V1.0.0

Structured Configuration-Style Prompts as Variance Control: An Analytical Framework for Cross-LLM Output Variance Reduction

Rowan Brad Quni-Gudzinas

QNFO-ULA

ABSTRACT

Large language models (LLMs) are increasingly used as interchangeable components in automated pipelines, yet the same natural-language instruction produces materially different outputs across models, sampling seeds, and runs. This variance blocks the use of LLMs where reproducible execution matters: batch document processing, schema-constrained data extraction, and multi-step agentic workflows. The specific gap is that prompt engineering is practiced as a literary craft with no predictive model of how prompt form controls output variance, and the largest existing variance decomposition of LLM creative output attributes 36.43% of originality variance to prompts but does not separate prompt form from prompt content. We develop an analytical framework that treats a prompt as a partial program whose unspecified parameters are filled in by each model's interpretation, models interpretation divergence as an input to the output-variance machinery of sensitivity analysis, and derives a variance-reduction factor equal to the share of interpretation variance carried by the parameters a structured prompt pins down. Under explicitly labeled illustrative assumptions, pinning four of six interpretation dimensions removes 82% of the prompt-attributable interpretation variance, bounded by the empirical decomposition to at most 29.9 percentage points of total variance; a 20-field schema at an assumed per-field compliance of 0.986 yields 75.4% full-compliance probability against 12.2% for an assumed prose baseline of 0.90 per field, a 3.57-fold reduction in violation probability. The framework converts prompt engineering into a specification discipline with measurable design targets and a falsifiable pilot protocol of 53 paired tasks.

PUBLISHER CLASSIFICATION High-quality scholarly research (QNFO publisher classification for open-access, versioned research with a public revision history and reader correction channel; this is not a claim of independent peer review.)

PEER-REVIEW STATUS Not independently peer-reviewed

CITE AS Quni-Gudzinas, R. B. (n.d.). Structured Configuration-Style Prompts as Variance Control: An Analytical Framework for Cross-LLM Output Variance Reduction. QNFO Preprint v1.0.0.

1. Introduction

Teams that wire LLMs into production pipelines face a reproducibility problem that has no accepted engineering answer. When the same extraction, transformation, or drafting instruction is given to two different models — or to the same model twice — the outputs differ in structure, in content, and sometimes in whether the task is performed at all. Downstream code that parses one model's output fails on another's; a workflow validated against one model breaks silently when a provider is swapped or a model is updated. The cost is concrete: every integration must either build model-specific parsing, accept silent failures, or pay for human review of every output. What is missing is a design lever that makes LLM behavior predictable across models, in the way a written specification makes behavior predictable across implementations of an API.

The dominant practice for controlling LLM behavior is prompt engineering: careful wording of natural-language instructions. This practice implicitly assumes that a prompt acts as a near-deterministic program — find the right phrasing and the model will execute it. Large-scale evidence now quantifies how far that assumption is from reality. A variance decomposition across 12 LLMs, 10 creativity prompts, and 100 samples per cell (N = 12,000) found that prompts ex-

plain 36.43% of variance in originality while model choice explains 40.94–51.25% and within-LLM sampling variance 10.56–33.70% [8]. The authors of that study conclude that prompting is "less about programming a machine and more about steering a probabilistic distribution" [8]. If prompts are weak, bounded levers even at their best, the practical question becomes: which property of a prompt makes its lever stronger, and can that property be engineered deliberately rather than discovered by trial?

This paper develops the conjecture that prompt *form* — not content — is that property. Prose-style instructions leave the interpretation of the task to the model: what counts as a section, what order to process items in, what to do on an edge case. Different models resolve these ambiguities differently, and the same model resolves them differently across seeds. Configuration-style prompts — explicit schemas, IF/THEN protocols, state machines, enumerated field lists — remove those degrees of freedom by construction. The claim is testable, and this paper does two things about it: it builds the formal machinery needed to state the claim precisely and predict its magnitude, and it specifies the experiment that would confirm or refute it. The machinery borrows from sensitivity analysis, where the influence of uncertain inputs on output variance is measured by Sobol indices and estimated efficiently from structured approximations [1], [5], and from work showing that the distribution of inputs, not just the map, controls output variance [2].

The contributions are:

- A formal model of a prompt as a partial program with an interpretation map per model, yielding a three-way variance decomposition (prompt-interpretation, model, sampling) aligned with the empirical decomposition of [8]; derived in Section 3 and Section 4.
- A variance-reduction factor for structured prompts: the reduction equals the fraction of interpretation variance carried by the free parameters that the structured form eliminates, expressed through total sensitivity indices in the manner of [1], [5]; derived in Section 4.
- Worked numerical predictions under explicitly labeled illustrative assumptions: an 82% reduction of the prompt-attributable interpretation variance when four of six interpretation dimensions are pinned, a bound of 29.9 percentage points on total-variance reduction, a schema-compliance calculation giving 75.4% versus 12.2% full-compliance probability at 20 fields, and a Bayes-factor amplification for cross-model verification protocols from 8 to 16 under a halved null-convergence probability; derived in Section 4 and reported in Section 5.
- A falsifiable pilot protocol — paired prose/structured prompt corpora, fixed temperature, multiple models, structural and semantic agreement metrics, sized at 53 paired tasks — with the specific outcome that would refute the central claim; specified in Section 7.

The framework also bears on adjacent problems: any method that raises inter-model agreement raises the evidential weight of multi-model verification panels, and the same variance-control logic applies wherever a stochastic generator must satisfy a conjunction of many independent constraints, including long-form drafting reliability. The result could matter wherever multi-model LLM panels, schema-constrained generation, or reproducible agentic execution are used.

2. Background and Related Work

The closest empirical anchor for this paper is the variance decomposition of LLM creative output by [8]. That study evaluated 12 LLMs on 10 creativity prompts with 100 samples each, for N = 12,000 outputs, partitioning variance into model, prompt, and within-LLM sources. Its headline numbers are that prompts explain 36.43% of variance in originality, comparable to model choice at 40.94%, while for fluency model choice dominates at 51.25% and within-LLM variance reaches 33.70% [8]. Its discussion states the premise this paper attacks directly: the field has operated on the premise that prompts act as close-to-deterministic programs, and the decomposition challenges that view [8]. What [8]

does not do is separate the *form* of a prompt from its *content* — all its prompts are prose variations — so its 36.43% conflates task framing with interpretive ambiguity. Our framework supplies that separation and predicts that the interpretation component, which structured forms eliminate, is the controllable part.

The machinery for predicting how input uncertainty becomes output variance comes from sensitivity analysis. [1] develops sparse polynomial chaos expansion (SN-PCE) for processes represented as directed acyclic graphs, constructs orthonormal polynomials and Sobol indices from the Hoeffding decomposition, and shows both theoretically and on two manufacturing processes and a flooding process that the sparse approximation identifies influential inputs with substantially fewer observations than Monte Carlo or naive PCE [1]. We reuse its central object — the Sobol index, the fraction of output variance attributable to an input — as the measure of how much output variance each unpinned interpretation parameter carries. [5] applies the same index family to feature selection: its total sensitivity index incorporates interactions between features, which most filter methods ignore, and on the Friedman dataset with known relevant features the total sensitivity index was highest for the relevant features, including under high noise when computed with neural-network models [5]. This matters for our framework because interpretation parameters interact: a model's choice of output format and its choice of section ordering are not independent levers, and a decomposition that ignores interaction would mis-rank which parameters a structured prompt must pin.

A second strand shows that the *distribution* of inputs, independent of the map, controls output variance. [2] studies parameterized quantum circuits and proves that for angle-encoded product states the input distribution's overlap with the circuit's dynamical Lie algebra — its algebraic purity — determines output variance: for the matchgate family every pure product input retains enough overlap to give output variance of order $1/n$, while for the off-diagonal family computational-basis inputs give exactly zero output for every parameter choice [2]. The structural lesson transfers: changing the distribution over interpretations (which a structured prompt does by shrinking the support) changes output variance even when the model is fixed. [4] provides the complementary network-side result, deriving the exact marginal output variance for $\mathcal{L}$-Lipschitz layers initialized with IID Gaussian weights via the Wishart distribution, and finding empirically that a weight scaling of $1/\sqrt{n}$ is less effective at controlling per-layer output variance than $10/\sqrt{n}$ [4]. Both works treat variance as a designed quantity — set by the input distribution or the parameterization — which is exactly the stance this paper takes toward prompt form. [9] reaches a similar design conclusion inside LLMs themselves: its LayerNorm Scaling rescales output variance inversely with layer depth so that deeper layers contribute meaningfully, showing that variance placement is an architectural lever with measurable performance consequences [9].

On the measurement side, [3] investigates the sequential sampling method for reducing the variance of a simulation output performance measure in crossdocking systems, targeting a desired 95% confidence interval half-width with a minimum number of replications [3]. The methodological point — that output variance determines how many runs any conclusion needs — governs the pilot protocol we specify in Section 7: the more structured prompts compress variance, the fewer paired runs are needed to establish the effect. [6] analyzes optimal experimental design for Bayesian regression and shows that popular entropy-based criteria have weak, sometimes uncontrollable dependence on the output values of existing samples, so directions of parameter space that contribute most to the output may not be emphasized; it proposes an output-weighted criterion that accelerates convergence [6]. We use its caution in Section 7: the pilot's prompt pairs should be selected to emphasize the interpretation dimensions that carry variance, not sampled uniformly from prompt space. [10] demonstrates the end-to-end pattern with Earth system models: machine-learning emulators make it practical to run global sensitivity analysis and connect forward uncertainty propagation to inverse parameter calibration, with parameter sensitivities depending on the output quantity, averaging window, and regime [10]. This is the template for our proposal to fit an emulator over (prompt form, model, seed) configurations and read off which prompt-form features carry variance.

Collectively, these works supply the three ingredients our framework needs: an empirical anchor that prompt variance is first-order [8], a variance-attribution formalism that handles interactions [1], [5], and a design tradition in which variance is a controlled output of input and architectural choices [2], [4], [9], with measurement protocols sized by variance [3], [6], [10].

3. Methods

3.1 The model

We model an LLM m as a conditional distribution $p_{\theta_m}(y \mid x)$ over outputs y given a prompt x, following the probabilistic framing of [8]. A prompt is not a function from intent to output; it is an *underspecified program*: a partial specification P with a set $F(P) = \{f_1, \ldots, f_k\}$ of free interpretation parameters — choices the prompt does not determine, such as output ordering, section granularity, edge-case handling, or format details. Each model m resolves these parameters through an interpretation map

$$g_m : P \mapsto \boldsymbol{\theta}_m = (\theta_{m,1}, \ldots, \theta_{m,k}) \in \Theta,$$

and the observed output is a draw $y \sim p_{\theta_m}(y \mid P)$. Sampling stochasticity is the conditional variance $\mathrm{Var}(y \mid \boldsymbol{\theta}_m)$; interpretation divergence is the spread of $\boldsymbol{\theta}_m$ across models and seeds.

3.2 Variance decomposition

Following the three-source partition of [8], we decompose the total output variance over the joint ensemble of models, prompts, and samples as

$$\mathrm{Var}[y] = \underbrace{\mathrm{Var}_m[\mathbb{E}(y \mid \boldsymbol{\theta}_m)]}_{\text{interpretation}} + \underbrace{\mathbb{E}_m[\mathrm{Var}(y \mid \boldsymbol{\theta}_m)]}_{\text{sampling}} + \mathrm{Var}_{\text{prompt content}},$$

where the first two terms are the law-of-total-variance split of the model-attributable variance and the third is the variance due to differing prompt content across the corpus. The empirical decomposition of [8] measures the sum of these components without separating interpretation from content; our framework predicts that structured forms act only on the first term.

3.3 Structured prompts as support restriction

A configuration-style prompt P_{struct} is one whose free-parameter set is smaller: $F(P_{\text{struct}}) \subset F(P_{\text{prose}})$, with the residual parameters either pinned to fixed values or constrained to a discrete schema (an enumerated field list, an IF/THEN protocol, a state machine). Formally, the structured prompt restricts the support of the interpretation distribution: where the prose prompt induces a distribution $g_m(P) \sim \pi$ over a product space $\Theta_1 \times \cdots \times \Theta_k$, the structured prompt induces a distribution concentrated on a submanifold where $k - k'$ coordinates are fixed. This is the same mechanism by which input-distribution restriction controls output variance in [2]: shrinking the input support toward states of high overlap with the intended behavior shrinks the output variance, without changing the map itself.

3.4 Sensitivity weighting of free parameters

Not all free parameters contribute equally to output variance. Following [1] and [5], we assign each free parameter f_i a total sensitivity index S_{T_i}, the fraction of output variance attributable to f_i including its interactions, with $\sum_i S_{T_i} \leq 1$ (equality when interactions are fully attributed). The prompt-attributable interpretation variance is then

$$V_{\text{int}} = \mathrm{Var}[y] \cdot \sum_{i=1}^{k} S_{T_i},$$

and a structured prompt that pins the parameters in a set $\bar{F}$ leaves

$$V'_{\text{int}} = \text{Var}[y] \cdot \sum_{f_i \in F \setminus \bar{F}} S_{T_i}.$$

The variance-reduction factor on the interpretation component is $R = V'_{\text{int}}/V_{\text{int}}$. This is the paper's central object: it converts "structured prompts reduce variance" into a computable quantity, estimable by the sparse-PCE machinery of [1] once pilot data exist.

3.5 Compliance as a conjunction of constraints

For schema-constrained outputs, success is a conjunction: all n fields must be present and well-formed. If per-field compliance is q, independent across fields, the full-compliance probability is q^n. Structured prompts raise q (the schema states each field explicitly) relative to a prose prompt that leaves field presence to interpretation.

3.6 Modeling convention and scope

Two modeling conventions were considered for quantifying how prompt form constrains interpretation: a token-density model, in which residual variance falls linearly with the fraction of syntactically constrained prompt tokens, and the sensitivity-weighting model of Section 3.4. We take the sensitivity-weighting model under the convention that prompt form acts by pinning specific high-sensitivity interpretation parameters rather than by generic token compression; Section 4.7 shows that the density model is quantitatively incoherent under the empirical anchor of [8], which motivates this choice.

This paper performs no experiment, measurement, or simulation on any model. Every numerical value in Section 4 is either (i) taken from a cited work and attributed to it, or (ii) computed here from explicitly labeled illustrative assumptions. The pilot study that would supply real estimates of the S_{T_i} is specified as a proposal in Section 7, not executed here.

4. Analysis

4.1 Consistency check of the empirical decomposition

The decomposition of [8] reports, for originality, prompt variance of 36.43%, model choice of 40.94%, and within-LLM variance in the range 10.56–33.70%. Summing the lower ends:

$$36.43\% + 40.94\% + 10.56\% = 87.93\%,$$

and the upper ends give $36.43\% + 51.25\% + 33.70\% = 121.38\%$. The components are therefore not strictly additive over the reported ranges — the model and within-LLM figures are ranges across metrics or conditions, and shares overlap or are normalized differently across outcome measures. We take from this only the order-of-magnitude facts that the paper states: prompt form (as prose) is a first-order source of variance, comparable to model identity, and within-LLM sampling variance is never negligible (at least 10.56%). Any variance-reduction mechanism that acts only on the prompt component can, at the extreme of the reported decomposition, address at most

$$\frac{36.43}{87.93} = 0.41435 \approx 41.4\%$$

of the total lower-bound variance — a bound we carry into Section 5.

4.2 Variance-reduction factor under illustrative sensitivity weights

We assume, as an illustrative configuration to be replaced by pilot estimates, that a prose task prompt leaves $k = 6$ interpretation dimensions with total sensitivity indices

$$S_{T_1} = 0.30, \quad S_{T_2} = 0.25, \quad S_{T_3} = 0.15, \quad S_{T_4} = 0.12, \quad S_{T_5} = 0.10, \quad S_{T_6} = 0.08,$$

which sum to $0.30 + 0.25 + 0.15 + 0.12 + 0.10 + 0.08 = 1.00$, i.e., the interpretation component accounts for the full prompt-attributable variance in this toy configuration. A structured prompt in the style of a configuration file pins the four largest contributors — output schema (f_1), processing order (f_2), section granularity (f_3), and edge-case protocol (f_4) — leaving f_5, f_6 free. The residual interpretation variance share is

$$R = \frac{S_{T_5} + S_{T_6}}{\sum_i S_{T_i}} = \frac{0.10 + 0.08}{1.00} = 0.18,$$

so the structured form removes $1 - R = 0.82$, i.e., 82%, of the prompt-attributable interpretation variance. Because the pinned set was chosen by descending S_{T_i} — the ranking logic of the total sensitivity index in [5], which correctly identified relevant features including under noise — this is the optimal pinning of four parameters under these weights. If instead the structured prompt pinned the four *smallest* contributors, the residual would be $(0.30 + 0.25)/1.00 = 0.55$, only a 45% reduction: the framework predicts that *which* degrees of freedom a structured form eliminates matters more than how many, a directly testable corollary.

4.3 Bound from the empirical decomposition

Combining Section 4.1 with Section 4.2: the interpretation component is at most the prompt share of total variance. At the lower-bound decomposition of [8], the prompt component is 36.43% of the 87.93% total, so the maximum total-variance reduction achievable by perfect elimination of interpretation divergence is

$$0.3643 \times 0.82 = 0.298726 \approx 0.299,$$

i.e., roughly 29.9 percentage points of total variance under the illustrative weights, with the remainder of total variance (model identity and sampling stochasticity) untouched by prompt form. This bound is the honest statement of the conjecture's reach: structured prompts are a lever on one component, not a cure for LLM stochasticity.

4.4 Schema-compliance arithmetic

Take the per-field compliance probability of a well-engineered structured pipeline to be $q_{\text{struct}} = 0.986$ and that of a prose baseline to be $q_{\text{prose}} = 0.90$; both values are illustrative assumptions standing in for data no cited source gives, with the structured value chosen at a per-constraint reliability level that a production pipeline plausibly requires. For a 20-field extraction schema (we assume $n = 20$ as a representative structured-output size), the full-compliance probabilities are:

$$P_{\text{prose}} = 0.90^{20} = e^{20 \ln 0.90} = e^{20 \times (-0.1053605)} = e^{-2.107210} = 0.121577 \approx 12.2\%,$$

$$P_{\text{struct}} = 0.986^{20} = e^{20 \ln 0.986} = e^{20 \times (-0.0140982)} = e^{-0.281964} = 0.754290 \approx 75.4\%.$$

The violation probabilities are $1 - 0.121577 = 0.878423$ and $1 - 0.754290 = 0.245710$, a ratio

$$\frac{0.878423}{0.245710} = 3.5743 \approx 3.57,$$

so under these assumptions the structured form reduces the per-document violation probability by a factor of 3.57. The independence assumption across fields is optimistic — field errors are plausibly correlated within a document — and Section 6 treats it as a limitation; the direction of the effect survives moderate correlation because both sides are affected.

4.5 Sensitivity of the compliance result to correlated field errors

The independence assumption of Section 4.4 is the strongest idealization in the compliance calculation. A standard way to absorb positive correlation among per-field failures is to deflate the number of effectively independent constraints. If the $n = 20$ field errors have equicorrelation ρ, the effective number of independent fields is

$$n_{\text{eff}} = \frac{n}{1 + (n-1)\rho}.$$

We assume $\rho = 0.05$ as a moderate-correlation illustration. Then

$$n_{\text{eff}} = \frac{20}{1 + 19 \times 0.05} = \frac{20}{1.95} = 10.2564,$$

and the full-compliance probabilities become

$$P^{(\rho)}_{\text{prose}} = 0.90^{10.2564} = e^{10.2564 \times (-0.1053605)} = e^{-1.080637} = 0.33935 \approx 33.9\%,$$

$$P^{(\rho)}_{\text{struct}} = 0.986^{10.2564} = e^{10.2564 \times (-0.0140982)} = e^{-0.144594} = 0.86538 \approx 86.5\%.$$

The violation probabilities are $1 - 0.33935 = 0.66065$ and $1 - 0.86538 = 0.13462$, a ratio

$$\frac{0.66065}{0.13462} = 4.907 \approx 4.91.$$

Under moderate positive correlation both arms improve, but the structured arm improves more in relative terms: the violation ratio rises from 3.57 (independent fields) to 4.91. At a stronger $\rho = 0.20$, $n_{\text{eff}} = 20/(1 + 19 \times 0.20) = 20/4.8 = 4.1667$, giving $P^{(\rho)}_{\text{prose}} = e^{4.1667 \times (-0.1053605)} = e^{-0.439002} = 0.64472$ and $P^{(\rho)}_{\text{struct}} = e^{4.1667 \times (-0.0140982)} = e^{-0.058743} = 0.94295$, a violation ratio of $0.35528/0.05705 = 6.23$. The direction of the Section 4.4 result therefore does not merely survive correlation; within this equicorrelation model it strengthens, because correlation compresses the number of independent chances to fail and the prose baseline has far more headroom to fail. The assumed value $q_{\text{struct}} = 0.986$ also coincides with the per-constraint reliability threshold $q^* \approx 0.986$ derived in [12] for multi-constraint drafting pipelines, which is the grounding for using that number as the structured-arm illustration rather than an arbitrary choice.

4.6 Amplification of cross-model verification evidence

The framework bears on multi-model verification panels of the kind formalized in [14], where agreement between independently trained LLMs on a multi-step argument is treated as Bayesian evidence for a claim's correctness. In that framework the synthesis-arm Bayes factor is the ratio of the convergence probability under the true hypothesis to the convergence probability under the null; its illustrative configuration gives a two-model Bayes factor $K_S = 8$, a two-model posterior of $8/17 \approx 0.4706$ from a 0.1 prior, and a requirement of $N = 4$ models to reach a posterior of at least 0.95 [14]. The framework also reports the sensitivity that governs the whole method: with the null's synthesis-convergence probability p_b^S varied from 0.20 to 0.02 at fixed $p_c^S = 0.4$, the two-model posterior spans approximately 0.182 to 0.690 [14]. The null's convergence probability is therefore the single quantity the panel's evidential weight hinges on.

Structured prompts act exactly on that quantity. If structured forms halve the probability that a null-hypothesis model converges on the same multi-step argument structure — because the argument structure is pinned by the prompt rather than reconstructed by each model — then, holding the true-hypothesis convergence probability fixed, the Bayes factor doubles:

$$K'_S = 2 \times K_S = 2 \times 8 = 16.$$

The two-model posterior from a 0.1 prior becomes

$$\text{prior odds} = \frac{0.1}{0.9} = \frac{1}{9}, \qquad \text{posterior odds} = \frac{1}{9} \times 16 = \frac{16}{9},$$

$$P'_2 = \frac{16/9}{1 + 16/9} = \frac{16}{25} = 0.64,$$

against $8/17 \approx 0.4706$ in the baseline configuration of [14]. The panel-size requirement shifts as well. A posterior of at least 0.95 requires posterior odds of at least $0.95/0.05 = 19$. With $K_N = (K'_S)^{N-1}$ under the cascade assumption:

- $N = 2$: odds $= 16/9 = 1.778 < 19$;
- $N = 3$: odds $= 16^2/9 = 256/9 = 28.444 \geq 19$, posterior $= 28.444/29.444 = 0.9661 \approx 0.966$.

So under the halved null-convergence assumption a three-model panel suffices where the baseline configuration of [14] requires four. For reference, the baseline cascade reproduces the reported values: at $N = 3$, odds $= 64/9 = 7.111$, posterior $= 7.111/8.111 = 0.8767$, matching the $P_3 \approx 0.8767$ of [14]; at $N = 4$, odds $= 512/9 = 56.889$, posterior $= 56.889/57.889 = 0.9827$, matching $P_4 \approx 0.9827$ [14]. The amplification claim is conditional on the halving assumption, which is precisely what the pilot protocol of Section 7 would measure; [14] itself flags that correlated errors from shared training corpora would deflate the effective per-model evidence, and structured prompts do not remove that shared-bias channel.

4.7 Size of the falsifying pilot

The pilot protocol of Section 7 is sized by the variance of the quantity it must estimate, following the principle that output variance determines how many runs any conclusion needs [3]. The binding estimate is the structured arm's full-compliance rate, expected near $P_{\text{struct}} = 0.754$ under the Section 4.4 assumptions. For a binomial proportion estimated with n paired tasks, the 95% confidence-interval half-width is approximately

$$h = \sqrt{\frac{z^2_{0.975}\, p(1-p)}{n}}, \qquad z_{0.975} = 1.96.$$

Using the conservative worst case $p = 0.5$ (which maximizes $p(1-p)$ and therefore the required sample size), the protocol size of $n = 53$ paired tasks gives

$$h = \sqrt{\frac{1.96^2 \times 0.25}{53}} = \sqrt{\frac{3.8416 \times 0.25}{53}} = \sqrt{\frac{0.9604}{53}} = \sqrt{0.018121} = 0.13461 \approx 0.135.$$

The expected effect against the prose baseline is $P_{\text{struct}} - P_{\text{prose}} = 0.754 - 0.122 = 0.632$, so the confidence-interval half-width is about $0.13461/0.632 = 0.213$, i.e., roughly 21% of the expected effect. Fifty-three paired tasks therefore resolve the predicted compliance gap with margin, while remaining small enough to run across several models and seeds; the same corpus supports the inter-model agreement and seed-consistency measurements the protocol specifies.

4.8 Incoherence of the token-density model

Section 3.6 set aside a rival convention in which residual output variance falls linearly with the fraction of syntactically constrained prompt tokens. The empirical anchor of [8] refutes this model quantitatively. Let c be the constrained-token fraction; the density model predicts a total-variance reduction of $c \times 100\%$. A configuration-style prompt plausibly constrains the majority of its tokens; even a conservative $c = 0.50$ predicts a 50% total-variance reduction. But Section 4.1 shows the prompt component is at most

$$\frac{36.43}{87.93} = 0.41435 \approx 41.4\%$$

of the lower-bound total variance, and prompt form can act only on the interpretation sub-component of that share. Any predicted reduction above 41.4 percentage points of total variance is impossible under the decomposition of [8], so the density model is falsified for every $c > 0.414$. The sensitivity-weighting model of Section 3.4 avoids this because its reduction is computed on the interpretation component only and then bounded by the prompt share, which is exactly the structure Section 4.3 used. This is why the paper adopts the sensitivity-weighting convention.

5. Results

All numbers below are computed in Section 4 from the explicitly labeled illustrative assumptions or taken from the cited works and attributed to them; no experiment, measurement, or simulation on any model was performed for this paper.

1. **Empirical anchor (attributed to [8]).** Prompts explain 36.43% of variance in originality across 12 LLMs, 10 prompts, and 100 samples per cell ($N = 12{,}000$), comparable to model choice at 40.94%; for fluency, model choice reaches 51.25% and within-LLM variance 33.70% [8]. The reported component shares are not additive over their ranges (lower-bound sum 87.93%, upper-bound sum 121.38%; Section 4.1).
2. **Maximum addressable share (Section 4.1).** The prompt-attributable component is at most $36.43/87.93 = 41.4\%$ of the lower-bound total variance.
3. **Variance-reduction factor (Section 4.2, illustrative weights).** Pinning the four highest-sensitivity of six interpretation dimensions leaves $R = 0.18$, an 82% reduction of the prompt-attributable interpretation variance; pinning the four lowest leaves $R = 0.55$, a 45% reduction. Which degrees of freedom are pinned matters more than how many.
4. **Total-variance bound (Section 4.3).** $0.3643 \times 0.82 = 0.298726 \approx 29.9$ percentage points of total variance, under the illustrative weights and the lower-bound decomposition.
5. **Schema compliance (Section 4.4, illustrative q values).** At $n = 20$ fields: $P_{\text{prose}} = 0.90^{20} = 12.2\%$, $P_{\text{struct}} = 0.986^{20} = 75.4\%$; violation ratio 3.57.
6. **Correlation robustness (Section 4.5).** Under equicorrelation $\rho = 0.05$: $P_{\text{prose}} = 33.9\%$, $P_{\text{struct}} = 86.5\%$, violation ratio 4.91; at $\rho = 0.20$ the ratio is 6.23. The structured advantage strengthens with moderate positive correlation in this model.
7. **Verification amplification (Section 4.6, conditional on the halving assumption).** Doubling the synthesis Bayes factor from $K_S = 8$ to 16 raises the two-model posterior from 0.4706 to $16/25 = 0.64$ and lowers the panel size needed for a posterior of at least 0.95 from $N = 4$ to $N = 3$ (posterior 0.966).
8. **Pilot sizing (Section 4.7).** Fifty-three paired tasks give a 95% confidence-interval half-width of $h = 0.135$ on a compliance rate, about 21% of the expected prose-to-structured effect of 0.632.
9. **Density-model refutation (Section 4.8).** The token-density model predicts impossible reductions for any constrained-token fraction above 0.414 and is rejected under the anchor of [8].

The honest summary is that the framework's predictions are conditional at every link: the sensitivity weights, the per-field compliance values, and the halved null-convergence probability are all assumptions awaiting the pilot's estimates. What the analysis establishes unconditionally is the structure — which component of variance prompt form can act on, what quantity measures the reduction, and what experiment would estimate it.

6. Discussion

The central claim of this paper is a conjecture with a specified test, not an established empirical result. The strongest internal evidence for it is structural: the decomposition of [8] shows prompts are a first-order variance source, the sensitivity machinery of [1] and [5] supplies a well-defined object for measuring which prompt-form features carry that variance, and the design tradition of [2], [4], and [9] shows variance is controllable by input and architectural choices in adjacent domains. None of this constitutes evidence that structured prompts in fact reduce cross-LLM variance at the predicted magnitude.

Limitations. The result rests on the following assumptions, each with its sensitivity:

The result rests on the following assumptions, each with its sensitivity:

- **The six-dimension sensitivity weight vector (Section 4.2).** The values $S_{T_1} = 0.30, \ldots, S_{T_6} = 0.08$ are illustrative, not estimated. The 82% reduction figure is sensitive to them: if the four pinned parameters carried only half the weight the assumption assigns them ($\sum_{i=1}^{4} S_{T_i} = 0.41$ instead of 0.82), the reduction would fall to 41%; if the interpretation space had ten dimensions with the top four carrying 0.60, the reduction would be 60%. The qualitative corollary — that pinning the highest-sensitivity parameters dominates pinning the lowest — is robust to any weight vector in which the indices are not all equal, but the headline magnitude is not.
- **The additivity bound from the decomposition of [8] (Section 4.1).** The 29.9-percentage-point bound uses the lower-bound sum 87.93% and the prompt share 36.43%. Because the reported component shares of [8] are not additive over their ranges (upper-bound sum 121.38%), the true prompt share of total variance could be lower than 36.43/87.93 = 41.4%, which would tighten the bound further; the bound is an upper limit, and the direction of the error is conservative.
- **The per-field compliance values (Section 4.4).** $q_{\text{struct}} = 0.986$ and $q_{\text{prose}} = 0.90$ are assumptions; the structured value is anchored to the per-constraint threshold $q^* \approx 0.986$ derived in [12], but the prose baseline of 0.90 has no source. The 3.57-fold violation ratio is sensitive to the baseline: at $q_{\text{prose}} = 0.95$, the prose full-compliance probability would be $0.95^{20} = e^{20 \times (-0.051293)} = e^{-1.025866} = 0.358486 \approx 35.8\%$, and the violation ratio would fall to $0.641514 / 0.245710 = 2.61$. The effect direction survives, the magnitude does not.
- **Field independence (Section 4.5).** The equicorrelation deflation model is a crude device; real schema errors correlate through shared formatting failures and context-window truncation, not through a single equicorrelation parameter. The model predicts the structured advantage strengthens with correlation, but a correlation structure concentrated entirely in fields the schema does not pin could reverse the comparison.
- **The halved null-convergence probability (Section 4.6).** The Bayes-factor amplification from 8 to 16 and the panel-size reduction from four models to three are conditional on the assumption that structured prompts halve p_b^S, the null-hypothesis convergence probability of [14]. No source measures this halving; it is the single most consequential untested assumption in the paper. If structured prompts reduce p_b^S by only 20% ($K'_S = 1.25 \times 8 = 10$), the two-model posterior would be $(10/9)/(1 + 10/9) = 10/19 \approx 0.526$, a modest gain, and four models would still be needed for a posterior of at least 0.95.
- **The pilot sizing (Section 4.7).** The $n = 53$ figure assumes a binomial proportion with worst-case variance and ignores pairing efficiency, model-count multiplication, and seed replication; the true required corpus is larger once multiple models and seeds are included, and 53 is a per-cell floor, not a total.

What would falsify the central claim. The claim is falsified by a specific observation from the pilot protocol of Section 7: if paired prose and structured prompts, run on the same tasks across at least three models at fixed temperature, show no reduction in inter-model output variance attributable to prompt form — operationally, if the structured arm's interpretation-variance share (estimated by the sparse-PCE machinery of [1] over the (prompt form, model, seed) configuration space) is not lower than the prose arm's by a margin exceeding the confidence half-width of Section 4.7 — the conjecture fails. A weaker falsifier targets the mechanism rather than the magnitude: if structured prompts reduce variance on schema-compliance metrics but not on semantic-agreement metrics, the support-restriction model of Section 3.3 is wrong about what structured forms pin, and the framework reduces to a formatting trick rather than an interpretation-control lever. A third falsifier is arithmetic: if a measured prompt-attributable variance share under structured forms exceeds the prose share of 36.43% reported in [8], the premise that structured forms shrink the prompt component is directly contradicted.

Failure modes. Three are visible from the model itself. First, structured prompts can *concentrate* rather than remove divergence: if a schema pins the format but leaves the semantic interpretation free, models may produce well-formed outputs with divergent content, and the variance moves to a component the compliance metric cannot see — the framework's own corollary from Section 4.2, that which parameters are pinned matters more than how many, cuts both ways. Second, over-constrained prompts can degrade task performance, trading variance for correctness; a schema that pins edge-case handling wrongly forces every model into the same wrong behavior, which is consistent output, not correct output. Third, the verification-amplification result of Section 4.6 inherits the shared-bias limitation that [14] itself states: correlated errors from overlapping training corpora deflate per-model evidence, and pinning the argument structure does nothing about a bias shared by all models.

Open questions. The framework leaves four questions open. (i) What is the actual dimensionality k of the interpretation space for realistic tasks, and are the sensitivity indices heavy-tailed enough that pinning a small set captures most of the variance? (ii) Do structured forms transfer across models without per-model tuning, or does each model require its own schema dialect — a question the pilot's cross-model design would answer? (iii) How does the interpretation-variance share scale with task complexity, and does the 82% illustrative reduction grow or shrink for tasks with more degrees of freedom? (iv) Does variance reduction interact with model capability, in the sense that stronger models resolve prose ambiguity more consistently, shrinking the gap the framework predicts?

7. Practical Implications and Engineering Pathways

(a) Who would use the result and what changes. The immediate users are engineering teams that integrate LLMs into automated pipelines: data-extraction services, document-processing workflows, and multi-step agentic systems. If the conjecture holds, the design decision that changes is the prompt-authoring standard: teams would write task prompts as configuration artifacts — explicit field schemas, enumerated processing orders, IF/THEN edge-case protocols — and version-control them like code, rather than iterating on prose. The measurable design target is the sensitivity-weighted pinning rule of Section 4.2: pin the highest-S_{T_i} interpretation parameters first, and accept that pinning the wrong parameters yields only a 45% reduction where pinning the right ones yields 82% under the illustrative weights. A second user group is the builders of multi-model verification panels of the kind formalized in [14]: if structured prompts halve the null-convergence probability, a three-model panel reaches the same evidential standard that otherwise requires four, a 25% reduction in per-verification inference cost under the cascade model of Section 4.6.

(b) Current readiness. This work is purely analytical and has not been peer reviewed; against the NASA Technology Readiness Level definitions it sits at TRL 1–2: the basic principles are stated, the variance-reduction factor is defined, and the quantitative predictions are worked out on paper, but no experimental evidence exists for the central conjecture.

The framework cannot exceed TRL 3 (proof of concept) until the pilot protocol below is run and the predicted variance reduction is observed on real models. No claim of deployment readiness, manufacturing readiness, or production validation is made or supportable.

(c) The smallest next experiment. The minimal de-risking study is the paired-corpus pilot specified in the research idea and sized in Section 4.7: a study would construct paired prompt corpora — one prose, one structured — describing identical tasks, run each through at least three LLMs under fixed temperature across multiple seeds, and measure inter-model agreement (structural similarity of outputs, semantic equivalence, schema-compliance rates) and intra-model consistency across seeds. The falsifying outcome is stated in Section 6: no structured-arm reduction in the prompt-attributable interpretation-variance share beyond the 0.135 confidence half-width of Section 4.7. A still smaller precursor, runnable before the full pilot, is a single-model, single-task compliance probe: generate $n = 53$ outputs each under the prose and structured prompts for one 20-field schema and compare the observed full-compliance rates against the predicted 12.2% and 75.4%; if the structured arm's observed rate falls below the prose arm's, the framework is refuted at the cheapest possible scale. This mirrors the minimal-experiment structure of [12], whose proposed pilot measures token counts, style violations, and rubric-scored depth, and whose falsifying thresholds are stated in advance.

(d) Binding constraints. The constraints that would bind first, in order: (i) *schema design cost* — writing a structured prompt that pins the right parameters requires knowing the S_{T_i} ranking, which no team has; the emulator-over-configurations template of [10] is the proposed route, and its cost is the dominant early expense; (ii) *evaluation data* — structural and semantic agreement metrics need validated definitions; the parser and paraphrase-robust matcher that [14] requires for argument-structure equivalence can fail and contaminate the convergence statistic in an unknown direction, and the same instruments would be needed here; (iii) *compute* — the pilot's cell structure (3 models × 2 prompt forms × 53 tasks × multiple seeds) is modest by LLM-evaluation standards, comparable in scale to the 12,000-output corpus of [8], so compute is unlikely to bind first; (iv) *correlated-error structure* — for the verification application, shared training corpora impose a floor on achievable evidence that prompt form cannot lower, as [14] states in its own limitations. Regulatory or safety constraints are not identified by any source here and remain an open question.

(e) Order-of-magnitude cost bound (projection). As a labeled projection with stated assumptions: if per-run inference cost is approximated by the cloud pricing illustration in [12] — $0.02 $per 1M tokens, with a task output of order 10^3 tokens — a single output costs of order $0.00002, and the full pilot of Section 4.7 (3 models × 2 arms × 53 tasks × an assumed 5 seeds = 1,590 runs) costs of order $0.03 in inference alone, excluding evaluation labor and metric development. This projection inherits every assumption of the pricing figure in [12] and is stated only to show that inference cost is not the binding constraint; the binding constraints are those of item (d).

(f) Direct risks and misuse paths. The main misuse path is the one stated in Section 6: consistent output is not correct output. A team that adopts structured prompts to pass compliance checks could pin wrong edge-case behavior into every model simultaneously, manufacturing uniform errors that are harder to detect than divergent ones — the failure is silent by construction. A second path follows from Section 4.6: if structured prompts raise inter-model agreement, agreement statistics become weaker evidence of correctness unless the null-convergence probability is re-measured under the structured regime; a verification panel that keeps the baseline $K_S = 8$ of [14] while adopting structured prompts would overstate its posterior confidence whenever the halving assumption fails in the direction of *raising* p_b^S for the structured arm's specific argument templates.

8. Conclusion

This paper built an analytical framework that treats a prompt as a partial program whose unspecified parameters each model fills by interpretation, and it derived a computable object — the variance-reduction factor R, equal to the share of interpretation variance carried by the parameters a structured prompt leaves free — that converts "structured prompts

reduce cross-LLM variance" from a craft intuition into a measurable design target. Under explicitly labeled illustrative assumptions, pinning the four highest-sensitivity of six interpretation dimensions removes 82% of the prompt-attributable interpretation variance, bounded by the empirical decomposition of [8] to at most 29.9 percentage points of total variance; a 20-field schema at the assumed per-field compliances yields a 3.57-fold reduction in violation probability, strengthening to 4.91 under moderate error correlation; and a halved null-convergence probability doubles the synthesis Bayes factor of [14] from 8 to 16, cutting the required verification panel from four models to three. The rival token-density model was shown to be quantitatively incoherent under the same empirical anchor, which motivates the sensitivity-weighting convention the paper adopts.

The honest status of the result is conditional at every quantitative link: the sensitivity weights, the compliance values, and the halving assumption all await the 53-task paired pilot specified in Section 7, whose falsifying outcome is stated in advance. What the paper establishes unconditionally is the structure of the problem — which variance component prompt form can act on, which quantity measures the reduction, which parameters to pin first, and what experiment would settle the question. If the pilot confirms the prediction, prompt engineering becomes a specification discipline with measurable targets; if it refutes the prediction, the framework's decomposition machinery still localizes where the variance lives, which is itself the information any remediation would need.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.55.12). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, openai/gpt-oss-120b, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: the execution sandbox was unavailable, so a second language model recomputed 15 quantitative claims step by step in text from the paper's stated inputs (no code was executed); 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] Identifying the Influential Inputs for Network Output Variance Using Sparse Polynomial Chaos Expansion. arXiv:1907.04266. https://arxiv.org/abs/1907.04266 [2] Unflattening by Flattening -- How Input Distributions Shape Output Variance in Angle-Encoded Circuits. arXiv:2610.01446. https://arxiv.org/abs/2610.01446 [3] An Investigation of the Sequential Sampling Method for Crossdocking Simulation Output Variance Reduction. arXiv:0803.1985. https://arxiv.org/abs/0803.1985 [4] LDLT L-Lipschitz Network Weight Parameterization Initialization. arXiv:2601.08253. https://arxiv.org/abs/2601.08253 [5] Orthogonal variance decomposition based feature selection.

arXiv:1910.09851. https://arxiv.org/abs/1910.09851 [6] Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples. arXiv:1907.07552. https://arxiv.org/abs/1907.07552 [8] Within-Model vs Between-Prompt Variability in Large Language Models for Creative Tasks. arXiv:2601.21339. https://arxiv.org/abs/2601.21339 [9] The Curse of Depth in Large Language Models. arXiv:2502.05795. https://arxiv.org/abs/2502.05795 [10] Connecting the forward problem to the inverse problem in uncertainty quantification of Earth system models using fast emulators. arXiv:2604.19500. https://arxiv.org/abs/2604.19500 [12] https://papers.qnfo.org/papers/close-slobphase1drafting-v10. QNFO: Quantitative Assessment of AI-Assisted Long-Form Drafting: Completion Rates and Reliability Thresholds. [14] https://papers.qnfo.org/papers/cross-model-llm-convergence-as-bayesian-evidence-for-scientific-claim-correctnes. QNFO: Cross-Model LLM Convergence as Bayesian Evidence for Scientific Claim Correctness: A Formal Framework, Worked Derivations, and a Calibration Protocol.