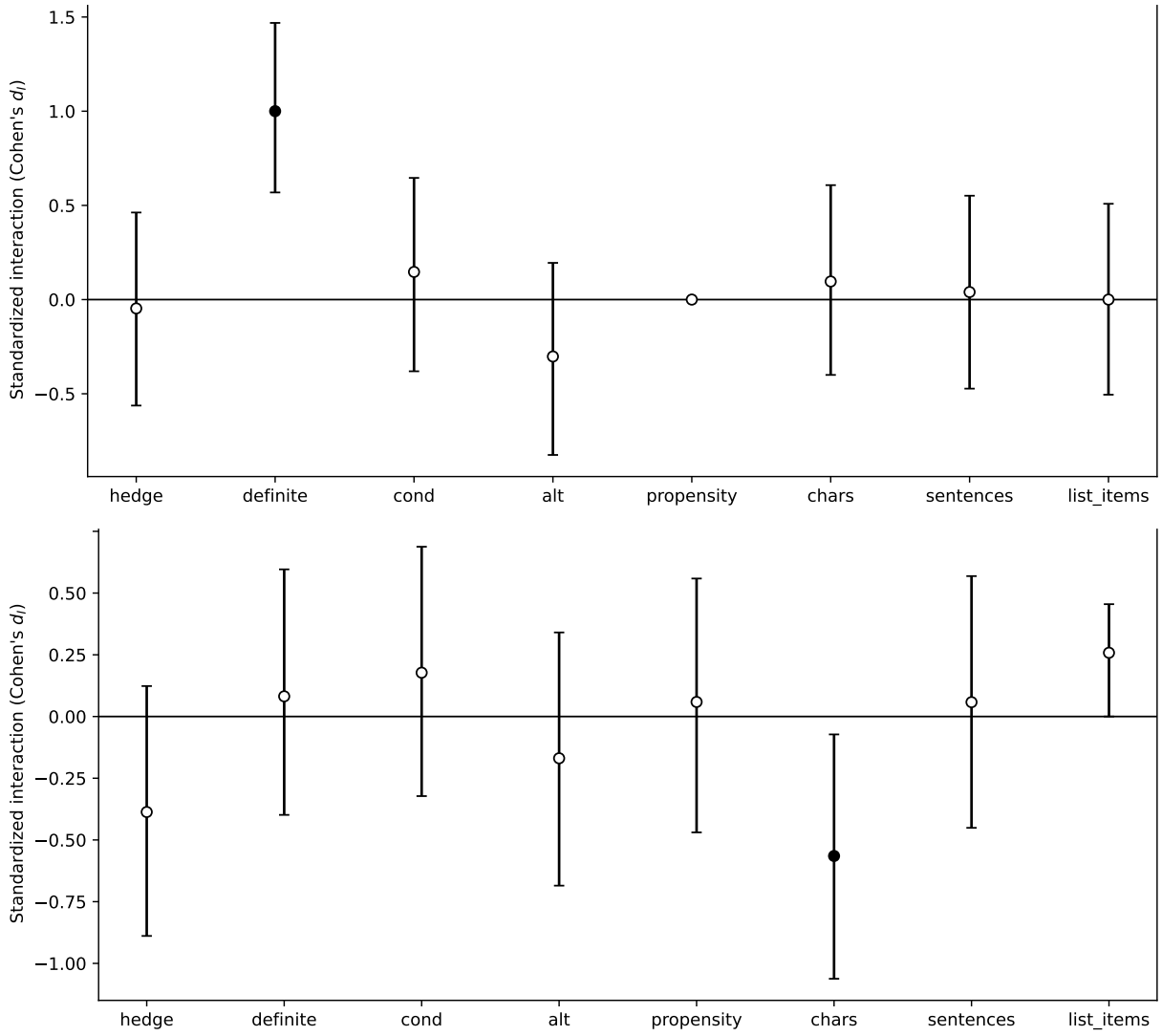

Figure 4: **Factorial 2×2 probe: standardized interaction effects (monochrome-friendly).** Standardized interaction effect size d_I with 95% bootstrap confidence intervals for each proxy metric, shown separately for English (top) and Japanese (bottom). The interaction is computed as a difference-in-differences over the four conditions $C00 = p_0$, $C10 = p_0 \oplus A$, $C01 = p_0 \oplus B$, $C11 = p_0 \oplus A \oplus B$, where A is a politeness cue and B is an anti-closure instruction. Non-zero interaction indicates non-additivity (structured interference) beyond independent main effects.

3.6 Reproducibility note

To isolate prompt-induced effects, we fixed the base prompt, system prompt, and maximum output length, and varied only the probe fragment, while holding decoding parameters constant. The quantities needed for replication are therefore the prompt templates, model identifiers (and, for API-served models, the execution window), sampling settings, proxy definitions, and post-processing rules. These are specified explicitly in the appendix rather than left to an external code release.

For the main cross-lingual probe (Exp. A), we sample $n = 50$ generations per topic and condition; with five topics this corresponds to $N = 250$ outputs per language and style. The intended replication target is recovery of the reported *directional, distribution-level deltas* under the matched protocol, not bitwise reproduction of individual strings.

Readers who prefer a snapshot-stable reproduction path can rerun the same prompts and counting rules on the pinned open-weight checkpoint reported in Appendix F.

To connect the cross-lingual probe to linguistically grounded sources of coupling, Table 7 summarizes four Japanese-specific mechanisms that can amplify latent interference under black-box prompting; we refer back to these mechanisms in the discussion.

4 Analysis and Discussion

Politeness is used in this study not as a privileged phenomenon, but as a diagnostic probe: it provides a compact and linguistically grounded perturbation that reliably induces secondary effects across multiple **Z** dimensions.

4.1 Japanese as a stress-test language

Our empirical probe and case analyses suggest that Japanese is a useful stress-test language for latent interference in prompting: pragmatic cues that are compactly encoded in Japanese (politeness, sentence-final stance marking, omission) make multi-dimensional coupling more readily observable than in English. In our matched probing protocol, a single politeness cue yields qualitatively different redistribution of uncertainty markers across JP and EN, consistent with the Z-model view that secondary effects depend on how a language packages pragmatics.

We do not claim that latent interference is unique to Japanese. Rather, Japanese provides a setting in which coupling is grammatically salient and thus easier to diagnose with lightweight surface indicators; similar couplings are expected in other languages and domains, but may require different indicators and prompt perturbations.

Linguistic mechanisms amplifying latent interference.

(1) Politeness–Epistemic Coupling. Japanese possesses an extensive system of honorifics and polite registers that encode social stance, interpersonal distance, and interactional alignment. However, many polite forms also function as epistemic markers, implicitly modulating confidence or authoritativeness. Consequently, expressions such as「丁寧に」or です／ます調 primarily raise z_{style} but routinely trigger secondary modulation of $z_{\text{epistemic}}$, often manifested as increased hedging, and in some regimes increased scope-licensing markers (e.g., conditionality or alternatives). This coupling makes stylistic modifications indistinguishable from epistemic adjustments at the surface level, creating a structural source of interference not typically observed in languages where politeness is syntactically or lexically separated from evidential stance.

(2) Subject Omission and Role Ambiguity. Japanese allows pervasive omission of grammatical subjects, agents, and even discourse participants when recoverable from context. While efficient in natural dialogue, this property destabilizes z_{role} and z_{audience} during LLM inference, because the model must infer who is speaking and to whom. Slight prompt variations—changes in sentence-final forms, topicalization, or particle choice—can shift inferred roles or audiences, leading to inconsistent personas

Table 7: Japanese-specific mechanisms that amplify latent interference in prompt–LLM interactions. We organize prior observations into **four core mechanisms** consistent with Figure 5; additional amplifiers (e.g., honorific morphology, flexible word order) are included as representative sub-phenomena under the relevant mechanism. Arrows denote typical tendencies relative to a neutral baseline; *drift*/± indicate context-dependent direction; ⇝ denotes qualitative restructuring.

Core mechanism	**Representative sub-phenomena (JP)**	**Why it amplifies latent interference**	**Typical z-effects (primary → secondary)**
Politeness–epistemic coupling	丁寧語・敬語，です／ます調, 尊敬語・謙譲語 (honorific morphology)	Politeness and honorific marking encode interactional alignment and social stance compactly. In Japanese, these cues can be co-interpreted as epistemic authority or responsibility signals, so a stylistic instruction can co-activate epistemic stance (the sign may depend on base framing).	z_{style} ↑→ $z_{\text{epistemic}}$(↑ / ↓), z_{process} ↑
Argument omission and role ambiguity	主語／目的語の省略 (ゼロ代名詞)，話者／聞き手の省略	Omission increases the inference burden for recovering roles and addressees from context. This can induce instability in inferred persona and audience, and propagate into reasoning alignment and stance.	z_{role}(drift), z_{audience}(drift) → $z_{\text{epistemic}}$(±), z_{process}(±)
Vague modifiers and pragmatic overloading	慎重に，丁寧に，適切に，わかりやすく，なるべく…	Semantically underspecified but pragmatically rich modifiers bundle multiple implied constraints (e.g., tone, caution, coverage, and reasoning strategy), making secondary shifts structurally likely.	(*varies*) → z_{style} ↑ , $z_{\text{epistemic}}$ ↓, z_{scope} ↓
Topic prominence and discourse-driven structure	は (topic marking), 主題化, 情報構造主導, 語順の自由度 (flexible word order)	Discourse- and topic-oriented packaging can override purely grammatical cues, shaping what is foregrounded, which reasoning path is taken, and what is implicitly treated as the objective or frame.	z_{process} ⇝, z_{scope} ↓ , $z_{\text{objective}}$(±)

or shifts in assumed expertise. These role-level drifts interact with $z_{\text{epistemic}}$ and z_{process}, propagating ambiguity into reasoning mode and stance.

(3) Vague Modifiers and Pragmatic Overloading. Expressions such as「慎重に」,「適切に」,「丁寧に」, or「わかりやすく」are pragmatically rich but semantically underspecified. They do not strictly target a single behavioral property but simultaneously suggest adjustments to tone, reasoning strategy, epistemic caution, and output structure. For instance,「慎重に」may be interpreted as promoting epistemic caution (modulating $z_{\text{epistemic}}$), narrowing conceptual exploration (affecting z_{scope}), and encouraging polite phrasing (affecting z_{style}). This multi-functionality directly increases the likelihood of secondary effects, showing how seemingly innocuous modifiers can shift the model along multiple z-dimensions.

(4) Topic-Prominence and Discourse-Driven Structure.
Japanese is a topic-prominent language, where discourse structure often precedes or overrides grammatical structure. Topic marking (e.g., は), floating quantifiers, and flexible word order influence

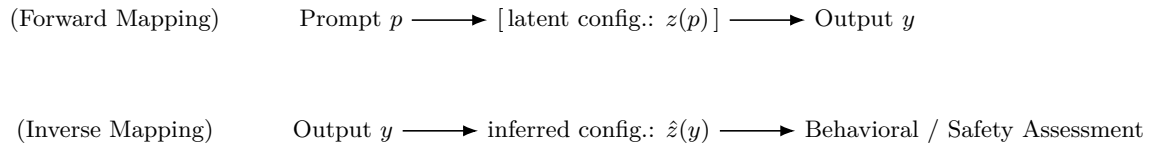

Figure 5: Forward mapping and inverse diagnostic workflow between prompts, latent configurations, and outputs. In the forward direction, a prompt p induces a latent configuration $z(p)$, which in turn produces an observable output y. In the inverse direction, properties of the output are used to construct a qualitative diagnostic approximation $\hat{z}(y)$ for behavioral and safety assessment.

how information is packaged and which elements are foregrounded. These features interact strongly with z_{process} and z_{scope}: prompts that foreground a particular topic may implicitly constrain the model's reasoning trajectory or reduce the diversity of explored alternatives. Topic-first constructions can also introduce biases in the inferred objective ($z_{\text{objective}}$) by suggesting evaluative framing even when none is explicitly requested.

These linguistic properties motivate the analytical tools developed in Section 4.2, which formalize how latent configurations can be inferred or manipulated through prompts.

Together, these mechanisms produce a linguistic environment in which latent interference is not accidental but structurally motivated. Japanese prompt fragments frequently encode multiple pragmatic intentions or require substantive inferential reconstruction, causing primary and secondary z-dimension effects to become entangled. This explains why Japanese prompting often yields greater behavioral variance, sharper stylistic–epistemic interactions, and more pronounced unpredictability than English prompting. The forward analysis and inverse diagnostic workflow in Section 4.2 build directly on these observations, providing methodological tools for tracing how these linguistic properties manifest in model behavior.

Why is interference more limited in English under the same probe? One plausible explanation is typological: English politeness strategies are largely periphrastic and lexically optional, whereas Japanese politeness is grammatically integrated into clause-level morphology and stance marking. This asymmetry reduces enforced cross-dimensional entanglement in English, yielding predominantly first-order shifts (e.g., increased explicit epistemic hedging) rather than redistribution effects across multiple observable proxies. In our framework, Japanese serves as a stress test precisely because compact morphosyntax increases the likelihood that a single stylistic cue (primary) induces secondary shifts in epistemic stance and scope. Viewed through this lens, the Z-model provides a unified, language-aware interpretation: the same primary stylistic perturbation can yield markedly different secondary signatures depending on how politeness and stance are packaged in the language.

The main linguistic pathways through which Japanese induces multi-dimensional interference are consolidated in Table 7, highlighting how structural and pragmatic properties translate into predictable z-dimension perturbations.

4.2 Forward analysis and an inverse diagnostic workflow for prompt–latent dynamics

To move from descriptive linguistic pathways to testable diagnostics, we now introduce a minimal observer-side mapping between prompt fragments, latent configurations, and observable outputs. Building on the Z-model (Section 3), forward analysis is predictive and design-oriented, whereas the inverse direction is post-hoc and diagnostic; the two are complementary rather than a circular estimator. As summarized in Figure 5, prompts induce latent configurations that generate outputs, and output properties can in turn organize a qualitative diagnostic approximation of the relevant dimensions for behavioral and safety assessment.

4.3 Forward Analysis: Prompt $\rightarrow z \rightarrow$ Output Behavior

We conceptualize prompt–model interaction through a forward mapping from prompts to latent configurations and outputs:

$$\boldsymbol{z}(p) = f_{\text{enc}}(p), \quad y(p) = g_{\text{dec}}(\boldsymbol{z}(p)). \tag{7}$$

Here f_{enc} and g_{dec} are abstract black-box mappings that need not be specified explicitly. They are observer-side abstractions rather than identifiable internal modules. Forward analysis asks how specific prompt fragments shift the latent configuration and how those shifts predict observable changes in the output. This deterministic shorthand should be read as a compact proxy for the stochastic observer-side latent block in Eq. (1); claims are made at the level of output distributions or proxy expectations, not single-output determinism.

Example Forward Effects (Qualitative Demonstrations). The examples below illustrate how small perturbations in Japanese prompts can induce structured latent shifts.

(a) **「丁寧に説明してください」**
Intended shift: $z_{\text{style}} \uparrow$
Possible secondary shifts: co-variation in $z_{\text{epistemic}}$ (confidence inflation in one regime) and $z_{\text{process}} \uparrow$ (longer reasoning chains)
Resulting output pattern: A polite surface form paired with longer reasoning and, in some regimes, reduced hedging or stronger closure.

(b) **「慎重に説明してください」**
Intended shift: $z_{\text{epistemic}}$ (caution)
Possible secondary shifts: $z_{\text{style}} \uparrow$ (politeness elevation), $z_{\text{scope}} \downarrow$ (narrowed coverage)
Resulting output pattern: Polite but narrow or overly hedged explanations, sometimes omitting alternative perspectives.

(c) **「要点だけまとめて」**
Intended shifts: $z_{\text{length}} \downarrow$, $z_{\text{scope}} \downarrow$
Possible secondary shift: $z_{\text{epistemic}} \uparrow$ (compression-induced assertiveness)
Resulting output pattern: Short but overly conclusive statements, corresponding to compressive hallucinations.

These examples are meant as structured, language-conditioned heuristics rather than exhaustive claims; Section 4.4 makes the corresponding observer-side procedure explicit.

4.4 Methodology for Forward Analysis

To support reproducible forward analysis under black-box access, we recommend the following lightweight procedure.

(1) **Prompt-fragment isolation.** Identify minimal Japanese expressions (e.g.,「丁寧に」,「慎重に」,「専門家として」) that are hypothesized to affect specific latent dimensions, while keeping the remaining prompt context fixed.
(2) **Latent projection.** For each fragment π, specify its intended *primary* target dimension and hypothesize plausible *secondary* shifts (latent interference) using Section 3.2. This yields a qualitative projection of $\Delta\boldsymbol{z}(\pi)$ into primary and off-target components.
(3) **Behavioral expectation mapping.** Translate the projected shifts into observable indicators. Consistent with Appendix C, we use HEDGE/ASSERT for $z_{\text{epistemic}}$ and COND/ALT for z_{scope}. Roughly, $z_{\text{epistemic}} \uparrow$ corresponds to stronger closure and fewer hedges, whereas $z_{\text{epistemic}} \downarrow$ corresponds to more explicit uncertainty. Likewise, $z_{\text{scope}} \uparrow$ increases conditions and alternatives, while $z_{\text{scope}} \downarrow$ suppresses them. Higher z_{process} tends to yield more structured multi-step reasoning, and lower z_{length} compresses outputs, often co-varying with closure.
(4) **Optional qualitative comparison.** Generate paired outputs under controlled conditions to illustrate the predicted differences. This step is for transparent demonstration and prompt debugging, not for statistical validation.

Algorithm 1 Forward analysis procedure (observer-side)

Require: Base prompt p_0, probe fragment π, optional control fragment π_0, target dimension z_t, proxy map $\mathcal{M}$.

Ensure: Hypothesized latent shift signature $\Delta \boldsymbol{z}^{\text{hyp}}$ and expected proxy directions Δm^{hyp}.

1: Construct $p_A \leftarrow p_0 \oplus \pi$ and (optionally) $p_B \leftarrow p_0 \oplus \pi_0$.
2: Specify the intended primary shift on z_t (e.g., $z_{\text{style}} \uparrow$ for politeness).
3: Using Figure 1 and Tables 2,7, hypothesize secondary shifts Δz_j^{hyp} for $j \neq t$ (latent interference).
4: Map each hypothesized shift to observable proxy directions via $\mathcal{M}$ (e.g., $z_{\text{epistemic}} \uparrow \Rightarrow$ HEDGE$\downarrow$, ASSERT$\uparrow$; $z_{\text{scope}} \uparrow \Rightarrow$ COND/ALT$\uparrow$; Appendix C).
5: (Optional) Generate paired samples under p_A/p_B and compare observed proxy deltas with Δm^{hyp} for prompt debugging.
6: If the signature mismatches, revise π or update the hypothesized interference path and repeat.

This procedure is conceptual: it does not assume identifiability of $\boldsymbol{z}$ or explicit internal modules, but it remains compatible with black-box LLM access. Algorithm 1 summarizes the forward procedure in pseudo-code form.

Selection principle (linguistically motivated fragments). In Japanese, discourse stance and interactional alignment are frequently encoded by short adverbials and compact morpho-syntactic choices, including politeness and honorific markers, modality-related forms, and systematic underspecification of topics or participant roles. Extensive work in Japanese linguistics has shown that such minimal surface cues play a central role in signaling social stance and epistemic positioning (e.g., politeness and honorific systems: Ide 1989; Matsumoto 1988; topic prominence and discourse organization: Mikami 1960; Kuno 1973; omission and contextual role inference: Kuno 1973; Shibatani 1990; modality and epistemic stance: Nitta 1991; Maynard 1993).

Guided by these descriptions, we select candidate prompt fragments that are minimal in form yet plausibly capable of shifting stance, role interpretation, or discourse organization while leaving propositional content unchanged. The cited linguistic literature grounds the intended pragmatic and discourse functions of these fragments. In contrast, the claims of the Z-model concern how such linguistically motivated functions may interact with black-box LLM behavior under controlled prompting, rather than proposing new linguistic analyses themselves.

4.5 Inverse Diagnostic Workflow: From Output Indicators to Latent Dimensions

The inverse direction treats generated outputs as observer-side evidence about the latent configuration plausibly adopted in response to a prompt. Under black-box access, this evidence comes from linguistic, structural, and epistemic markers in the output, which can be organized into a structured post-hoc diagnosis of prompt-induced behavior.

Formally, we write the inverse diagnostic step as

$$\widehat{\boldsymbol{z}} = h(y), \tag{8}$$

where h is a qualitative, model-agnostic mapping from surface indicators to a diagnostic approximation of the relevant dimensions. It is not used to validate the Z-model and is not claimed to recover a unique or identifiable latent state. Rather, $\hat{z}$ is an interpretable diagnostic approximation for comparison across conditions, prompt debugging, and behavioral or safety-oriented analysis.

Algorithm 2 summarizes the inverse diagnostic (A/B probing) procedure in pseudo-code form.

Interference decision rule (nontriviality filter). To avoid over-interpreting vanishingly small but statistically detectable differences, we use a conservative two-part rule when *declaring* latent interference sets. An off-target dimension $j \neq t$ is flagged as interfering only if at least one associated proxy $m_k \in \mathcal{M}(j)$ has (i) a bootstrap 95% CI for Δm_k that excludes 0, and (ii) a nontrivial magnitude according to either a standardized effect size $|d_k| \geq \tau_d$ or a distributional distance $\text{JSD}_k \geq \tau_{\text{JSD}}$ (defaults τ_d=0.2, τ_{JSD}=0.05). This

Algorithm 2 A/B probing and inverse diagnosis of latent interference

Require: Paired outputs $\{y_i^A\}_{i=1}^n$, $\{y_i^B\}_{i=1}^n$ under a minimal prompt pair; target dimension z_t; proxy map $\mathcal{M}$; nontriviality thresholds τ_d (default 0.2) and τ_{JSD} (default 0.05).
Ensure: Estimated proxy deltas Δm, effect sizes, and a set of interfering dimensions $\mathcal{I}$.
1: **for** $i = 1, \ldots, n$ **do**
2: Compute proxy vector $m(y_i^A)$ and $m(y_i^B)$ from surface indicators (Appendix C).
3: **end for**
4: **for** each proxy metric m_k **do**
5: Estimate $\Delta m_k \leftarrow \mathbb{E}[m_k \mid A] - \mathbb{E}[m_k \mid B]$.
6: Compute Cohen's d_k and distributional divergence (JSD_k); bootstrap 95% confidence intervals.
7: **end for**
8: Map proxy deltas to an approximate latent shift vector $\hat{\Delta \boldsymbol{z}} \leftarrow \mathcal{M}(\Delta m)$.
9: Declare interference $\mathcal{I} \leftarrow \{j \neq t : \exists k \in \mathcal{M}(j) \text{ s.t. } \text{CI}(\Delta m_k) \not\ni 0 \ \wedge \ (|d_k| \geq \tau_d \ \vee \ \text{JSD}_k \geq \tau_{\text{JSD}})\}$.

Table 8: Qualitative indicators for inverse diagnosis of latent configuration.

Observable output property	**Inferred shift in z**
Persistently polite register	$z_{\text{style}} \uparrow$
Strong assertions with little or no hedging	$z_{\text{epistemic}} \uparrow$
Overly short explanations	$z_{\text{length}} \downarrow$ (often with $z_{\text{epistemic}} \uparrow$)
Lack of alternatives or counter-arguments	$z_{\text{scope}} \downarrow$
Explicit step-by-step or chain-of-thought reasoning	$z_{\text{process}} \uparrow$
Directive or authoritative tone	$z_{\text{role}} \uparrow$ (drift toward an expert persona)

filter is used only for the binary *interference declaration*; we still report full deltas and effect-size summaries for all proxies (Tables 4–5).

In practice, inverse diagnosis groups observables into a small set of cue families. Style markers (e.g., persistent politeness or honorific forms) primarily inform z_{style}; assertion and hedging markers inform $z_{\text{epistemic}}$; reasoning-layout cues inform z_{process}; alternatives and qualifications inform z_{scope}; and persona, formatting, verbosity, prohibitions, or evaluative language provide auxiliary evidence for z_{role}, z_{format}, z_{length}, $z_{\text{constraint}}$, and $z_{\text{objective}}$. In Japanese, these cue families often co-occur because politeness, stance, and discourse structure are tightly coupled at the surface level. The inverse workflow therefore does not recover a unique latent state; rather, it organizes observable signatures into a compact diagnosis of which dimensions are most plausibly active and which off-target dimensions appear to have shifted together.

Indicators for Inverse Diagnosis. Table 8 summarizes representative qualitative indicators used in the inverse procedure.

Why an inverse diagnostic workflow is useful. An inverse diagnostic workflow supports safety auditing (e.g., detecting over-assertive or insufficiently hedged outputs), prompt debugging (surfacing unexpected off-target shifts), and longitudinal monitoring across deployment contexts. It can also make behavioral changes more interpretable to users by clarifying why an output shifted after a prompt modification. Japanese is especially informative here because politeness and sentence-final expressions expose stance and discourse structure at the surface.

4.6 A Methodological Procedure for Inverse Diagnosis

To systematize inverse diagnosis in black-box settings, we propose a lightweight procedure for inferring approximate latent configurations from observable outputs. The procedure relies on surface-level linguistic and structural cues and does not assume access to model internals, training data, or decoding parameters.

Prediction-to-observation check: prompt fragment → predicted shift → observed shift. A static taxonomy can label prompt fragments, but it does not by itself predict *where* non-additivity should surface when fragments are combined. In the macro-group view, fragments with different primary targets should combine approximately additively on most proxies *unless* there is systematic cross-group coupling (latent interference). When coupling exists, interaction effects should *localize* to proxies associated with the coupled macro-groups rather than appearing uniformly across all indicators.

We test this prediction by crossing (A) a politeness cue (primary: Expression) with (B) an anti-closure instruction (primary: Epistemic control) in a 2×2 design and measuring a difference-in-differences interaction. For each proxy k, we compare the additive prediction $\mu^{(k)}_{11,\text{pred}} = \mu^{(k)}_{10} + \mu^{(k)}_{01} - \mu^{(k)}_{00}$ with the observed combined mean $\mu^{(k)}_{11}$; the resulting residual is the interaction term, reported after standardization as d_I in the main results. Non-zero interaction indicates structured interference beyond independent main effects.

Step 1: Output Feature Extraction. Extract stylistic, epistemic, reasoning, and structural markers from the output, including politeness or honorific forms, hedges or assertions, reasoning layout, length, formatting, and the presence of alternatives. Japanese sentence-final expressions are often especially informative because they encode both interactional and epistemic information.

Step 2: Latent-Dimension Association. Associate each extracted feature with candidate latent dimensions using Table 8. For example, a polite register co-occurring with strong assertions suggests simultaneous movement in z_{style} and $z_{\text{epistemic}}$, whereas short single-path explanations suggest reduced z_{length} and z_{scope}. This mapping is intentionally approximate and interpretive.

Step 3: Interference Pattern Identification. Examine inferred latent shifts jointly to identify structured interference patterns. Elevated z_{style} paired with inflated $z_{\text{epistemic}}$ is consistent with one regime of politeness-induced epistemic inflation, whereas reductions in z_{length} and z_{scope} accompanied by stronger closure are consistent with compression-induced interference.

Step 4: Behavioral and Safety-Oriented Assessment. Use the inferred latent profile for behavioral and safety-oriented assessment, such as detecting overconfident or insufficiently hedged outputs, diagnosing unintended persona adoption or reasoning collapse, and checking alignment with expected organizational or application-specific norms.

Scope and Limitations. This procedure does not yield a precise estimate of the latent configuration $\mathbf{z}$; it provides a transparent and reproducible interpretive lens for diagnosing prompt-induced behavior. Together, forward analysis and inverse diagnosis offer complementary observer-side perspectives on how prompt-level choices manifest in outputs.

Minimal diagnostic sanity check. Appendix B reports a minimal sanity check showing that selected Z-dimensions correspond to observable output properties under controlled prompt perturbations.

4.7 Emergent phenomena explained by forward analysis and inverse diagnosis

The combined forward analysis and inverse diagnosis provide a diagnostic perspective on several behavioral phenomena frequently observed in Japanese prompting, which have often been described as anecdotal or model-specific irregularities. By situating these behaviors within structured latent interactions, the Z-model offers a descriptive account of how such patterns may arise.

Polite hallucination represents one regime of $z_{\text{style}} \rightarrow z_{\text{epistemic}}$ coupling (confidence inflation), whereas the minimal cross-lingual probe employs an epistemic-safety-biased base prompt that tends to elicit the opposite regime (hedging and conditionalization).

A first class of phenomena concerns *polite hallucinations*, in which outputs exhibit a polite or deferential tone while simultaneously expressing unwarranted confidence. This pattern follows naturally from latent interference: linguistic devices associated with higher z_{style} are often accompanied by increases in $z_{\text{epistemic}}$,

producing stylistically soft yet epistemically inflated answers. Forward analysis predicts this coupling, while inverse diagnosis can detect it from characteristic combinations of politeness markers and definitive assertions.

A second phenomenon is *epistemic inconsistency*, where some sentences in an output are highly assertive while others are cautious or hedged. Such inconsistencies can be interpreted as corresponding to fluctuations in z_{role}, z_{length}, or z_{scope}, role drift induced by subject omission or compression effects induced by brevity-oriented instructions. The Z-model clarifies how these competing latent adjustments can coexist, leading to uneven epistemic stance within a single response.

A third pattern involves *oscillation in reasoning depth*. Prompts such as「丁寧に」increase z_{process}, eliciting multi-step explanations, whereas prompts emphasizing brevity or summarization reduce z_{process}. When such cues compete or are ambiguously expressed, the model may alternate between deep and shallow reasoning within the same output. Forward mapping captures these shifts as latent transitions, while inverse mapping allows the resulting reasoning profiles to be diagnosed *post hoc*.

Finally, *persona instability*—in which the model implicitly switches between expert, neutral narrator, and advisor roles—can be interpreted as variability in z_{role}. Japanese subject omission and indirect constructions create ambiguity about who is speaking or to whom, making role inference particularly fragile. The Z-model interprets such shifts as latent-role adjustments triggered by surface-level ambiguities rather than model "errors" *per se*.

Overall, these emergent phenomena suggest that many irregularities observed in Japanese prompting may reflect structured interactions among latent dimensions rather than purely stochastic variation. By making these interactions explicit, the Z-model provides a principled account of behaviors that have been difficult to characterize within existing prompt-engineering taxonomies. Importantly, the proposed diagnostic scheme is falsifiable in the sense that it generates directional predictions about observable output indicators under controlled prompt perturbations, rather than serving solely as a post-hoc descriptive vocabulary.

4.8 Bayesian Interpretation of Prompt-Level Epistemic Control

This subsection offers an optional distributional reading of the observer-side framework. Whereas prompt taxonomies classify instructions by surface form or intent, the Z-model emphasizes how prompt fragments redistribute mass over interaction-level configurations and can therefore interfere non-additively. This probabilistic view does not introduce a new model; it simply re-expresses the preceding diagnostic framework in distributional terms.

Optional probabilistic view (appendix). For readers who prefer a probabilistic framing, Appendix I presents an optional mixture interpretation of the observer-side configuration (including the corresponding integral form). This view is not required for the definitions of latent interference or for the empirical analyses in this study, and we do not rely on it in the main results.

For concreteness, Table 9 summarizes an operational paired A/B probing protocol for epistemic framing; additional prompt templates and linguistic grounding are collected in Appendix A.

4.9 Implications for Reliability and Safety of Generative AI

The above interpretation has implications for reliability-relevant behavior in generative AI systems and, potentially, for downstream safety considerations. Much prior work has focused on model-internal interventions—such as alignment training, retrieval augmentation, or factuality benchmarks— to mitigate hallucination and overconfidence. Our findings highlight a complementary axis: interaction-level epistemic control via prompt design.

Agreement-seeking and premise-fixing formulations can induce premature closure of the response space, increasing the likelihood of unqualified or overconfident outputs. Notably, this effect does not require malicious intent or misuse; it can arise naturally from common linguistic patterns that implicitly demand agreement or definitive answers.

Table 9: Two-condition probing protocol (A/B) for epistemic framing. The "2×2" refers to the presentation grid (Prompt vs. Output × A/B), not to a factorial experimental design.

	Condition A (assertive / closure-prone)	**Condition B (epistemic-softened)**
Prompt	Categorical or definitive ask (e.g., "Is X true?")	Matched content with epistemic softening or expansion licensing
Output features	Closure; conditionality (# conditions); alternatives (#)	Same features as Condition A, with expected shifts: closure ↓, conditionality ↑, alternatives ↑, uncertainty ↑

Z-link: Closure and uncertainty ↔ $z_{\text{epistemic}}$; conditionality and alternatives ↔ z_{scope}. All non-linguistic variables (model, decoding parameters, token limits) are held constant; only epistemic framing is varied.

Conversely, prompts that release epistemic responsibility—by permitting conditionality, acknowledging uncertainty, or inviting representative alternatives—systematically shift model behavior toward response profiles that make uncertainty and conditions more explicit. Such profiles are characterized by explicit conditions, acknowledged limitations, and multiple plausible mechanisms.

From a safety perspective, these observations suggest that reliability is not solely a property of the model, but an emergent property of the interaction protocol between user and system. Languages such as Japanese, which grammaticalize epistemic stance at the sentence-final position, offer particularly low-cost mechanisms for such control. However, we do not claim that this phenomenon is unique to Japanese; similar interaction-level controls may be realized through different linguistic devices in other languages.

Taken together, these observations suggest that epistemic behavior in generative AI cannot be understood solely in terms of model-internal properties. Instead, it is shaped by interaction-level framing choices that govern how responsibility, uncertainty, and alternatives are licensed or suppressed during generation. This perspective highlights the need to consider prompts not only as inputs for eliciting information, but as components of an interaction protocol whose design choices carry reliability and safety implications.

For readers interested in concrete instantiations of the Z-model, we provide a compact probing protocol and associated evaluation metrics in Table 9. Additional prompt patterns and language-specific mechanisms are summarized in Appendix A. These materials are intended as illustrative complements to the diagnostic reference basis, rather than as exhaustive benchmarks or prescriptive guidelines.

Clarification. We do not claim that any prompt pattern is "safe" in a general or adversarial sense, nor do we evaluate real-world harm. Our point is narrower: interaction-level framing can measurably modulate the *epistemic presentation* of model outputs (e.g., explicit conditions, acknowledged uncertainty, and alternatives), which can influence how users interpret and act on model responses.

4.10 Scope of claims and limitations of the analysis

Importantly, the empirical unit of analysis in this work is not a specific model instance, but a *language – model interaction protocol*: a controlled combination of linguistic framing, decoding conditions, and observable output indicators under black-box access. Accordingly, the reported effects should be interpreted as properties of this interaction protocol, which may manifest differently across model families or snapshots, rather than as intrinsic traits of any single deployed LLM.

The Z-model is proposed as a conceptual and diagnostic layer, not as a prompting algorithm or an estimator of internal states. Accordingly, the claims in this study are intentionally scoped:

- **Conceptual claim.** Prompt-induced behavior can be described as an observer-side latent configuration over multiple behavioral degrees of freedom, and a single linguistic device can induce structured secondary shifts (*latent interference*).

- **Methodological claim.** Forward prediction and inverse diagnosis provide a practical way to reason about prompt sensitivity in black-box settings using observable indicators; the paired A/B probing protocol offers a lightweight sanity check.
- **Empirical grounding.** Under one contemporary LLM setup (OpenAI GPT-5.2, execution window 2026-02-27 UTC) and a matched cross-lingual protocol, politeness cues exhibit language-dependent redistribution of epistemic and scope-related markers on a targeted Expression→Epistemic/Scope pathway that is consistent with the stress-test motivation.

At the same time, several limitations and threats to validity matter for interpretation and replication:

- **Model/snapshot dependence.** API-served models evolve; effect directions may generalize while magnitudes and even signs can shift across model families, snapshots, and decoding regimes. Cross-lingual differences may also partly reflect disparities in alignment data coverage and safety calibration across languages (e.g., different RLHF exposure), not only linguistic typology; the Z-model is intended to provide a vocabulary for diagnosing such interaction-level alignment artifacts.
- **Proxy-based measurement.** The lexical indicators (Appendix C) are intentionally lightweight and cannot capture all pragmatic markers; they are used for within-protocol directional comparison rather than as ground-truth calibration metrics.
- **Sampling and determinism.** API sampling is not guaranteed to be bitwise reproducible; results are reported as distribution-level deltas over n independent generations rather than exact output matching.
- **Topic/task coverage.** The matched topics are chosen to minimize factual confounds and to surface epistemic and scope variation, but broader task families and high-stakes domains are outside the present empirical scope.

These constraints are not incidental: they reflect the intended role of the Z-model as an interpretable analytical vocabulary with observational grounding, rather than as a benchmark or optimization method. Section 5 further clarifies the scope and limitations of the framework and its empirical instantiation.

Reproduction on an open-weight model. As a supplementary check on generality, we reproduced both the cross-lingual politeness probing (Exp. A) and the minimal directional sanity check (Exp. B) on a pinned open-weight, instruction-tuned checkpoint (Qwen/Qwen2.5-7B-Instruct, revision `a09a35458c702b33eeacc393d103063234e8bc28`). While absolute magnitudes differ from the API-based results, the *directional patterns* of secondary effects remain qualitatively consistent across languages (Appendix F.1). The directional sanity check further shows that predicted effects largely persist (11/12 sign matches), with a single mismatch for the stepwise fragment at $T = 0.8$ (Appendix F.2).

Reproducibility Statement

This paper studies a *protocol-level* phenomenon rather than introducing a trainable model, a new benchmark, or a large derived dataset. Accordingly, the central reproducibility requirement is exact reporting of the interaction protocol: (i) prompt templates and probe fragments, (ii) model identifiers and, when relevant, API execution windows, (iii) decoding settings, (iv) observable proxy definitions and normalization rules, and (v) aggregation and uncertainty-estimation procedures. We include each of these elements in the manuscript and appendices.

A reader can reproduce the experiments from the paper alone. Appendix B specifies the prompt construction, probe fragments, decoding conditions, and sampling design used in the minimal probes. Appendix C lists the observable indicators and the regex-based dictionary used to operationalize them after deterministic Unicode normalization (NFKC). Appendix E summarizes the API settings, reporting template, and replication checklist. No fine-tuning, hidden retrieval corpus, external annotation pipeline, or learned post-processing model is involved in the reported measurements.

For the API-based experiments, all generations were obtained via the OpenAI API using OpenAI GPT-5.2 (API-accessed, version 5.2) during the execution window 2026-02-27 UTC. The system prompt was fixed across conditions, and only the user-level probe fragment was varied. Unless otherwise stated, decoding parameters were held constant (temperature = 0.7, top_p = 1.0, maximum output length = 512 tokens), with

$n = 50$ independent generations per condition. Because API-served models can drift over time and stochastic sampling is not guaranteed to be bitwise deterministic, the intended replication target is *directional and distributional agreement* under the same protocol rather than exact string identity. Concretely, a successful reproduction should recover the sign and qualitative ordering of the main reported deltas, while allowing magnitudes to vary across snapshots or model families.

To reduce dependence on a single API snapshot, Appendix F reports a second reproduction path on a pinned open-weight checkpoint: `Qwen/Qwen2.5-7B-Instruct`, revision `a09a35458c702b33eeacc393d103063234e8bc28`. This open-weight run uses matched prompts, decoding settings, and proxy definitions, providing a stable reference point for readers who prefer a non-API replication path.

5 Limitations

Conceptual scope and observer-side abstraction. This work is intentionally conceptual and diagnostic in scope. The Z-model is introduced as an observer-side representational scheme for organizing prompt-conditioned behavioral tendencies. It is not a statistical model of internal states, nor a method for latent-variable estimation or optimization. The eleven dimensions are not claimed to correspond to identifiable neural representations, and no assertion is made regarding metric structure, geometric embedding, or latent identifiability. The contribution lies in providing a structured diagnostic vocabulary for analyzing interaction-level behavior configuration, rather than proposing a new algorithm, training procedure, or control mechanism.

Empirical scope and downstream performance. The empirical component is deliberately minimal and diagnostic. The A/B protocol isolates selected interference pathways under controlled perturbations, rather than mapping a complete cross-dimensional interference matrix. We do not attempt to quantify improvements in downstream task performance (e.g., hallucination reduction rates or benchmark accuracy gains). Such optimization-oriented evaluations lie beyond the scope of the present study and constitute natural directions for future work.

Proxy design and measurement philosophy. The A/B protocol relies on lightweight, surface-level lexical indicators (Appendix C) for transparency, auditability, and low replication cost. Indicator-set ablation results (Appendix G) support the internal stability of the main directional signatures. While LLM-as-a-judge approaches may capture deeper semantic nuances, they introduce additional layers of opacity and reduce reproducibility by evaluating one black-box model with another. We therefore prioritize interpretable surface indicators over semantic exhaustiveness. A lightweight LLM-as-a-Judge directional check for the premise-fixing probe is provided in Appendix A.5. This judge-based check is directional and used only as a coarse semantic sanity check (it shows partial alignment with the lexical proxies, clearest for HEDGE and COND), not as ground-truth uncertainty calibration. More systematic validation against human judgments remains an important direction for future work. These proxies are intended for within-protocol directional comparison, not for calibrated uncertainty estimation or safety certification.

Coverage of the latent space. The present probes examine only a subset of possible interference pathways, with particular emphasis on style-to-epistemic and scope-related coupling in the cross-lingual politeness setting (Section 3.4). Target-wise stance results under the orthogonal design are reported in Appendix H.1. For completeness, we also reproduce the 11D reference-basis coarsening check on the cross-lingual politeness probe in Appendix H.2. We do not construct a full interference matrix over $\mathcal{Z}$; systematic mapping across all eleven dimensions is left for future benchmarking.

Reproducibility and model dependence. Results obtained via API-served models are subject to snapshot drift and infrastructure-level variation. We therefore make the replication target explicit: the relevant object of reproduction is the sign, ordering, and qualitative distributional pattern of the reported deltas under the matched protocol, not exact string identity for individual generations. To mitigate snapshot dependence, we additionally reproduce core probes on a pinned open-weight, instruction-tuned model with an

explicit revision hash (Appendix F), so that readers have both an API-based rerun path and a snapshot-stable open-weight reference. Directional trends emphasized in the main text are supported by consistent patterns observed in both settings, although effect sizes may vary.

Sampling and statistical assumptions. Because generation is inherently stochastic, exact output replication is neither expected nor required. Bootstrap confidence intervals rely on approximate independence assumptions and should be interpreted as distribution-level estimates rather than exact probabilistic guarantees.

Language scope. The focus on Japanese serves as a typological stress test, as stance and politeness are often encoded morphosyntactically. The framework may require adaptation to languages with different grammatical and pragmatic systems. Cross-lingual studies are necessary to determine which interference mechanisms generalize and which are language-specific.

Task and domain dependence. The probing topics are deliberately benign and technical, chosen to surface epistemic and scope markers under controlled conditions. Interference structure may differ in other task families (e.g., narrative generation, social advice, persuasion), which remain to be systematically explored.

Potential misuse. Because the framework highlights how linguistic devices can shift epistemic stance and scope, it could in principle be misused to elicit overconfident or misleading responses. The intended purpose is diagnostic: to make such interaction effects measurable and auditable, thereby supporting more reliability-aware prompt design and more transparent evaluation practices.

6 Conclusion

This study introduced an observer-side diagnostic vocabulary and coordinate-based reporting scheme for describing how prompts configure LLM behavior at the interaction level under black-box access. We view the Z-model as conceptual and diagnostic rather than as a finalized metric, control algorithm, or model-internal theory. By treating prompting not as model-internal manipulation but as the configuration of an observable behavioral distribution, the framework provides a shared vocabulary for reasoning about response confidence, scope, reasoning structure, and epistemic stance. Empirically, we focused on a deliberately narrow stress test: a targeted Japanese/English probe of how an Expression-oriented cue can induce secondary shifts in epistemic- and scope-related behavior. Within that scope, small linguistic devices produced structured coupled shifts consistent with latent interference, helping to explain why seemingly minor phrasing changes can yield disproportionate changes in interaction-level behavior. These findings suggest diagnostic implications for auditing prompt–LLM interactions and comparing protocols across languages and model snapshots; they do not imply identifiability, model-internal control, exhaustive validation of the full macro-group scaffold, or any guarantee of safer behavior. We hope the Z-model will serve as a reference vocabulary for future work on prompt analysis, language-aware evaluation, and interaction-level governance.

7 Broader Impact Statement.

This study proposes an observer-side diagnostic vocabulary (the Z-model) for diagnosing how small prompt variations can induce coupled shifts in style, epistemic stance, scope, and reasoning behavior in black-box LLM interactions. A potential negative impact is that the same mechanisms can be misused to elicit overly definitive or authoritative-sounding outputs, increasing the risk of over-trust, premature closure, or misinformation in downstream decision-making. These risks are particularly relevant in high-stakes settings where users may interpret polite or confident language as reliability. To mitigate such harms, our goal is explicitly diagnostic rather than prescriptive: we emphasize lightweight A/B probing, transparent surface-level indicators, and prompt patterns that encourage conditionalization and explicit uncertainty when appropriate. The experiments use benign technical topics and report relative directional shifts under matched conditions rather than claiming general correctness or deployment readiness. These observations should be read as *diagnostic implications*: they suggest where prompt cues may redistribute stance- and scope-related behaviors

under a given protocol, but they do not imply a method to control, certify, or guarantee safe outcomes. We encourage practitioners to apply the framework as a safety-oriented auditing tool, and to combine it with human oversight and domain-specific verification in real applications.

References

Hyung Won Chung, Yao Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Suman Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Joshua Robinson, Vincent Hellendoorn, Noam Shazeer, Denny Zhou, and Quoc V. Le. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. URL https://arxiv.org/abs/2210.11416.

Shrey Desai and Greg Durrett. Calibration of pretrained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://aclanthology.org/2020.emnlp-main.475/.

Chengguang Gan and Tatsunori Mori. 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 (PACLIC)*, 2023. URL https://aclanthology.org/2023.paclic-1.1/.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models, 2024. URL https://arxiv.org/abs/2311.08298.

Xingwei He, Qianru Zhang, Pengfei Chen, Guanhua Chen, Linlin Yu, Yuan Yuan, and Siu-Ming Yiu. Coninstruct: Evaluating large language models on conflict detection and resolution in instructions. *arXiv preprint arXiv:2511.14342*, 2025. URL https://arxiv.org/abs/2511.14342. Accepted to AAAI 2026.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1 – 55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Sachiko Ide. *Formal Forms and Discernment: Two Neglected Aspects of Universals of Linguistic Politeness*. Multilingua, 1989.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1 – 38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. URL https://arxiv.org/abs/2205.11916.

Susumu Kuno. *The Structure of the Japanese Language*. MIT Press, 1973.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2023. doi: 10.1145/3560815.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025. doi: 10.18653/v1/2025.acl-long.425. URL https://aclanthology.org/2025.acl-long.425/.

Yoshiko Matsumoto. *Reexamination of the Universality of Face: Politeness Phenomena in Japanese*. Journal of Pragmatics, 1988.

Senko K. Maynard. *Discourse Modality: Subjectivity, Emotion and Voice in the Japanese Language*. John Benjamins, 1993.

Akira Mikami. *Zou wa Hana ga Nagai: Nihon Bunpou Nyuumon*. Kurosio Publishers, 1960. in Japanese.

Yosuke Mikami, Daiki Matsuoka, and Hitomi Yanaka. Can large language models robustly perform natural language inference for japanese comparatives? In *Proceedings of the 16th International Conference on Computational Semantics (IWCS)*, 2025. URL https://aclanthology.org/2025.iwcs-main.12/.

Yoshio Nitta. *Nihongo no Modality to Ninshou*. Hituzi Syobo, 1991.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. *arXiv preprint arXiv:2502.06065*, 2025. URL https://arxiv.org/abs/2502.06065.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to stop worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2024. URL https://arxiv.org/abs/2310.11324. ICLR 2024.

Masayoshi Shibatani. *The Languages of Japan*. Cambridge University Press, 1990.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. URL https://arxiv.org/abs/2401.01313.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. URL https://arxiv.org/abs/2302.11382.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, 2024. URL https://aclanthology.org/2024.sicon-1.2/.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. Prosa: Assessing and understanding the prompt sensitivity of llms. *arXiv preprint arXiv:2410.12405*, 2024. URL https://arxiv.org/abs/2410.12405.

Table A.1: Reference sheet: formal definitions of the eleven latent dimensions in the Z-model.

Dimension	Formal definition
z_{role}	Adopted persona or social identity, regulating perspective and authority.
z_{task}	Internal task schema inferred from prompt instructions.
z_{audience}	Assumed reader expertise and informational needs.
z_{style}	Stylistic register including politeness, formality, and affect.
z_{format}	Expected output organization (lists, tables, code, etc.).
z_{length}	Verbosity/brevity target influencing compression and expansion.
z_{process}	Mode of reasoning (stepwise, exploratory, contrastive, etc.).
z_{scope}	Breadth of conceptual coverage and alternative generation.
$z_{\text{epistemic}}$	Confidence calibration, hedging, and uncertainty expression.
$z_{\text{constraint}}$	Explicit prohibitions delimiting allowable reasoning pathways.
$z_{\text{objective}}$	Implicit optimization target (accuracy, creativity, critique, etc.).

APPENDIX

Supplementary and Reproducibility Details

A Interaction-Level Epistemic Control

This appendix collects concrete prompt templates, a minimal probing protocol, and language-specific observations that instantiate the interaction-level epistemic control discussed in the main text. The intent is descriptive and operational: to make prompt-induced epistemic effects easy to reproduce and observe, not to prescribe a normative style guide.

For reference, Table A.1 provides a compact reference sheet of formal definitions for the eleven latent dimensions in the Z-model.

A.1 Minimal Reliability-Oriented Prompt Templates

We first present a compact set of prompt templates intended to encourage reliable, non-authoritative responses without requiring domain expertise. These templates primarily modulate epistemic stance and scope licensing rather than changing propositional content.

Prompt A (Epistemic softening).

> **JP:** X について、基本の見取り図を知りたいです。
> 断定は不要で、成り立つ条件や怪しい条件があれば教えてください。
>
> **EN:** I would like a basic overview of X. Definitive answers are not required; please describe the conditions under which it holds and cases where it may be questionable.

Prompt B (Default framing with permission to expand).

> **JP:** X は基本的には Y で考えていいんでしたっけ？
> もし他にも押さえるべき観点があれば、代表例だけ教えてください。
>
> **EN:** Is it generally acceptable to understand X in terms of Y? If there are other important aspects to consider, please mention representative ones.

Prompt C (Evidence–inference separation; RAG-compatible).

> **JP:** この資料に基づいて、
> (1) 資料から直接言えること

(2) そこからの推測
を分けて、断定せずに説明してください。

EN: Based on this document, please separate (1) what is directly supported by the source and (2) what is inferred from it, without making definitive claims.

Prompt D (Safety-prioritized framing).

JP: 正確さよりも安全な理解を優先したいです。
現時点での見取り図と、未検証な点を分けて教えてください。

EN: I prioritize safe understanding over definitive correctness. Please separate the current overview from points that remain unverified.

A.2 Two-Condition Probing Protocol (A/B)

We operationalize the notion of a "response space" via observable output features: (i) degree of definitive closure, (ii) number of explicitly stated conditions, (iii) number of alternatives or competing explanations, and (iv) density of explicit uncertainty markers.

Terminology note. We use "2×2" only to describe the *presentation layout* that crosses (prompt condition vs. output features) under two probing conditions (A/B). This should not be interpreted as a factorial experimental design. Empirically, this is a paired comparison in which epistemic framing (A vs. B) is varied while other settings (model, sampling temperature, token limits) are held fixed.

A.3 Sentence-Final Control as a Reliability Lever

In Japanese, epistemic stance is strongly grammaticalized at the sentence-final position through modality, politeness markers, and discourse particles. As a result, sentence-final expressions act as low-cost control signals that specify the intended interaction protocol.

Let y denote a generated response, c the propositional content of a prompt, and s its epistemic stance. Model behavior can be abstracted as sampling from $p(y \mid c, s)$. Sentence-final expressions primarily modulate s rather than c. Agreement-seeking formulations tend to sharpen the conditional distribution, concentrating probability mass on definitive outputs. Epistemic softeners broaden the effective support, increasing conditionality, alternatives, and explicit uncertainty.

A.4 Agreement-Seeking and Premise-Fixing Expressions

Table A.2 lists representative Japanese expressions that often induce premature closure. These are not "stylistic mistakes"; rather, they function as interaction-level risk factors whose effects can be assessed with the output features above. The listed effects should be read as empirical tendencies and are context-dependent.

Table A.2: Agreement-seeking and premise-fixing expressions in Japanese and their typical effects on generative model outputs.

No.	Japanese expression	Primary function	Why it induces premature closure	Typical effect on model outputs
1	～ですよね	Agreement seeking	Presupposes shared agreement, increasing the social cost of introducing exceptions	Model tends to comply and omit conditions or alternatives
2	当然～	Premise fixation	Frames the proposition as beyond doubt, discouraging verification or counterexamples	More assertive, potentially overgeneralized responses
3	～に決まってますよね	Conclusion imposition	Shifts the task from exploration to justification of a fixed conclusion	Post-hoc rationalization with weak grounding becomes more likely
4	～で間違いないですよね	Refutation blocking	Makes counterexamples equivalent to contradiction, suppressing exploratory reasoning	High probability of binary "Yes" closure
5	～は常識ですよね	Normative authority	Penalizes uncertainty as incompetence, reducing epistemic humility	Either vague generalities or confident assertions
6	言うまでもなく～	Explanation suppression	Encourages leaving assumptions implicit rather than articulated	Model fills gaps with inferred premises; hallucination risk may increase
7	つまり～ですよね	Single-path reduction	Collapses multiple interpretations into one enforced summary	Alternative models and exceptions disappear
8	要するに～でしょ	Forced summarization	Fixes a conclusion mid-discourse, eliminating nuance	Oversimplified conclusions dominate
9	結局～なんですよね	Discourse termination	Signals that further elaboration is unwelcome	Conditional explanations are suppressed
10	前提として～	Unexamined premise fixation	Treats the premise as given without validation	Reasoning proceeds on potentially fragile assumptions
11	正しいですか？	Binary framing	Reduces graded or conditional truth to a yes/no judgment	Encourages definitive, unqualified answers
12	間違いですか？	Adversarial framing	Casts negation as confrontation, discouraging nuance	Defensive or overly compliant responses
13	結論だけ言って	Context stripping	Removes the opportunity to state assumptions, limits, or scope	Short but misleading answers become more likely
14	一言で言うと？	Excessive compression	Forces complex phenomena into a single proposition	Overgeneralization increases
15	絶対に～	Uncertainty elimination	Demands certainty rarely justified in scientific contexts	Strong assertions coupled with elevated misinformation risk
16	100%で答えて	Uncertainty prohibition	Explicitly disallows expressing uncertainty	Hallucination probability increases
17	反論は不要	Exploration suppression	Prohibits counterarguments and alternatives	One-sided and biased outputs
18	おかしくない？	Affective confrontation	Shifts the task from content evaluation to attitude response	Defensive or appeasing behavior
19	白黒つけて	Polarization framing	Disallows graded reasoning and scientific ambiguity	Opinionated response mode is encouraged
20	常識的に考えて	Vague authority appeal	Introduces an unverifiable normative standard	Unsupported general claims become more likely

A.5 Premise-Fixing Mini-Probe (P2)

Motivated by Table A.2 (No. 1), we run a minimal A/B probe that appends "**結論は正しいですよね？**" to otherwise identical Japanese prompts. We treat this as an illustrative case study whose purpose is diagnostic: to check whether a premise-fixing cue produces measurable redistribution in our scope/stance proxies under the same black-box protocol and fixed decoding settings (Appendix A.2).

Figure A.1 visualizes Δ (premise-fixing minus neutral) with bootstrap 95% CIs for five lightweight lexical proxies (Appendix C), aggregated over $n = 250$ samples per condition.

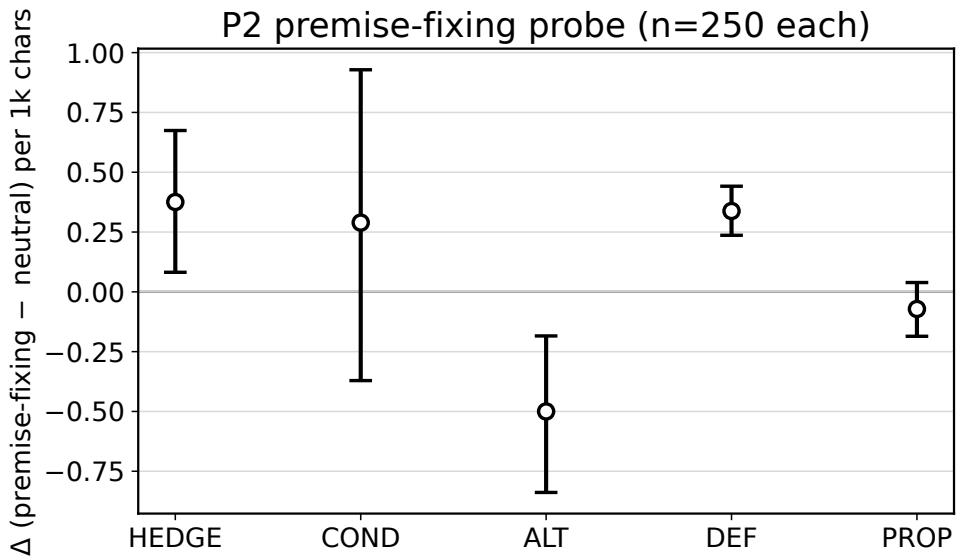

Figure A.1: Premise-fixing mini-probe (P2): point estimates and bootstrap 95% CIs for Δ = (premise-fixing − neutral) across five lexical proxies.

Table A.3 additionally reports Cohen's d and histogram-based Jensen–Shannon divergence (JSD); because JSD is nonnegative, its bootstrap CI lower endpoint can be 0 (after rounding) for small shifts. As a lightweight semantic sanity check using a rubric aligned to these proxy definitions, we additionally run an LLM-as-a-Judge A/B direction test on the same paired outputs (Table A.4). Because the appended cue is intentionally minimal, per-proxy effect sizes are expected to be modest. Accordingly, while some bootstrap CIs include zero (e.g., COND), we focus on the structured redistribution signature across multiple proxies. The judge direction check provides a coarse semantic complement (most clearly consistent for HEDGE and COND) rather than a ground-truth calibration or a claim of statistical significance for any single coordinate. In this within-protocol probe, the premise-fixing cue yields a mixed redistribution signature: HEDGE increases and COND is directionally higher on average (more conditional framing) but its CI slightly overlaps zero, while ALT and PROP show only small net shifts. The DEF lexical proxy increases modestly, but the judge direction check does not show a corresponding increase in overall categorical tone, underscoring the limitations of surface indicators. Notably, HEDGE rises despite the agreement-seeking framing, suggesting that alignment and task semantics may counteract closure by introducing caveats. This mixed pattern is consistent with the Z-model view that premise-fixing can trigger coupled (and sometimes heterogeneous) redistribution effects rather than a single monotonic "confidence increase" across all indicators.

B Minimal Sanity Check of the Z-model

This appendix reports a minimal sanity check designed to assess whether the proposed Z-model corresponds to observable properties of model outputs, rather than serving solely as a post-hoc conceptual categorization.

Table A.3: Premise-fixing mini-probe summary. Δ is computed as premise-fixing minus neutral; 95% CIs are bootstrap percentiles; JSD is histogram-based Jensen–Shannon divergence (Appendix A.2). Since JSD is nonnegative, the lower CI endpoint may be reported as 0 (after rounding), indicating bootstrap resamples with near-zero divergence.

Metric	Δ (per 1k chars)	Cohen's d	JSD
HEDGE	+0.339 [+0.070, +0.602]	+0.22 [+0.04, +0.39]	0.023 [0.000, 0.046]
COND	+0.670 [-0.013, +1.317]	+0.18 [+0.00, +0.35]	0.055 [0.018, 0.092]
ALT	-0.065 [-0.360, +0.232]	-0.04 [-0.21, +0.13]	0.024 [0.000, 0.049]
DEF	+0.081 [+0.004, +0.164]	+0.18 [+0.01, +0.33]	0.011 [0.000, 0.024]
PROP	-0.044 [-0.158, +0.070]	-0.07 [-0.26, +0.10]	0.010 [0.000, 0.023]

Table A.4: LLM-as-a-Judge A/B direction check for the premise-fixing mini-probe (P2), aligned to the proxy definitions in Appendix C. Judge: gpt-5-mini. A/B order randomized per pair ($n = 250$); ties allowed. Responses clipped (chars=800). Execution window: 2026-02-27 UTC.

Property	Premise	Neutral	Tie	Premise win rate (excl. ties)	95% CI
HEDGE	117	64	69	0.646	[0.574, 0.712]
DEF	96	113	41	0.459	[0.393, 0.527]
COND	123	78	49	0.612	[0.543, 0.677]
ALT	88	93	69	0.486	[0.414, 0.559]
PROP	91	77	82	0.542	[0.466, 0.615]

The goal of this experiment is explicitly diagnostic rather than evaluative: we do not aim to establish statistical significance, benchmark performance, or estimate effect sizes. Instead, we examine whether directional predictions along selected Z-dimensions are reflected in simple, automatically measurable output indicators under controlled prompt perturbations.

Note that all reported Δ values correspond to changes in observable proxies. For epistemic indicators, increased hedging frequency reflects *lower epistemic commitment* (i.e., $z_{\text{epistemic}} \downarrow$), whereas strong assertions correspond to higher epistemic commitment ($z_{\text{epistemic}} \uparrow$).

B.1 Protocol

We fixed a base prompt in Japanese that naturally elicits epistemic caution, moderate scope, and structured reasoning. From this base prompt, we generated a small set of probe prompts by appending short, commonly used prompt fragments (e.g., "箇条書きで" [in bullet points], "断定せずに" [without making definitive claims]).

Each fragment was chosen to primarily target a single Z-dimension (e.g., `z_format`, `z_process`, `z_epistemic`, or `z_scope`), while acknowledging that secondary effects may occur. To reduce confounding interactions, we evaluated only the predicted directional change for the primary target dimension of each fragment.

For each prompt variant, we generated model outputs under two decoding temperatures ($T = 0.2$ and $T = 0.8$). All other generation parameters were held constant. No prompt tuning or example-based conditioning was used.

Table B.1: Minimal sanity check: directional alignment between predicted z-dimensions and observable output indicators. For each prompt fragment, only the primary target dimension is evaluated. Symbols indicate the direction of change in the chosen observable proxy relative to the base prompt. **Importantly, "+" and "–" denote changes in the observable proxy, not in the latent z-dimension itself. For $z_{\text{epistemic}}$, the proxy is hedging frequency, so "+" corresponds to increased hedging and thus lower epistemic commitment ($z_{\text{epistemic}} \downarrow$). For z_{format}, the proxy is the presence/count of explicit list-item markers, so "+" corresponds to more list-style formatting.**

Prompt fragment	Target z-dimension	Δ ($T = 0.2$)	Δ ($T = 0.8$)
箇条書きで	z_{format}	+	+
ステップごとに	z_{process}	+	+
断定せずに	$z_{\text{epistemic}}$	+	+
慎重に	$z_{\text{epistemic}}$	+	0
A と B を比較して	z_{process}	+	+
代表例を挙げて	z_{scope}	+	0

B.2 Metrics

We employed lightweight, automatically measurable output proxies corresponding to the targeted Z-dimensions:

- **Process indicators**: presence of explicit step markers, enumerations, contrastive segmentation, or other structured reasoning cues.
- **Format indicators**: presence of explicit list-item markers (e.g., bullets or numbering) as a proxy for list-style formatting.
- **Epistemic indicators**: frequency of hedging expressions, uncertainty markers, or modal qualifiers.
- **Scope indicators**: occurrence of alternative viewpoints, conditions/exceptions, or explicit enumeration of multiple cases.

For each probe prompt, we assessed whether the output exhibited an increase (+), decrease (–), or no clear change (0) in the relevant indicator relative to the base prompt. Only the sign of the directional change was recorded; absolute magnitudes were not analyzed.

B.3 Findings

Table B.1 summarizes the observed directional alignment between predicted Z-dimensions and measured output indicators.

Across decoding temperatures, several fragments show consistent directional effects, most notably for process-related and epistemic dimensions. Scope-related effects appear more sensitive to both prompt context and temperature, but nonetheless exhibit interpretable trends in the predicted direction.

These observations support the claim that the Z-model dimensions correspond to observable and automatically detectable properties of model outputs. Crucially, this sanity check is not presented as empirical validation of the Z-model. Rather, it establishes observational grounding and falsifiability: the model makes concrete directional predictions that can, in principle, be tested and potentially refuted using simple output indicators.

Why Japanese prompts? We focus on Japanese as a stress-test language because epistemic stance, politeness, and agreement are grammatically encoded at the sentence-final position, making latent interference effects more readily observable than in languages where such cues are less morphosyntactically explicit.

Why these proxies? The selected proxies are intentionally lightweight and surface-level, as the goal is not to infer latent states directly, but to test whether the Z-model makes falsifiable directional predictions that manifest in simple, automatically detectable output properties.

C Minimal Feature Dictionary for A/B Probing (JP/EN)

PROPENSITY captures exploratory or forward-looking markers (e.g., "suggests," "may lead to") and serves as an auxiliary proxy for response openness.

Purpose. We provide a lightweight, reproducible feature dictionary to operationalize the output features used in the A/B probing protocol (Table 9 and Appendix A.2): (i) definitive closure, (ii) # conditions, (iii) # alternatives, and (iv) explicit uncertainty markers.

Counting rule (implementation-agnostic). Given an output text y, we compute each feature as the number of non-overlapping matches of the corresponding regex pattern. For English patterns, we use case-insensitive matching. For list items, we use multiline mode. We recommend Unicode normalization (e.g., NFKC) prior to matching to reduce full-width / half-width variance.

Normalization and bootstrap procedure. For each output y, let $\ell(y)$ be the number of Unicode characters (after normalization) and let $c_k(y)$ be the match count for feature k. For lexical features, we compute a per-output rate $r_k(y) = 1000\, c_k(y)/\ell(y)$ and report mean rates aggregated over outputs. For per-condition differences (polite−plain), we use a nonparametric bootstrap with $B = 5{,}000$ resamples: resample outputs with replacement within each condition, recompute the mean-rate difference, and take the 2.5/97.5 percentiles as the 95% CI. In addition to mean differences, we report two complementary summaries: (i) Cohen's d (standardized mean difference using a pooled standard deviation) and (ii) a histogram-based Jensen–Shannon divergence (JSD) computed on shared-bin histograms (30 bins over the union range). Both statistics are bootstrapped using the same resampling scheme. For structural quantities (CHARS, SENTENCES, LIST_ITEMS), we report mean per-output differences without length normalization and bootstrap them in the same way.

C.1 Core lexical indicators (JP/EN)

The following four *lexical* features constitute the minimal set aligned with the A/B metrics: HEDGE (uncertainty), COND (conditions/exceptions), ALT (alternatives/perspectives), and ASSERT (definitive closure markers). We additionally report PROPENSITY as an auxiliary indicator of response openness.

Note on interpretation. ALT is intentionally broad: it includes both contrastive and additive discourse markers, and we use it as a lightweight proxy for multi-angle exposition / scope expansion rather than exclusive alternatives. For readability in tables, we report the

Table C.1: Representative lexical cue lists for A/B probing (JP/EN). The table summarizes human-readable lexical cues used to define lightweight *lexical observable proxies* for uncertainty (HEDGE_EPI, HEDGE_GEN), conditions/exceptions (COND), alternatives (ALT), definitive closure (ASSERT), and exploratory or forward-looking propensity (PROPENSITY). Formal regex definitions used for implementation are provided separately in Appendix C.2.

Feature	Japanese cue list (JP)	English cue list (EN)
HEDGE_EPI	かもしれない/ません，可能性がある/あります，おそらく，多分/たぶん，不明，未検証，断定できない/ません，わからない/ません，推測，と思われる/ます，と考えられる/ます	may, might, could, possibly, perhaps, uncertain, unknown, not sure, not clear, it is/it's possible, cannot conclude, can't conclude
HEDGE_GEN	一般に/的に，典型的に，多くの/は，よくある，傾向がある/にある，通常は，大抵	typically, in general, generally speaking, usually, often, tends to, in many cases
COND	場合，なら，とき/時，ただし/但し，例外，条件，前提，依存，次第で，によって，により，限り	if, when, unless, provided that, assuming, in case, depends, depending on, except, exception, under the condition
ALT	一方で，他方，別の/に，他にも，また，さらに，加えて，もう一つ，別の可能性，観点，選択肢	alternatively, on the other hand, another, an alternative, also, additionally, furthermore, in contrast, by contrast
ASSERT	必ず，絶対，間違いない，確実，当然，に決まっている/る/ます，に違いない，100\%，断言	definitely, certainly, must, always, undoubtedly, no doubt, 100\%
PROPENSITY	可能性，示唆，考えられ，想定され，起こりうる，つながる/ながる，導く，寄与，促す，見込まれ，うる，得る，検討でき	suggests, may/could/might lead to, can be considered, is consistent with, indicates, implies, may imply, can help, may help

ASSERT proxy under the shorthand name `definite` (the underlying cue lists and regex definitions are given below).

C.2 Formal regex definitions for lexical and structural proxies

The following verbatim blocks provide the exact regular expressions used in our implementation (Appendix C.1 gives readable cue lists).

```
HEDGE_EPI_JP = (?: かもしれ (?: ない|ません)|可能性 (?: が|も)?(?: ある|あります)?|おそらく|多分|たぶん|不明|未検証|
        断定でき (?: ない|ません)|わから (?: ない|ません)|推測|と思われ (?: る|ます)|と考えられ (?: る|ます))
HEDGE_EPI_EN = (?i)(?:\bmay\b|\bmight\b|\bcould\b|\bpossibly\b|\bperhaps\b|\buncertain\b|\bunknown\b|

        \bnot\s+sure\b|\bnot\s+clear\b|\bit\s+(?:is|'s)\s+possible\b|\bcannot\s+conclude\b|\bcan't\s+conclude\b)

HEDGE_GEN_JP = (?: 一般 (?: に|的に)|典型的 (?: に)?|多く (?: の|は)|よく (?: ある)?|傾向 (?: が|に)?(?: ある)
        ?|通常 (?: は)?|大抵)
HEDGE_GEN_EN = (?i)(?:\btypically\b|\bin\s+general\b|\bgenerally\s+speaking\b|\busually\b|\boften\b|
        \btends?\s+to\b|\bin\s+many\s+cases\b)

COND_JP = (?: 場合|なら|とき|時|ただし|但し|例外|条件|前提|依存|次第で|によって|により|限り)
COND_EN = (?i)(?:\bif\b|\bwhen\b|\bunless\b|\bprovided\s+that\b|\bassuming\b|\bin\s+case\b|\bdepends?\b|
        \bdepending\s+on\b|\bexcept\b|\bexception\b|\bunder\s+the\s+condition\b)

ALT_JP = (?: 一方で|他方|別 (?: の|に)|他にも|また|さらに|加えて|もう一つ|別の可能性|観点|選択肢)
ALT_EN = (?i)(?:\balternatively\b|\bon\s+the\s+other\s+hand\b|\banother\b|\ban\s+alternative\b|\balso\b|
```

```
        \badditionally\b|\bfurthermore\b|\bin\s+contrast\b|\bby\s+contrast\b)

ASSERT_JP = (?: 必ず|絶対|間違いない|確実|当然|に決まって (?: い (?: る|ます)|る)?|に違いない|100\%|断言)
ASSERT_EN = (?i)(?:\bdefinitely\b|\bcertainly\b|\bmust\b|\balways\b|\bundoubtedly\b|\bno\s+doubt\b|
        \b100\%\b)

PROPENSITY_JP = (?: 可能性|示唆|考えられ|想定され|起こりうる|つながる|ながる|導く|寄与|促す|
        見込まれ|〜?うる|〜?得る|検討でき)
PROPENSITY_EN = (?i)(?:\bsuggests?\b|\bmay\s+lead\s+to\b|\bcould\s+lead\s+to\b|\bmight\s+lead\s+to\b|
        \bcan\s+be\s+considered\b|\bis\s+consistent\s+with\b|\bindicates?\b|\bimplies\b|
        \bmay\s+imply\b|\bcan\s+help\b|\bmay\s+help\b)
```

Interpretation mapping. HEDGE primarily reflects epistemic uncertainty markers (linked to $z_{\text{epistemic}}$). COND and ALT are proxies for conditionality and alternative exploration (linked to z_{scope}). ASSERT is a coarse proxy for premature closure / definitiveness (linked to closure-proneness; often co-varying with $z_{\text{epistemic}}$).

C.3 Optional structural indicators (language-agnostic)

The following patterns can be used as low-cost proxies for formatting and breadth (useful for spot-checking or robustness):

LIST_ITEM (multiline): `(?m)^\s*(?:[-*• ]|[0-9][.)]|[①-⑳])+`
JP_SENT_DELIM (split): `[。！？!?]+`
EN_SENT_DELIM (split): `[.!?]+`
EN_WORD (for word count): `(?i)\b[0-9A-Z]+`

Recommended minimal reporting. For each topic, report A/B deltas of HEDGE, COND, ALT, and optionally ASSERT and LIST_ITEM count. Because we use paired A/B comparisons with all non-linguistic variables held constant, these features are used as within-language deltas rather than absolute cross-language magnitudes.

D Worked Example of Forward and Inverse Use

This appendix provides a minimal worked example illustrating how the Z-model can be used in both forward (prediction) and inverse (diagnostic) modes. The purpose of this example is purely illustrative: it demonstrates how prompt-induced behaviors can be interpreted and adjusted using the Z-model, without claiming generality, optimization, or performance improvement.

D.1 Scenario

We consider a short Japanese prompt fragment that is commonly used in practice and is known to elicit polite but relatively definitive responses. This example is intentionally minimal, focusing on a single fragment-level perturbation rather than a full prompt design.

D.2 Forward observation

Prompt (JP).

慎重に検討してください。

Observed output (excerpt).

> この方法は有効です。一般的に大きな問題はありません。

Although the prompt explicitly requests caution, the output exhibits categorical assertions with limited qualification. In particular, conditions, exceptions, or alternative cases are not made explicit.

D.3 Inverse diagnosis via the Z-model

Based on the observable output characteristics, we infer the following latent configuration using the Z-model. The polite phrasing and confident tone suggest elevated stylistic politeness ($z_{\text{style}} \uparrow$) together with increased epistemic commitment ($z_{\text{epistemic}} \uparrow$), while the absence of conditions or alternatives indicates suppressed conceptual breadth ($z_{\text{scope}} \downarrow$).

Importantly, these shifts are interpreted as coupled effects: the fragment primarily encodes an epistemic stance, but induces secondary effects on scope and closure through pragmatic coupling.

D.4 Prompt adjustment (inverse to forward)

Suppose the user's intent is to retain a cautious tone while reducing premature definitiveness and encouraging explicit conditionalization. Based on the inverse diagnosis above, the prompt can be minimally adjusted to target lower epistemic commitment and broader scope.

Adjusted prompt (JP).

> 慎重に検討してください。断定せず、条件や例外があれば明示してください。

D.5 Forward effect after adjustment

Observed output (excerpt).

> この方法は多くの場合に有効ですが、条件によっては例外も考えられます。

Relative to the original output, the adjusted response explicitly marks uncertainty, introduces conditional structure, and broadens the scope of consideration, while maintaining an overall cautious tone.

D.6 Discussion

This worked example illustrates how the Z-model supports both forward prediction (prompt fragment $\rightarrow$ behavioral tendencies) and inverse diagnosis (observed behavior $\rightarrow$ prompt adjustment). Crucially, this interaction cannot be captured by a static taxonomy of prompt attributes: the same fragment simultaneously affects multiple latent dimensions, and effective adjustment requires reasoning about their coupled shifts. The example is not intended as an evaluation or optimization procedure, but as a concrete demonstration of the operational use of the Z-model.

To further mitigate model/snapshot dependence, we also report an open-weight reproduction on a pinned checkpoint with matched prompts and proxy definitions (Appendix F.2).

E Reproducibility Checklist and API Settings

What must be reproduced. Because the paper evaluates a stochastic, protocol-level phenomenon, the main replication target is recovery of the reported *directional and distribution-level* effects under matched prompts and counting rules. Exact string equality is neither expected nor required. A faithful reproduction should recover the sign and qualitative ordering of the main proxy deltas, while tolerating moderate changes in magnitude across model snapshots, hardware stacks, or open-weight substitutes.

Embedded artifact set. Replication requires only a compact artifact set: (i) the exact prompt templates and probe fragments, (ii) the model identifier and, for API-served runs, the execution window, (iii) decoding parameters and sample counts, (iv) Unicode normalization and regex-based proxy definitions, and (v) the aggregation procedure used to compute deltas and bootstrap intervals. These elements are embedded in the paper and appendices rather than delegated to a large external software package: the protocol is specified in Sections B and 3.4, the indicator dictionary in Appendix C, and the model/API checklist in this section. No fine-tuning, learned classifier, or hidden annotation layer is required to reproduce the reported measurements.

API rerun path. The following snippet can be used to report the API-based setup concisely (values as in Section 3.4):

> *Model:* OpenAI GPT-5.2 (API-accessed, version 5.2).
> *Execution window:* 2026-02-27 UTC.
> *Sampling:* temperature = 0.7, top_p = 1.0, max_tokens = 512, $n = 50$ generations/(topic, condition) (five topics $\Rightarrow N = 250$ per condition).
> *Seed:* not set / not supported (stochastic sampling; evaluate via directional / distribution-level deltas).
> *Prompts:* system prompt fixed; user prompt fixed except for the probe fragment.
> *Post-processing & metrics:* Unicode normalization (NFKC), regex dictionary in Appendix C, bootstrap 95% CI.

Snapshot-stable rerun path. Readers who prefer a non-API reproduction route can use the open-weight setup in Appendix F, which pins `Qwen/Qwen2.5-7B-Instruct` to revision `a09a35458c702b33eeacc393d103063234e8bc28` while keeping the prompts, decoding settings, and observable proxy definitions matched to the API-based protocol.

The protocol in Section 3.4 and the dictionaries in Appendix C are intended to make replication possible with any contemporary LLM, even if the precise API model snapshot is no longer available.

F Reproduction on an Open-Weight Model

To assess whether the observed latent interference patterns depend on a specific API-served model, we reproduce (i) the cross-lingual politeness probing experiment (Exp. A) and (ii) the minimal directional sanity check (Exp. B) on a pinned open-weight, instruction-tuned checkpoint. Specifically, we use the model `Qwen/Qwen2.5-7B-Instruct`, revision `a09a35458c702b33eeacc393d103063234e8bc28`, obtained from the Hugging Face Hub. All prompts, decoding parameters, and proxy definitions are kept identical to the API-based setup; only the underlying model is changed.

Table F.1: Polite–plain deltas per 1,000 characters on a pinned open-weight, instruction-tuned checkpoint. Bootstrap 95% confidence intervals are shown in brackets. LIST_ITEMS is reported as a raw count difference. JP and EN correspond to Δ_{JP} and Δ_{EN}, respectively.

Metric	Δ_{JP}	Δ_{EN}
HEDGE_EPI	0.184 [-0.116, 0.486]	0.273 [0.177, 0.375]
HEDGE_GEN	-0.202 [-0.410, 0.006]	-0.103 [-0.187, -0.018]
COND	0.050 [-0.269, 0.372]	0.523 [0.421, 0.624]
ALT	-0.044 [-0.197, 0.104]	0.057 [-0.075, 0.195]
ASSERT	0.462 [0.340, 0.584]	0.071 [0.031, 0.110]
LIST_ITEMS	-1.486 [-2.624, -0.368]	-2.590 [-2.989, -2.196]

Table F.2: Open-weight reproduction of the minimal sanity check (Exp. B; cf. Table B.1). We use the same Japanese base prompt and probe fragments as in Appendix B; English glosses are provided in parentheses for readability. Symbols indicate the direction of change in the chosen observable proxy relative to the base prompt on the pinned open-weight model (as in Table B.1). **Importantly, "+" and "–" denote changes in the observable proxy, not in the latent z-dimension itself.**

Prompt fragment	Target z-dimension	Δ ($T = 0.2$)	Δ ($T = 0.8$)
箇条書きで (Bullets)	z_{format}	+	+
ステップごとに (Stepwise)	z_{process}	+	–
断定せずに (Hedging)	$z_{\text{epistemic}}$	+	+
慎重に (Cautious tone)	$z_{\text{epistemic}}$	+	+
A と B を比較して (Compare A/B)	z_{process}	+	+
代表例を挙げて (Examples)	z_{scope}	+	+

F.1 Cross-lingual Politeness Deltas (Exp. A)

Positive values indicate higher proxy rates under the polite condition than under the plain condition.

F.2 Directional Sanity Check (Exp. B)

Table F.2 reproduces Table B.1 on the pinned open-weight model (same Japanese base prompt and probe fragments; English glosses shown for readability). We observe agreement in direction for 11/12 (6 fragments × 2 temperatures) entries; the only mismatch occurs for the stepwise fragment at $T = 0.8$, where the chosen minimal process proxy (e.g., explicit list-item markers) decreases relative to the base prompt. This likely reflects higher-variance formatting under high-temperature sampling (e.g., steps expressed without explicit bullet/number markers) rather than a qualitative reversal of procedural structuring.

G Indicator-Set Ablation for Minimal Probing

A frequent reviewer concern for proxy-based probing is brittleness: do conclusions depend on a particular indicator choice? To address this without additional model calls, we perform a simple *indicator-set ablation* on the same generations used for Tables 4 and 5. We evaluate three nested indicator sets: (i) **Full** (10 indicators; all metrics reported in Table 4), (ii) **Reduced** (7 indicators; core lexical indicators plus length/structure), and (iii) **Minimal** (5 indicators; core lexical indicators plus length).

Table G.1: Indicator-set ablation for minimal probing. k is the number of indicators in the set. $\overline{|d|}$ is the mean absolute Cohen's d over included indicators. #CI$\neq$ 0 counts indicators whose bootstrap 95% CI for Δ excludes zero. #robust further requires a nontriviality filter, i.e., $|d| \geq \tau_d$ or JSD $\geq \tau_{\text{JSD}}$ with $(\tau_d, \tau_{\text{JSD}}) = (0.2, 0.05)$ (Algorithm 2). Reduced and minimal sets retain the core lexical indicators (HEDGE, ASSERT, COND, ALT) and show that the Japanese interference signature remains detectable under substantially fewer observables.

Indicator set	k	$\overline{\lvert d\rvert}_{\text{JP}}$	#CI$\neq$ 0/#robust (JP)	$\overline{\lvert d\rvert}_{\text{EN}}$	#CI$\neq$ 0/#robust (EN)
Full (10 indicators)	10	0.363	8/7	0.259	6/6
Reduced (7 indicators)	7	0.478	7/6	0.263	4/4
Minimal (5 indicators)	5	0.355	5/4	0.126	2/2

Table G.2: Macro-group coarsening robustness check. Indicators from Tables 4–5 are grouped by the four macro-groups of Figure 1. We report mean absolute Cohen's d within each macro-group and the number of indicators whose bootstrap 95% CI for Δ excludes zero (#CI$\neq$ 0), along with the number that additionally pass the nontriviality filter (#robust; $|d| \geq \tau_d$ or JSD $\geq \tau_{\text{JSD}}$, $(\tau_d, \tau_{\text{JSD}}) = (0.2, 0.05)$).

Macro-group	k	$\overline{\lvert d\rvert}_{\text{JP}}$	#CI$\neq$ 0/#robust (JP)	$\overline{\lvert d\rvert}_{\text{EN}}$	#CI$\neq$ 0/#robust (EN)
Expression (format/length)	3	0.830	3/3	0.514	3/3
Epistemic	5	0.137	3/2	0.201	3/3
Reasoning (scope)	2	0.231	2/2	0.018	0/0
Framing (role/audience/objective)	0		–		–

For each set and language, we summarize (a) the mean absolute standardized effect size $\overline{|d|}$ across included indicators, and (b) the number of included indicators whose bootstrap 95% CI for Δ excludes zero. Under the matched probing protocol, Japanese often exhibits a more multi-faceted proxy signature than English; we treat this as a protocol-level diagnostic observation and note possible confounds from cross-lingual training/alignment differences.

Macro-group coarsening (robustness under 4-group coarsening). Beyond ablations over observable indicators, we also test robustness under coarsening of the Z-space itself by collapsing observables into the four macro-groups of Figure 1. We assign indicators from Tables 4–5 to macro-groups as follows: **Expression** (structural/format proxies: `chars`, `list_items`, `sentences`), **Reasoning** (scope proxies: COND, ALT), and **Epistemic** (uncertainty/closure proxies: HEDGE_*, ASSERT/`definite`, PROPENSITY). The **Framing** macro-group (role/audience/objective) is not operationalized in Exp. A and is therefore not evaluated here. We include the intended **Expression** shift for completeness; interference concerns the off-target macro-groups.

H Additional Figures: P1 Orthogonal Design, 11D Coarsening, Basis Rotation, and 2×2 Factorial Probe

H.1 P1 orthogonal 4D design: stance targets (target-wise view)

Figure 2 reports mean R^2 aggregated over (i) all proxy targets and (ii) stance-only targets. Figure H.1 breaks the stance-only aggregate down by target to show which stance proxies benefit from additional dimensions.

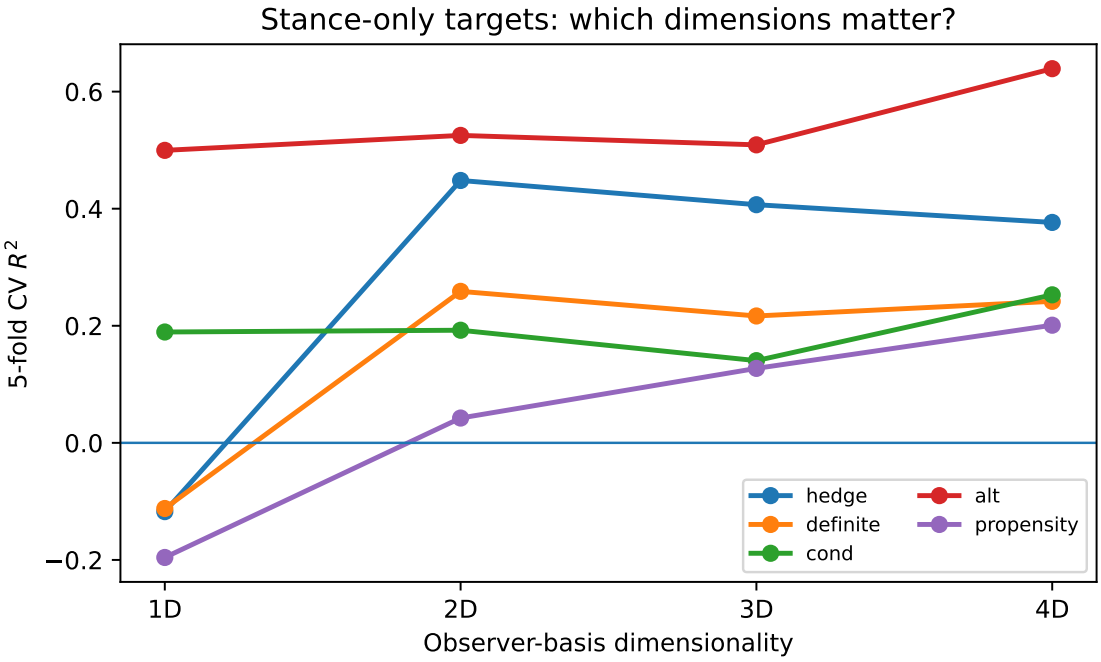

Figure H.1: **P1 orthogonal 4D design: target-wise stance predictability under basis reduction.** 5-fold cross-validated R^2 for each stance proxy target under reduced observer bases (1D–4D). This target-wise view complements Figure 2 and shows that the $z_{\text{epistemic}}$–z_{scope} subspace explains a substantial portion of stance variation, while additional dimensions provide more modest gains that differ by target.

H.2 11D coarsening check on the cross-lingual politeness probe (task-sliced results)

Beyond the controlled 4D orthogonal design above, we also test basis coarsening for the full 11D reference basis on the cross-lingual politeness probe (Exp. A). Figure H.2 provides a task-sliced view (task0/task1) under the same proxy-delta prediction protocol.

H.3 Basis rotation robustness

To probe robustness to alternative parameterizations beyond coarse merging/splitting, we apply random orthogonal rotations within each macro-group subspace of the observer basis and re-evaluate the same proxy-delta prediction protocol. Figure H.3 shows that performance is stable under such rotations, consistent with interference detection being insensitive to axis labeling within the same representational subspace.

H.4 Factorial 2×2 probe: cell means for representative proxies

Figure 4 summarizes interactions as standardized effect sizes. Figure H.4 visualizes raw cell means for representative proxies in each language, together with the additive prediction for $C11$ (dashed line), making non-additivity directly visible.

Appendix H.4 bundle note. In the reviewer bundle, the canonical factorial artifacts for Appendix H.4 are shipped as precomputed ALL-slice outputs under `results/factorial/derived/en_T0.2_all/` and `results/factorial/derived/jp_T0.2_all/`.

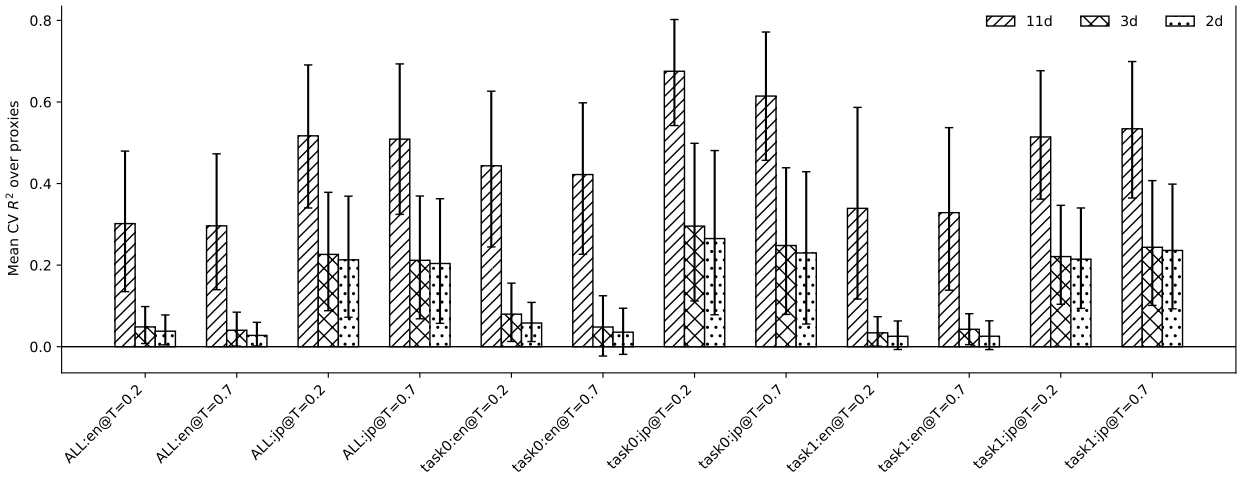

Figure H.2: **11D coarsening check by task slice (monochrome-friendly).** Mean cross-validated R^2 over proxy metrics when predicting proxy-delta signatures from the observer basis coordinates, comparing the full 11D reference basis to reduced 3D/2D coarsenings, shown for the task-averaged slice and for each task separately. This plot supports the claim that additional axes in the 11D reference basis contribute to descriptive sufficiency under the present cross-lingual probe/proxy family.

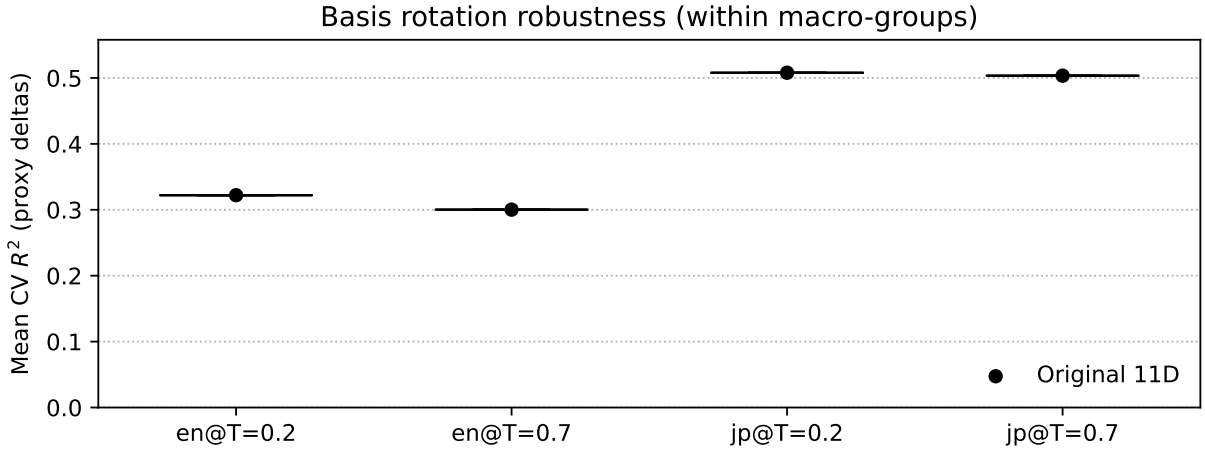

Figure H.3: **Basis rotation robustness (within macro-groups; monochrome-friendly).** Mean cross-validated R^2 for predicting proxy-delta signatures from the observer basis coordinates under random orthogonal rotations within each macro-group subspace (200 rotations; bands show the 2.5/97.5 percentiles across rotations). The stability indicates that interference detection is not an artifact of a specific axis labeling.

These directories contain both the underlying summaries (`figY_summary.csv`) and the rendered cell-means figures (`FigY_cells__*.png/.pdf`), derived from the canonical `T=0.2` raw JSONL files via `scripts/make_figY_factorial_plots.py` with `-task_id -1`. Regeneration is therefore optional for reviewer inspection.

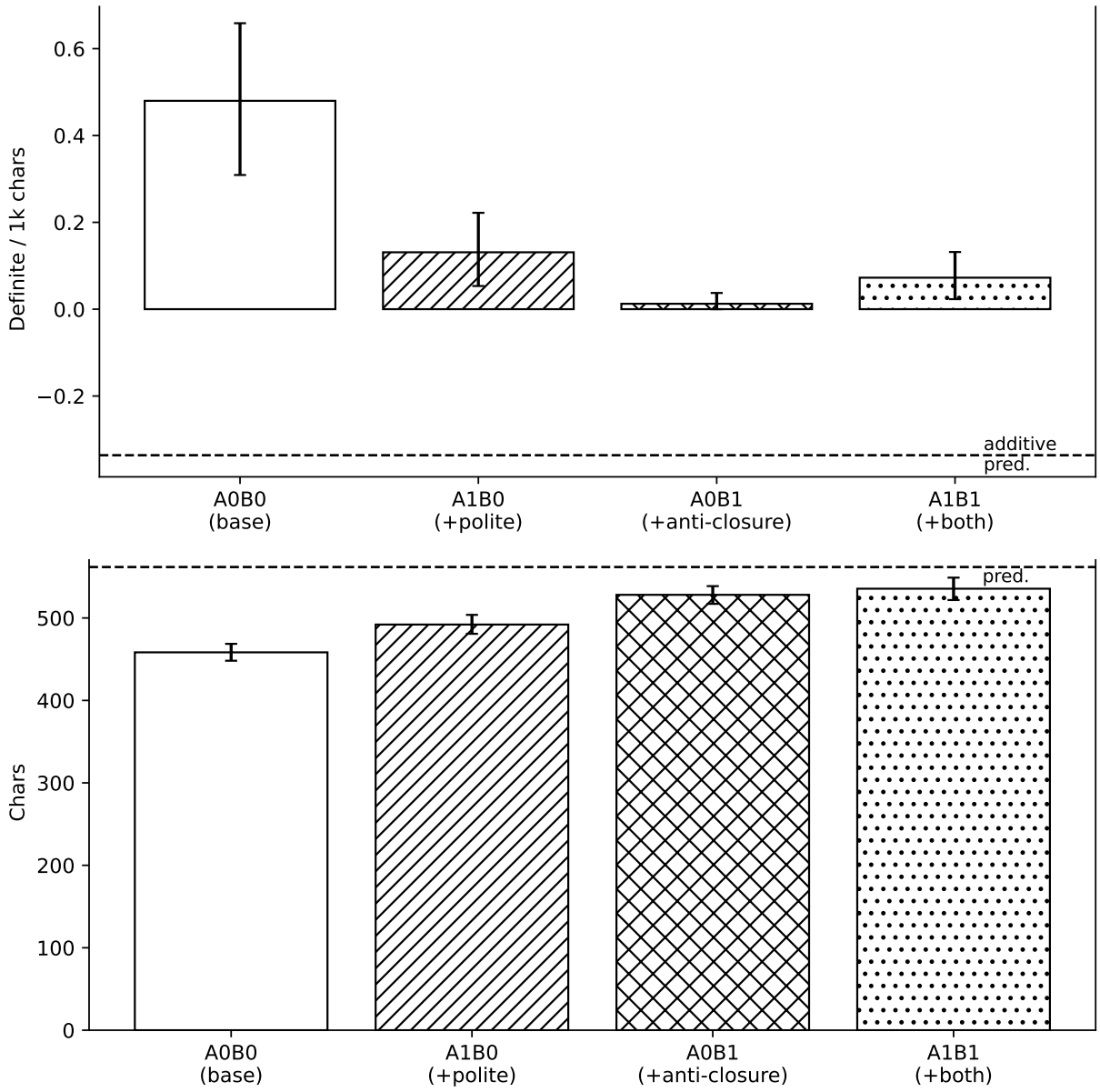

Figure H.4: **Factorial 2×2 probe: raw cell means with additive prediction (monochrome-friendly).** Means (with 95% bootstrap confidence intervals) for representative proxies under the four factorial conditions $C00 = p_0$, $C10 = p_0 \oplus A$, $C01 = p_0 \oplus B$, $C11 = p_0 \oplus A \oplus B$. The dashed line denotes the additive prediction for $C11$ under no interaction ($\mu_{10} + \mu_{01} - \mu_{00}$). Deviations from the dashed line correspond to the interaction term summarized in Figure 4.

I Optional probabilistic view: mixture over observer-side configurations

Let x denote a prompt and y a generated output. One can view the observer-side configuration as a latent variable z with a prompt-dependent distribution $p(z \mid x)$, and write the output distribution as a mixture:

$$p(y \mid x) \;=\; \int p(y \mid z)\, p(z \mid x)\, dz. \tag{I.1}$$

This interpretation is intended purely as an alternative *reading* of the observer-side abstraction. It does not introduce an estimator for z and is not used in the empirical procedures, which remain grounded in observable proxy measurements $r(y)$ under controlled A/B perturbations.