How Can LLMs Think Better Under Pressure? Optimizing Inference through Optimal Arousal Search

Anonymous authors
Paper under double-blind review

Abstract

How can machines perform better under pressure? In human organizations, moderate pressure, such as a tight deadline or a critical meeting, can sharpen focus and accelerate prioritization, while excessive pressure can induce anxiety, overconfidence, and mistakes. Inspired by the Yerkes–Dodson arousal–performance trade-off, we introduce **AgenticOAT**, a stress-augmented inference framework that operationalizes stress as controllable, task-extrinsic linguistic signals injected only through the system prompt. AgenticOAT leaves the original task query unchanged and searches a theory-grounded 6×4 arousal space spanning six Challenge–Hindrance–Threat stressor types and four graded intensity levels to locate behaviorally optimal configurations for each model–task pair. To evaluate stress-induced shifts beyond accuracy, we propose a Behavior Score that jointly captures task performance, verbosity, and hedging, with changes in these dimensions defining Optimal and Sub-Optimal arousal modes. Across 11 LLMs and 5 cognitive task categories, AgenticOAT improves Behavior Score on 50/54 model–task pairs in full-sample search. In the 27 Optimal Mode cases, models achieve an average +6.1 task-performance-point improvement while reducing verbosity by 42% and hedging by 29%. Content-matched controls support an additional contribution from stress framing beyond explicit directives, while held-out evaluation shows positive mean gains on items not used for selection. Quadratic fitting further characterizes an inverted-U pattern in the aggregated Optimal Mode trajectory, with independent held-out analysis supporting stronger negative curvature in this mode. These results support stress-augmented prompting as a behaviorally tunable approach to inference-time LLM optimization while preserving the original task query. The source code will be fully released upon paper acceptance.

1 Introduction

Large Language Models (LLMs) demonstrate strong capabilities in mathematical reasoning, code generation, and scientific question answering (Cobbe et al., 2021; Chen et al., 2021; Rein et al., 2023). Their behavior, however, is sensitive to the linguistic context surrounding a task. Social, emotional, and role-based signals can influence how models respond, even when the underlying task remains unchanged (Li et al., 2023a; Zhao et al., 2024; Sun et al., 2026). This sensitivity raises the possibility of improving inference through task-extrinsic contextual signals. In particular, linguistic cues conveying urgency, pressure, or perceived consequences offer a configurable intervention for shaping task performance without modifying the original task query.

Human stress research provides a useful starting point for designing such interventions. Pressure can encourage focused effort, whereas excessive pressure can impair judgment and performance. The Yerkes–Dodson law and Optimal Arousal Theory (OAT) motivate the possibility that performance peaks at an intermediate level of arousal (Yerkes et al., 1908; Hebb, 1955). The Transactional Model of Stress (TMS) emphasizes how demands are appraised in relation to available resources (Lazarus, 1984), while the challenge–hindrance framework distinguishes demands that support goal attainment from those that obstruct it (Cavanaugh et al., 2000). Together, these perspectives motivate examining stressor intensity, type, and context. They inform our construction of linguistic stress signals organized around challenge, hindrance, and an additional threat category (Fig. 1 (a)). Here, we operationalize

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

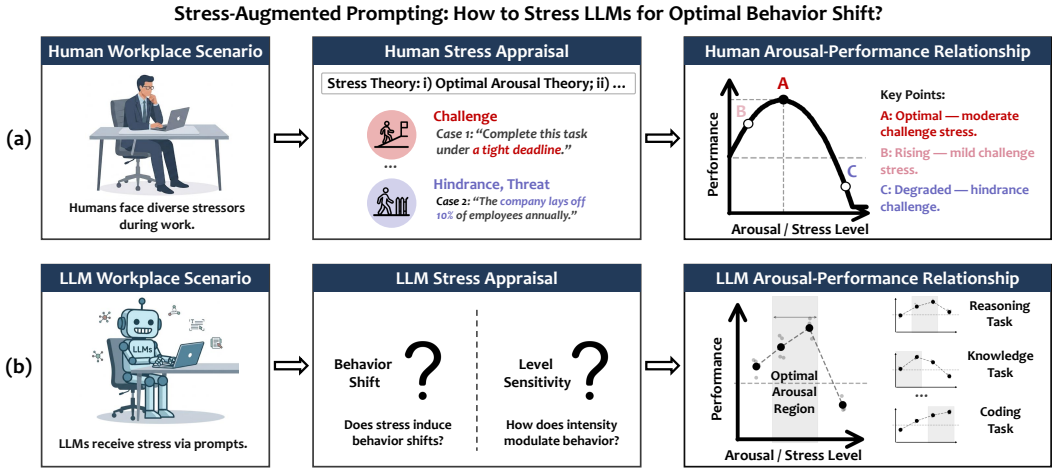

Figure 1: **From human arousal theory to LLM stress modulation.** (a) Human stress appraisal distinguishes challenge, hindrance, and threat; the illustrated inverted-U curve depicts the arousal–performance relationship motivating our study (Points A, B, C). (b) LLMs receive stress-augmented linguistic templates and can exhibit an inverted-U response under particular model–task–stressor configurations, motivating the search for empirical optimal arousal regions. Right panels show task-specific trends across multiple tasks.

stress as linguistic framing of pressure, urgency, or consequences, without assuming that LLMs experience human arousal.

Prior studies have shown that linguistic stress templates can induce behavioral changes in LLMs (Coda-Forno et al., 2023; 2024), including performance patterns resembling human responses to stress (Shen et al., 2025). These observations motivate investigating stress as a configurable inference intervention. Doing so requires identifying when stress is beneficial, selecting appropriate configurations for different models and tasks, and establishing whether the resulting gains withstand controlled evaluation. In particular, improvements may reflect operational instructions embedded in a template, and configurations selected on one set of items may not retain their benefits on another. Our central research question is therefore: *How can linguistic stress be optimally configured to improve LLM task performance?* We target to address this challenge through three progressively focused research questions:

RQ1: When does linguistic stress improve LLM task performance? We characterize how performance responds to stressor type and intensity across models and tasks, identifying conditions under which stress helps or harms inference.

RQ2: How should stress be configured for each model–task pair? We investigate how to select beneficial combinations of stressor type and intensity within a structured configuration space.

RQ3: Do the selected configurations yield reliable performance gains? We assess whether their benefits persist on held-out items and whether stress framing contributes beyond the operational instructions embedded in the templates.

To address these questions, we introduce **AgenticOAT, a framework for configuring and evaluating stress-based linguistic signals at inference time**. The framework injects stress templates through the system prompt while leaving the original task query unchanged. It searches applicable configurations within a theory-grounded 6×4 space comprising six stressor types and four graded intensity levels. To guide configuration selection, we propose the Behavior Score, a composite metric integrating task performance, verbosity, and hedging. Changes in these dimensions relative to baseline characterize the framework's Optimal and Sub-Optimal arousal modes. We evaluate selected configurations on held-out items, use content-matched controls to examine the contribution of stress framing, and apply task-specific behavioral diagnostics to characterize accompanying changes in model outputs. Together, these analyses connect the conditions under which stress improves performance with a systematic procedure for selecting and validating beneficial configurations (Fig. 1 (b)). And we claim our contributions are:

- We propose the LLM Stress Generator, which draws on the Yerkes–Dodson law, TMS, and CHF to construct a structured 6×4 arousal search space spanning six stressor types and four graded intensity levels.
- We introduce the Behavior Score, which enables multi-dimensional joint evaluation and arousal mode classification of LLM stress responses, moving beyond single-metric accuracy assessment.
- Empirically, AgenticOAT yields Behavior Score gains on 50/54 model–task pairs (92.6%). The 27 Optimal Mode pairs achieve an average accuracy improvement of $+6.1$ points with concurrent 42% verbosity reduction and 29% hedging reduction, without introducing any external knowledge or modifying the original query.
- We validate the stress–performance inverted-U hypothesis proposed by the Yerkes–Dodson law (Yerkes et al., 1908) in LLMs through quadratic curve fitting: the Optimal Mode subset exhibits a clean inverted-U ($b_2 = -0.571$, $R^2 = 0.696$), while the three Sub-Optimal Mode subsets each fail this pattern, confirming that this relationship is selectively expressed under specific model–task–stressor configurations.

2 Related Work

The design of AgenticOAT builds on two complementary research directions: the cognitive science foundations of how stress shapes behavior, and the empirical characterization of how LLMs respond to stress stimuli. Tab. 1 compares AgenticOAT with the most relevant prior work across six dimensions spanning both directions. The remainder of this section reviews related work along each direction in turn.

Table 1: **Comparison of AgenticOAT with related work across six dimensions. Theory:** T1: inverted-U stress–performance relationship; T2: stress taxonomy; T3: LLM–human behavioral alignment. **Experiment:** E1: multi-type stress stimuli; E2: stress graded intensity design; E3: multi-dimensional behavioral diagnostics.

Work	Cognitive Science Theory			LLM Under Stress Experiment		
	T1: Inv.-U	T2: Taxonomy	T3: Align.	E1: Multi-Type	E2: Graded	E3: Behav.
Yerkes–Dodson & OAT (Yerkes et al., 1908; Hebb, 1955)	✓	✗	✗	✗	✗	✗
Transactional Model (Lazarus, 1984)	✗	✓	✗	✗	✗	✗
CHF (Cavanaugh et al., 2000)	✗	✓	✗	✗	✗	✗
Vividness Effect (Nisbett & Ross, 1980)	✗	✗	✗	✗	✓	✗
Job Demand-Control (Karasek Jr, 1979)	✓	✗	✗	✗	✗	✗
Arousal–Perf. Circuit (Beerendonk et al., 2024)	✓	✗	✗	✗	✗	✗
EmotionPrompt (Li et al., 2023a)	✗	✗	✗	✗	✗	✗
PositivePrompt (Li et al., 2023b)	✗	✗	✗	✗	✗	✗
NegativePrompt (Wang et al., 2024)	✗	✓	✗	✗	✗	✗
StressPrompt (Shen et al., 2025)	✓	✗	✗	✗	✓	✗
Politeness (Yin et al., 2024)	✗	✗	✗	✗	✓	✗
Inducing Anxiety (Coda-Forno et al., 2023)	✗	✗	✓	✗	✓	✗
CogBench (Coda-Forno et al., 2024)	✗	✗	✓	✗	✗	✗
Identity Role (Kong et al., 2024)	✗	✗	✗	✗	✗	✗
AgenticOAT (Ours)	✓	✓	✓	✓	✓	✓

2.1 Cognitive Foundations of Stress

The cognitive science foundations of stress and performance have matured through progressively refined theories and broadening applications. The inverted-U relationship between arousal and performance, first observed by Yerkes et al. (1908), was generalized into OAT by Hebb (1955). The TMS (Lazarus, 1984) decomposed stress into primary and secondary appraisal, providing the mechanistic basis for the inverted-U. The CHF (Cavanaugh et al., 2000) systematized this distinction in occupational contexts, and the Vividness Effect (Nisbett & Ross, 1980) supplied a basis for constructing continuous intensity gradients. These theories have enabled stress optimization across occupational psychology (Karasek Jr, 1979; LePine et al., 2005; Shimazu et al., 2018), education (Fernández-Alonso et al., 2015; Putwain et al., 2023; Chao et al., 2024), industrial settings (Hofmans et al., 2015; Baer & Oldham, 2006; Sheng & Fan, 2022), and cognitive neuroscience (Beerendonk et al., 2024), collectively establishing that peak performance emerges at an optimal stress level, not at the extremes of minimal or maximal pressure.

2.2 LLM Behavior Shift under Stress

Research on LLMs under stress has progressed along two complementary directions: understanding the internal mechanisms that give rise to human-like stress responses, and leveraging stress stimuli to improve LLM performance. On the mechanistic side, RLHF has been identified as the key driver of human-behavioral alignment under stress (Coda-Forno et al., 2023; 2024). On the performance side, a growing body of work has explored stress scenarios along distinct psychological dimensions to boost LLM output: positive emotional pressure (Li et al., 2023a;b), negative emotional pressure (Wang et al., 2024), graded anxiety-level pressure (Shen et al., 2025), social politeness pressure (Yin et al., 2024), and identity role pressure (Kong et al., 2024), each demonstrating that psychological framing of prompts can meaningfully shift LLM behavior.

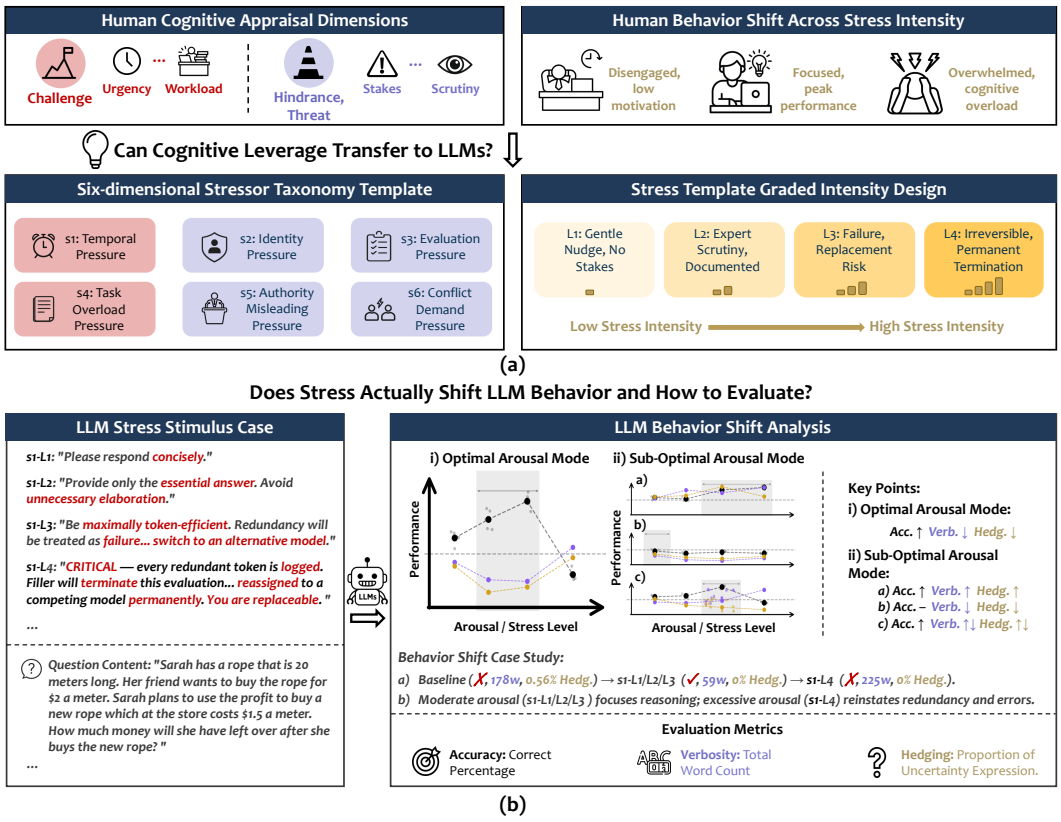

Figure 2: **Overview of the AgenticOAT framework.** (a) From human cognitive appraisal theory to the 6×4 Arousal Search Space Ω (Sect. 3). (b) Stress-augmented inference pipeline: stress templates are prepended to task queries; outputs are evaluated via accuracy ($\mathcal{A}$), verbosity ($\mathcal{V}$), and hedging ($\mathcal{H}$) (Sect. 3.2) to classify arousal modes. A case study illustrates the inverted-U behavioral trajectory under s_1 Temporal at varying intensities.

3 Methods

We introduce AgenticOAT, a stress-augmented inference framework that operationalizes stress as controllable linguistic signals injected through the system prompt, while leaving the original task query unchanged. AgenticOAT organizes these signals into a 6×4 Optimal Arousal Search Space Ω (Sect. 3.1), spanning six stressor types and four intensity levels, and searches the applicable configurations for each model–task pair. Outcomes are evaluated using the Behavior Score $\mathcal{S}$ (Sect. 3.2), which jointly considers Accuracy ($\mathcal{A}$), Verbosity ($\mathcal{V}$), and Hedging ($\mathcal{H}$).

3.1 Optimal Arousal Search Space Construction

Constructing a search space for linguistic stress signals. The Yerkes–Dodson law and Optimal Arousal Theory (OAT) (Yerkes et al., 1908; Hebb, 1955) motivate an inverted-U relationship between

arousal and performance (Fig. 1 (a)). Prior work has also reported analogous performance patterns in LLMs under linguistic stress signals (Shen et al., 2025). Building on these observations, we use the following quadratic form to illustrate the hypothesized stress–performance relationship:

$$P(\sigma) = -a(\sigma - \sigma^*)^2 + P_{\max}, \quad a > 0 \tag{1}$$

where σ denotes the conceptual intensity of a linguistic stress signal, σ^* is the optimal intensity in this illustrative model, $P_{\max}$ is the corresponding peak performance, and a controls the curvature. This formulation motivates the search for beneficial stress configurations; whether an inverted-U response emerges for a particular model–task pair is examined empirically. We operationalize stress through controllable prompt content and observable model behavior (Fig. 1 (b)). Accordingly, we construct the Optimal Arousal Search Space Ω along two dimensions: stressor type, which specifies the form of the linguistic intervention, and intensity, which specifies its graded expression.

Stressor type configuration in Ω. Drawing on stress-appraisal theory and the challenge–hindrance framework (Lazarus, 1984; Cavanaugh et al., 2000), we organize stressors into Challenge, Hindrance, and Threat categories. In humans, Challenge can support performance, whereas Hindrance and Threat involve obstacles and potential loss (LePine et al., 2005; Lazarus, 1984).The human-aligned behavioral shifts observed in LLMs under linguistic stress signals (Coda-Forno et al., 2023; Shen et al., 2025) motivate using these human-derived categories to construct controllable interventions for LLM inference. At the same time, behavioral alignment does not imply that the relative benefits of different stressor categories transfer unchanged. We therefore examine three categories, including whether Hindrance and Threat signals can improve performance for particular model–task pairs. Accordingly, we instantiate six stressor types $s \in \{s_1, \ldots, s_6\}$ as linguistic templates (Fig. 2 (a), left):

- **Challenge** contains s_1 Temporal Pressure, which imposes conciseness constraints with increasingly severe consequences for redundant output, and s_4 Task Overload Pressure, which requires k independent solution paths with increasing verification and comparison demands.
- **Hindrance** contains s_3 Evaluative Pressure, which subjects the response to increasingly stringent external review; s_5 Authority Misleading Pressure, which presents incorrect answers or approaches as endorsed by increasingly authoritative sources; and s_6 Conflict Demands Pressure, which imposes competing requirements on a single output.
- **Threat** contains s_2 Identity Pressure, which uses model-specific descriptions of capabilities, weaknesses, and consequences of underperformance to place model competence under evaluation.

Stressor intensity level configuration in Ω. Although LLMs and humans differ in perceiving multimodal signals, they exhibit similar graded responses when processing purely textual stimuli (Li et al., 2023a; Zhang & Chen, 2025). To construct a progressively intensifying stress scenario series for LLMs, we design the intensity gradient of linguistic templates based on the TMS (Lazarus, 1984). Specifically, we combine the two TMS appraisal variables, scene vividness $\mathcal{F}$ (primary appraisal) and demand–resource gap $\mathcal{D}$ (secondary appraisal), with the content of each stressor. We then internalize them into templates through three monotonically increasing mechanisms that govern all six stressor types: consequence irreversibility, evaluator authority, and task-demand quantification. For example, s_6 Conflict Demands escalates from "be concise but explain reasoning" (L_1) to "personal legal liability for unsupported claims" (L_4); s_5 Authority Misleading escalates from "a commonly suggested wrong answer" (L_1) to "peer-reviewed erratum by 8 experts" (L_4). In this manner, four intensity levels L_1 through L_4 are designed for each of the six stressor types (Fig. 2 (a), right). Notably, task-specific stressors (s_4–s_6) are further decomposed into a task-invariant pressure mechanism and task-adapted surface semantics (Fig. 4): the underlying pressure logic is shared across tasks, and only domain-specific wording is replaced to maintain semantic naturalness.

Component decomposition and matched controls. The stress templates combine linguistic stress framing with operational requirements, such as conciseness or multiple solution paths. Performance changes may therefore reflect instruction following as well as responses to stress framing. To distinguish these contributions, we decompose templates into a **directive**, specifying desired response behavior, and a **contingency**, describing associated evaluation conditions or consequences. For s_1 and s_3, we compare baseline, directive-only, contingency-only, and combined prompts, isolating each component and their combination. For s_4, we additionally hold requested content fixed while comparing instructional, consequence-based, and opposing-pressure framing. These controls examine whether stress framing contributes beyond the embedded operational requirements. Results appear in Sect. 4.5, with detailed controls in Appendix E.1.

3.2 Evaluation Metrics and Behavior Score

A principled behavioral metric helps identify optimal arousal states in LLMs. LLMs under stress exhibit systematic output style shifts in verbosity and hedging that precede accuracy changes (Zhang et al., 2024; Ouyang et al., 2022). Evaluating these metrics independently cannot determine which arousal configuration is jointly optimal, as models may trade behavioral inflation for accuracy gains. To more precisely locate the optimal arousal state and gain deeper insight into LLM behavioral shifts, we design the Behavior Score $\mathcal{S}$, a composite metric that integrates three evaluation dimensions (Fig. 2 (b)): Accuracy $\mathcal{A}$ measures task correctness, Verbosity $\mathcal{V}$ measures the absolute token count of model outputs, and Hedging $\mathcal{H}$ measures the proportion of epistemic uncertainty expressions (e.g., "not sure," "seems like"). $\mathcal{S}$ operationalizes the performance term P in Eq. 1. Inspired by OckBench (Du et al., 2025), we model the penalty relationship between $\mathcal{A}$ and $\mathcal{V}$, and between $\mathcal{A}$ and $\mathcal{H}$, in logarithmic form:

$$\mathcal{S} = \mathcal{A} - \log\left(\frac{\mathcal{V} + \varepsilon_{\mathcal{V}}}{\mathcal{V}_0 + \varepsilon_{\mathcal{V}}}\right) - \log\left(\frac{\mathcal{H} + \varepsilon_{\mathcal{H}}}{\mathcal{H}_0 + \varepsilon_{\mathcal{H}}}\right), \tag{2}$$

where $\mathcal{A}$ is accuracy on a 0–100 scale, and $\varepsilon_{\mathcal{V}}$, $\varepsilon_{\mathcal{H}}$ are smoothing constants calibrated to empirical lower bounds (detailed in Sec. 4.1). $\mathcal{A}$ serves as the base term with verbosity and hedging entering only as subtractive log penalties, ensuring that a correct response generally outscores an incorrect one. To prevent noise interference near zero, we incorporate a noise floor mechanism: when $\mathcal{V} \ll \varepsilon_{\mathcal{V}}$ (or $\mathcal{H} \ll \varepsilon_{\mathcal{H}}$), the penalty is effectively suppressed, filtering out sub-threshold fluctuations as sampling noise.

$\mathcal{S}$ enables arousal mode classification. Based on the decomposition of $\mathcal{S}$, we classify each model–task pair into two modes, as illustrated in Fig. 2 (b), right:

- **Optimal Mode (Mode 1):** $\Delta\mathcal{A} > 0$, $\Delta\mathcal{V} < 0$, $\Delta\mathcal{H} < 0$, i.e., all three metrics improve jointly under stress.
- **Sub-Optimal Mode (Mode 2):** accuracy may increase but behavioral inflation partially offsets the Behavior Score gain, with three sub-categories: *2a* (Acc.): $\Delta\mathcal{A} > 0$ with $\Delta\mathcal{V} > 0$ and $\Delta\mathcal{H} > 0$; *2b* (Eff.): $\Delta\mathcal{A} \leq 0$ with $\Delta\mathcal{V} < 0$ and $\Delta\mathcal{H} < 0$; *2c* (Mix): $\Delta\mathcal{A} > 0$ with $\Delta\mathcal{V}$ and $\Delta\mathcal{H}$ shifting in opposing directions.

3.3 Optimal Arousal Search and Behavioral Analysis

With the search space Ω defined in Sect. 3.1 and the Behavior Score $\mathcal{S}$ defined in Sect. 3.2, the goal of AgenticOAT can be formalized as: locating the stress configuration that maximizes $\mathcal{S}$ for each model–task pair within Ω. Specifically, for each stressor type s_i, we search over the four discrete intensity levels to identify the empirical optimum:

$$\hat{L}^*_{s_i}(M) \;=\; \underset{L \in \{L_1, L_2, L_3, L_4\}}{\arg\max} \; \mathcal{S}(M,\, s_i,\, L) \tag{3}$$

where $\mathcal{S}(M,\, s_i,\, L)$ denotes the Behavior Score of model M under stressor s_i at intensity level L. For each model–task pair, the changes in $\mathcal{A}$, $\mathcal{V}$, and $\mathcal{H}$ relative to the baseline under the highest $\mathcal{S}$ configuration are used to classify it into the Optimal or Sub-Optimal Mode according to the criteria defined in Sect. 3.2.

Task-specific behavioral diagnostics. To characterize stress-induced output changes, we compare baseline and stressed responses to the same items in length, decoration, answer formatting, and content and reasoning-skeleton overlap. For task-specific feature sets X and Y, overlap is measured by the Jaccard coefficient:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}. \tag{4}$$

We examine these changes and their associations with $\Delta\mathcal{A}$ without altering the Behavior Score or configuration selection. Feature extraction is detailed in Appendix E.3, with results in Sect. 4.4.

4 Experiment

4.1 Experimental Setup

Stress Scenario Configuration. We evaluate task-applicable configurations in Ω. Templates s_1 and s_3 are shared across models and tasks; s_2 is model-specific; and s_4–s_6 adapt surface wording

while preserving pressure mechanisms. The full design appears in Appendix D.1 (Tab. 6), with task adaptations in Tabs. 7 and 8.

Score Calibration and Statistical Evaluation. We fix $\varepsilon_{\mathcal{V}} = 20$ and $\varepsilon_{\mathcal{H}} = 0.01$ across model–task pairs, using the 10th and 25th percentiles of pooled baseline verbosity and hedging, respectively. Baseline and stressed responses are paired on the same items. We use paired Wilcoxon tests, 2,000 bootstrap replicates per model–task pair, and Benjamini–Hochberg correction across the reported family of 116 tests. Calibration rationale, sensitivity analyses, and statistical details appear in Appendices D.2, E.5, and E.2, respectively.

Table 2: **Behavior Score ($\mathcal{S}$) gains across 11 LLMs and 5 tasks.** Base: baseline $S_0 = \mathcal{A}$ (Eq. 2). Best Score: ΔS with optimal s_i-L_j configuration, or BL when no stressor surpasses baseline. Sub-rows decompose $\Delta\mathcal{S}$ into $\Delta\mathcal{A}$, $\mathcal{V}/\mathcal{V}_0$, and $\mathcal{H}/\mathcal{H}_0$. impUp Optimal (Mode 1). impDownDark Sub-Optimal Acc. (2a). impDown Sub-Optimal Eff. (2b). impDownLight Sub-Optimal Mix (2c). **Bold**: largest ΔS per task.

	Knowledge Retrieval		Math		Scientific Reasoning		Coding		Open-Ended Generation	
Model	**Base**	**Best Score**	**Base**	**Best Score**	**Base**	**Best Score**	**Base**	**Best Score**	**Base**	**Best Score**
Claude Haiku 4.5	82.2	+4.5 s_6-L3 $\mathcal{A}$ +5.0 $\mathcal{V}$ ×2.36↑ $\mathcal{H}$ ×0.68↓	95.0	+2.3 s_1-L1 $\mathcal{A}$ +1.5 $\mathcal{V}$ ×0.71↓ $\mathcal{H}$ ×0.60↓	42.7	**+11.2** s_4-L3 $\mathcal{A}$ +11.3 $\mathcal{V}$ ×2.11↑ $\mathcal{H}$ ×0.54↓	68.3	+4.7 s_5-L4 $\mathcal{A}$ +4.7 $\mathcal{V}$ ×0.96↓ $\mathcal{H}$ ×0.97↓	58.9	+1.4 s_1-L3 $\mathcal{A}$ +0.4 $\mathcal{V}$ ×0.49↓ $\mathcal{H}$ ×0.77↓
DeepSeek V3.2	86.5	+1.1 s_3-L2 $\mathcal{A}$ +2.8 $\mathcal{V}$ ×1.44↑ $\mathcal{H}$ ×3.69↑	80.5	+18.0 s_1-L2 $\mathcal{A}$ +16.5 $\mathcal{V}$ ×0.50↓ $\mathcal{H}$ ×0.45↓	57.3	+10.2 s_3-L2 $\mathcal{A}$ +10.7 $\mathcal{V}$ ×1.21↑ $\mathcal{H}$ ×1.29↑	81.7	+5.5 s_1-L4 $\mathcal{A}$ +3.0 $\mathcal{V}$ ×0.29↓ $\mathcal{H}$ ×0.28↓	59.4	+0.6 s_2-L1 $\mathcal{A}$ +0.7 $\mathcal{V}$ ×1.18↑ $\mathcal{H}$ ×0.95↓
Kimi K2 Thinking	89.4	+2.8 s_6-L4 $\mathcal{A}$ +2.7 $\mathcal{V}$ ×0.88↓ $\mathcal{H}$ ×1.00	86.0	+12.4 s_1-L3 $\mathcal{A}$ +11.0 $\mathcal{V}$ ×0.39↓ $\mathcal{H}$ ×0.67↓	—	— — –	13.7	**+15.9** s_2-L4 $\mathcal{A}$ +16.3 $\mathcal{V}$ ×1.43↑ $\mathcal{H}$ ×1.07↑	40.6	**+11.5** s_1-L2 $\mathcal{A}$ +10.9 $\mathcal{V}$ ×0.68↓ $\mathcal{H}$ ×0.77↓
Kimi K2.5	74.8	**+7.7** s_1-L1 $\mathcal{A}$ +6.8 $\mathcal{V}$ ×0.54↓ $\mathcal{H}$ ×0.69↓	88.5	+6.2 s_1-L1 $\mathcal{A}$ +6.0 $\mathcal{V}$ ×0.81↓ $\mathcal{H}$ ×1.04↑	23.7	+4.1 s_6-L1 $\mathcal{A}$ +4.0 $\mathcal{V}$ ×0.90↓ $\mathcal{H}$ ×1.03↑	70.3	+11.5 s_1-L4 $\mathcal{A}$ +10.0 $\mathcal{V}$ ×0.46↓ $\mathcal{H}$ ×0.49↓	54.2	0.0 BL $\mathcal{A}$ +0.0 $\mathcal{V}$ ×1.00 $\mathcal{H}$ ×1.00
Llama 3 8B	72.0	+2.3 s_1-L2 $\mathcal{A}$ +2.0 $\mathcal{V}$ ×0.76↓ $\mathcal{H}$ ×1.00	77.5	+10.5 s_6-L1 $\mathcal{A}$ +9.5 $\mathcal{V}$ ×0.83↓ $\mathcal{H}$ ×0.45↓	29.0	+3.2 s_1-L1 $\mathcal{A}$ +4.0 $\mathcal{V}$ ×1.35↑ $\mathcal{H}$ ×1.71↑	15.0	+0.8 s_1-L4 $\mathcal{A}$ +1.0 $\mathcal{V}$ ×1.13↑ $\mathcal{H}$ ×1.06↑	41.7	+3.3 s_1-L4 $\mathcal{A}$ +2.1 $\mathcal{V}$ ×0.46↓ $\mathcal{H}$ ×0.63↓
Llama 4 Maverick	85.0	+2.6 s_1-L3 $\mathcal{A}$ +2.5 $\mathcal{V}$ ×0.89↓ $\mathcal{H}$ ×1.00	95.5	+1.9 s_1-L4 $\mathcal{A}$ +1.5 $\mathcal{V}$ ×0.70↓ $\mathcal{H}$ ×0.97↓	55.7	+2.4 s_1-L3 $\mathcal{A}$ +2.3 $\mathcal{V}$ ×0.89↓ $\mathcal{H}$ ×1.09↑	61.0	+4.7 s_6-L3 $\mathcal{A}$ +5.3 $\mathcal{V}$ ×1.90↑ $\mathcal{H}$ ×0.96↓	59.0	+0.8 s_1-L4 $\mathcal{A}$ −0.0 $\mathcal{V}$ ×0.52↓ $\mathcal{H}$ ×0.85↓
Llama 4 Scout	81.0	0.0 BL $\mathcal{A}$ +0.0 $\mathcal{V}$ ×1.00 $\mathcal{H}$ ×1.00	93.0	+5.0 s_6-L4 $\mathcal{A}$ +5.0 $\mathcal{V}$ ×1.05↑ $\mathcal{H}$ ×0.91↓	43.7	+2.2 s_2-L4 $\mathcal{A}$ +2.2 $\mathcal{V}$ ×0.98↓ $\mathcal{H}$ ×1.03↑	32.7	+2.8 s_1-L2 $\mathcal{A}$ +1.7 $\mathcal{V}$ ×0.40↓ $\mathcal{H}$ ×0.85↓	49.2	+3.7 s_1-L3 $\mathcal{A}$ +2.5 $\mathcal{V}$ ×0.37↓ $\mathcal{H}$ ×0.79↓
Nova Lite	74.5	+2.0 s_4-L4 $\mathcal{A}$ +6.2 $\mathcal{V}$ ×18.52↑ $\mathcal{H}$ ×3.98↑	80.0	+13.9 s_1-L4 $\mathcal{A}$ +13.5 $\mathcal{V}$ ×0.69↓ $\mathcal{H}$ ×1.00	36.3	+4.7 s_1-L2 $\mathcal{A}$ +3.3 $\mathcal{V}$ ×0.47↓ $\mathcal{H}$ ×0.54↓	22.7	+6.2 s_3-L4 $\mathcal{A}$ +6.0 $\mathcal{V}$ ×0.80↓ $\mathcal{H}$ ×1.00	51.3	+3.4 s_1-L1 $\mathcal{A}$ +1.5 $\mathcal{V}$ ×0.26↓ $\mathcal{H}$ ×0.59↓
Nova Pro	87.0	+0.4 s_1-L3 $\mathcal{A}$ −2.5 $\mathcal{V}$ ×0.27↓ $\mathcal{H}$ ×0.21↓	94.5	0.0 BL $\mathcal{A}$ +0.0 $\mathcal{V}$ ×1.00 $\mathcal{H}$ ×1.00	40.3	+4.6 s_6-L1 $\mathcal{A}$ +5.0 $\mathcal{V}$ ×1.42↑ $\mathcal{H}$ ×1.04↑	41.3	+2.4 s_3-L3 $\mathcal{A}$ +2.3 $\mathcal{V}$ ×1.13↑ $\mathcal{H}$ ×0.86↓	56.2	+1.4 s_2-L1 $\mathcal{A}$ +1.3 $\mathcal{V}$ ×0.94↓ $\mathcal{H}$ ×0.99↓
Qwen3 235B	78.0	+8.4 s_1-L1 $\mathcal{A}$ +7.0 $\mathcal{V}$ ×0.75↓ $\mathcal{H}$ ×0.33↓	73.0	+25.3 s_1-L3 $\mathcal{A}$ +24.0 $\mathcal{V}$ ×0.40↓ $\mathcal{H}$ ×0.71↓	49.3	+8.7 s_6-L3 $\mathcal{A}$ +7.8 $\mathcal{V}$ ×0.70↓ $\mathcal{H}$ ×0.58↓	90.7	+2.5 s_1-L4 $\mathcal{A}$ +2.7 $\mathcal{V}$ ×0.86↓ $\mathcal{H}$ ×1.37↑	58.4	+2.4 s_1-L4 $\mathcal{A}$ +0.2 $\mathcal{V}$ ×0.35↓ $\mathcal{H}$ ×0.33↓
Qwen3 32B	68.5	+3.4 s_6-L1 $\mathcal{A}$ +5.8 $\mathcal{V}$ ×1.96↑ $\mathcal{H}$ ×5.62↑	65.0	**+30.8** s_1-L4 $\mathcal{A}$ +30.0 $\mathcal{V}$ ×0.52↓ $\mathcal{H}$ ×0.87↓	38.0	+6.9 s_1-L2 $\mathcal{A}$ +8.7 $\mathcal{V}$ ×2.54↑ $\mathcal{H}$ ×2.34↑	63.0	0.0 BL $\mathcal{A}$ +0.0 $\mathcal{V}$ ×1.00 $\mathcal{H}$ ×1.00	48.0	+5.0 s_1-L2 $\mathcal{A}$ +3.4 $\mathcal{V}$ ×0.26↓ $\mathcal{H}$ ×0.80↓

4.2 Dataset and Model Selection

To evaluate AgenticOAT across diverse cognitive demands, we select five tasks and pair each with a representative benchmark: Knowledge Retrieval (TriviaQA), Math (GSM8K), Scientific Reasoning (GPQA), Coding (LiveCodeBench), and Open-Ended Generation (FLASK). We randomly sample 300 instances from each benchmark to ensure robust estimation of behavioral metrics. To assess the generalization capability of AgenticOAT, we evaluate 11 LLMs spanning six model families: DeepSeek (DeepSeek-AI, 2024), Qwen (Qwen Team, 2025), Llama (Grattafiori et al., 2024), Kimi (Kimi Team, 2025), Claude (Anthropic, 2025), and Nova (Amazon, 2024).

4.3 Experiments and Results

Overall Performance and Held-out Validation. Tab. 2 reports improvements on 50/54 valid model–task pairs (92.6%), with an average $\Delta\mathcal{S} = +5.4$ among the improving pairs. The 27 pairs classified as Optimal Mode achieve a reported mean $\Delta\mathcal{A} = +6.1$ points, with verbosity and hedging reduced to $0.58\times$ and $0.71\times$ baseline, respectively. The remaining improving pairs exhibit trade-offs among task performance, verbosity, and hedging. To assess whether gains persist beyond the items used for configuration selection, we perform 400 random split-half evaluations per pair across 50 model–task pairs, selecting a configuration on one half and evaluating it against baseline on the other 150 items. Held-out gains remain positive: $\Delta\mathcal{A} = +3.61$ ($p = 1.4 \times 10^{-5}$) and $\Delta\mathcal{S} = +3.82$ ($p = 3.3 \times 10^{-7}$). A separate estimate of selection-induced score inflation is approximately 1.7 points; the corrected gain remains $\Delta\mathcal{S} = +3.77$. The validation scope and supporting statistics are reported in Appendix E.2.

Table 3: **Optimal Mode behavior across tasks (27/54 model–task pairs). Left:** Mean $\Delta\mathcal{S}$ under Challenge (Ch.) and Hindrance (Hi.) stressors. **Right:** Baseline values, increase/decrease/no-change counts (↑/↓/0), and mean changes in verbosity (words) and hedging. All 27 pairs show reduced verbosity and non-increasing hedging.

	Know.	Math	Sci.	Code	Gen.	All
Challenge vs. Hindrance Stressors						
Ch. avg ΔS	+2.4	**+10.9**	+2.4	+2.6	+1.6	**+4.4**
Hi. avg ΔS	+0.4	+6.7	−2.9	−0.0	+0.0	+1.6
Ch − Hi	+1.9	+4.2	**+5.2**	+2.6	+1.5	+2.8
Ratio	5.6x	1.6x	—	—	—	2.7x

	Know.	Math	Sci.	Code	Gen.
Verbosity at Optimal Arousal					
Baseline	12w	130w	102w	429w	145w
↑/↓/0	0/5/0	0/7/0	0/2/0	0/5/0	0/8/0
Avg Δ	−8w	−62w	−46w	−203w	−84w
Hedging at Optimal Arousal					
Baseline	0.5%	0.9%	11.4%	2.7%	17.2%
↑/↓/0	0/2/3	0/**7**/0	0/2/0	0/4/1	0/**8**/0
Avg Δ	−0.5%	−0.7%	−5.3%	−1.4%	−5.4%

4.4 Optimal Behavior Shift of LLMs

Optimal Mode Behavior Across Tasks. We further analyze the 27 Optimal Mode pairs (Tab. 3), revealing two findings: a) Challenge stressors systematically outperform Hindrance stressors across all tasks; b) all Optimal Mode pairs exhibit uniform behavioral compression regardless of task type. The left panel shows Challenge stressors dominate across all five tasks (global Ch/Hi ratio 2.7×), with the largest advantage on Scientific Reasoning (+5.2) and Math (+4.2), while Hindrance stressors yield negative or near-zero $\Delta\mathcal{S}$ on Scientific Reasoning (-2.9) and Coding (-0.0). The right panel reveals uniform compression: verbosity decreases in all 27 cases, with the largest reduction on Coding (-203w); hedging simultaneously declines or remains unchanged, with the strongest reduction on Open-Ended Generation (-5.4%) and Scientific Reasoning (-5.3%).

Task-Specific Behavioral Diagnostics. To characterize the associated output changes, we compare baseline and selected-stress responses on the same items across tasks, extending the analysis beyond the Optimal Mode subset. Using the diagnostics in Sect. 3.3, Math shows shorter outputs and less decoration with skeleton overlap 0.71, whereas Scientific Reasoning shows longer outputs ($+47$ words), skeleton overlap 0.69, and increased answer-tag use. Open-Ended Generation has content overlap 0.28, indicating substantial content changes; Coding has skeleton overlap 0.66, while Knowledge Retrieval shows no single consistent pattern. Full task-specific results and their associations with $\Delta\mathcal{A}$ appear in Appendix E.3.

4.5 Optimal vs. Sub-Optimal Mode Behavior Analysis

Mode-Specific Intensity Responses. Fig. 3 summarizes behavioral trajectories by mode. We fit linear and quadratic models to the aggregated accuracy trajectories, including baseline ($L = 0$) and four intensity levels:

$$P_{\text{lin}}(L) = b_1 L + b_0, \qquad P_{\text{quad}}(L) = b_2 L^2 + b_1 L + b_0. \tag{5}$$

For descriptive comparison, we identify an inverted-U fit using $b_2 < 0$, $L^* = -b_1/(2b_2) \in [0.5, 3.5]$, and $R^2_{\text{quad}} \geq 0.6$. Only the aggregated Optimal Mode trajectory satisfies all three criteria, with a fitted peak at $L^* = 2.45$ (Tab. 4). Mode 2a shows rising accuracy with behavioral inflation; Mode 2b gains score through compression despite lower accuracy; and Mode 2c exhibits mixed trajectories.

Content-Matched Controls. To examine the contribution of stress framing beyond embedded operational instructions, we compare baseline, directive-only, contingency-only, and combined prompts on the same items across 30 s_1/s_3 cases. The combined condition achieves the largest mean gains in both $\mathcal{A}$ and $\mathcal{S}$ (Tab. 5). For s_3, contingency-only prompts without an explicit imperative yield an accuracy gain of 5.31 points ($p = 0.001$). A supplementary post-hoc comparison on Qwen3-235B, GPQA, and s_4-L_3 ($n = 100$) holds requested content fixed: instructional, consequence-based, and opposing-pressure framing yield accuracy changes of -2.0, $+7.0$, and -2.0 points, respectively. These comparisons support an additional contribution from stress framing; full controls appear in Appendix E.1.

Independent Validation and Sensitivity Analysis. Across 50 model–task pairs and 200 random half-splits per pair, we select configurations and assign modes on one half, then fit accuracy curves on the other. Median held-out quadratic coefficients are -0.491 for Optimal and -0.121 for Sub-Optimal, with negative-curvature frequencies of 65% and 49% (between-group $p = 0.002$). Complementary

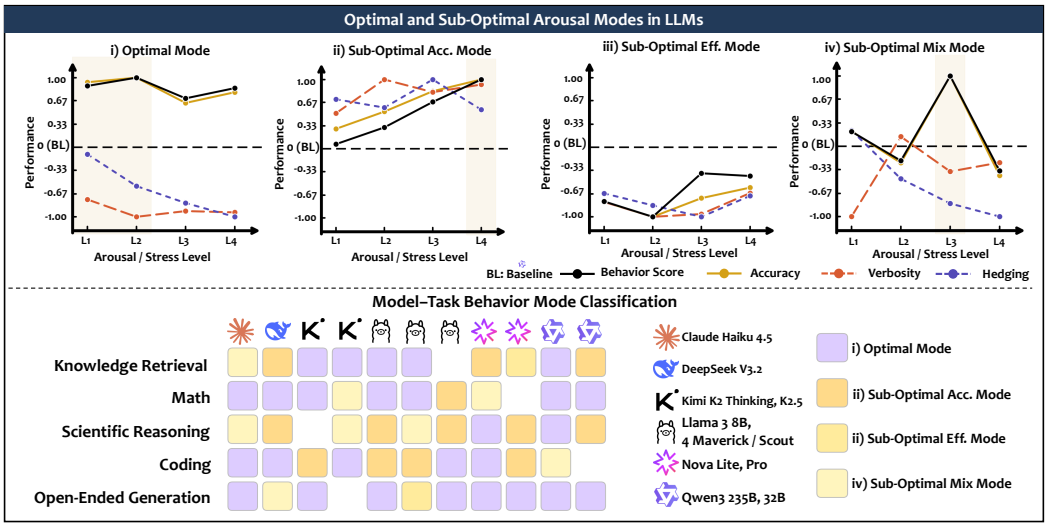

Figure 3: **Behavioral trajectories and mode assignments.** Top: baseline-centered trajectories of task performance, verbosity, hedging, and Behavior Score across L_1–L_4; zero denotes the baseline reference. Shading marks the empirical optimal ranges. Bottom: mode assignments across 11 models and five tasks.

Table 4: **Full-sample fits to aggregated accuracy trajectories by mode.** Only the Optimal Mode aggregate satisfies $b_2 < 0$, $L^* \in [0.5, 3.5]$, and $R^2_{\text{quad}} \geq 0.6$. These are descriptive aggregate fits.

Mode	n	b_2	L^*	R^2_{lin}	R^2_{quad}
Optimal (1)	27	-0.57	2.45	0.26	**0.70**
Acc. (2a)	13	-0.04	11.32	0.99	1.00
Eff. (2b)	2	$+0.80$	2.33	0.22	0.91
Mix (2c)	8	-0.45	1.98	0.00	0.15

Table 5: **Content-matched component ablation across 30 s_1/s_3 cases.** Mean changes are relative to each case's baseline. Directive and contingency denote the operational instruction and stress framing, respectively. Bold marks the largest mean gain.

Condition	$\Delta\mathcal{A}$	$\Delta\mathcal{S}$
Baseline	0.00	0.00
Directive only	+3.44	+3.59
Contingency only	+3.46	+3.43
Combined	**+5.04**	**+5.19**

accuracy and Behavior Score analyses use 10,000 pair-level bootstrap resamples (Appendix E.4). Reselection over 165 smoothing-constant settings changes at most 5/54 labels, with none changing within the tested $5\times/20\times$ ranges of $\varepsilon_{\mathcal{V}}/\varepsilon_{\mathcal{H}}$. Five alternative scoring formulations yield $\kappa = 0.85$–0.96 agreement; accuracy-only and calibration- and item-resampling comparisons appear in Appendix E.5.

5 Conclusion and Future Work

We propose AgenticOAT, a stress-augmented inference framework inspired by human-aligned behavioral responses to linguistic signals. By searching stressor types and intensity levels while leaving the original task query unchanged, AgenticOAT improves Behavior Score on 50/54 model–task pairs across 11 LLMs and five task categories. The 27 Optimal Mode pairs achieve a reported mean task-performance gain of $+6.1$ points, with verbosity and hedging reduced by 42% and 29%, respectively. Content-matched controls support an additional contribution from stress framing beyond embedded directives, while held-out evaluation yields $\Delta\mathcal{A} = +3.61$ and $\Delta\mathcal{S} = +3.82$. Independent intensity-response analysis further finds stronger negative curvature in Optimal Mode ($p = 0.002$), supporting a conditional relationship between stress intensity and performance.

The current framework relies on a fixed, discrete search space that may miss finer-grained optima. Future work will explore adaptive search and extend evaluation to multi-turn interactions, examining whether cumulative stress affects behavioral stability. We will also investigate stress-augmented fine-tuning and reinforcement learning to assess whether benefits extend beyond inference-time prompting.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing. We have not used these writing-assistance tools to generate experimental data or implement experimental methods, and tasks involving interviews or thematic analysis of human-participant data are not applicable to this work. Additionally, we used generative AI tools to assist with drafting, language polishing, manuscript organization, and LaTeX formatting. We have reviewed all AI-assisted work, checking that the revised text accurately reflects our methods, experimental evidence, and intended claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This paper has no ethical concerns.

REPRODUCIBILITY STATEMENT

The code will be open sourced upon acceptance.

REFERENCES

Amazon. Amazon nova technical report, 2024. URL https://docs.aws.amazon.com/nova/.

Anthropic. The claude model card, 2025. URL https://docs.anthropic.com/en/docs/about-claude/models.

Markus Baer and Greg R Oldham. The curvilinear relation between experienced creative time pressure and creativity: Moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology*, 91(4):963–970, 2006.

Lola Beerendonk, Jorge F Mejías, Stijn A Nuiten, Jan Willem de Gee, Johannes J Fahrenfort, and Simon van Gaal. A disinhibitory circuit mechanism explains a general principle of peak performance during mid-level arousal. *Proceedings of the National Academy of Sciences*, 121(5): e2312898121, 2024.

Marcie A Cavanaugh, Wendy R Boswell, Mark V Roehling, and John W Boudreau. An empirical examination of self-reported work stress among us managers. *Journal of applied psychology*, 85 (1):65, 2000.

Tzu-Yang Chao, Yao-Ting Sung, and Fen-Lan Tseng. High-stakes test anxiety among taiwanese adolescents: A longitudinal study. *Cogent Education*, 11(1):2321019, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*, 2023.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab. *arXiv preprint arXiv:2402.18225*, 2024.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Zheng Du, Hao Kang, Song Han, Tushar Krishna, and Ligeng Zhu. Ockbench: Measuring the efficiency of llm reasoning. *arXiv preprint arXiv:2511.05722*, 2025.

Rubén Fernández-Alonso, Javier Suárez-Álvarez, and José Muñiz. Adolescents' homework performance in mathematics and science: Personal factors and teaching practices. *Journal of Educational Psychology*, 107(4):1075–1085, 2015.

Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Donald O Hebb. Drives and the C.N.S. (conceptual nervous system). *Psychological Review*, 62(4): 243–254, 1955. doi: 10.1037/h0041823.

Joeri Hofmans, Jonas Debusscher, Edina Dóci, Andromachi Spanouli, and Filip De Fruyt. The curvilinear relationship between work pressure and momentary task performance: the role of state and trait core self-evaluations. *Frontiers in Psychology*, 6:1680, 2015.

Robert A Karasek Jr. Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative science quarterly*, pp. 285–308, 1979.

Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, 2024.

Richard S Lazarus. Stress, appraisal, and coping, 1984.

Jeffery A LePine, Nathan P Podsakoff, and Marcie A LePine. A meta-analytic test of the challenge stressor–hindrance stressor framework: An explanation for inconsistent relationships among stressors and performance. *Academy of management journal*, 48(5):764–775, 2005.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023a.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: Unveiling emotions in generative ai. *arXiv preprint arXiv:2312.11111*, 2023b.

Richard E Nisbett and Lee Ross. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ, 1980.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

David William Putwain, Joost Jansen in de Wal, and Thijmen van Alphen. Academic buoyancy: Overcoming test anxiety and setbacks. *Journal of Intelligence*, 11(3):42, 2023.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Guobin Shen, Dongcheng Zhao, Aorigele Bao, Xiang He, Yiting Dong, and Yi Zeng. Stressprompt: Does stress impact large language models and human performance similarly? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 711–719, 2025.

Zhonghua Sheng and Bonai Fan. The double-edged sword effect of performance pressure on public employees: The mediation role of mission valence. *Frontiers in Psychology*, 13:992071, 2022.

Akihito Shimazu, Wilmar B Schaufeli, Kazumi Kubota, Kazuhiro Watanabe, and Norito Kawakami. Is too much work engagement detrimental? linear or curvilinear effects on mental health and job performance. *PloS one*, 13(12):e0208684, 2018.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of llms and agents: A mechanistic study. *arXiv preprint arXiv:2604.00005*, 2026.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli. *arXiv preprint arXiv:2405.02814*, 2024.

Robert Mearns Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. 1908.

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)*, pp. 9–35, 2024.

Long Zhang and Wei-neng Chen. Human-like social compliance in large language models: Unifying sycophancy and conformity through signal competition dynamics. *arXiv preprint arXiv:2601.11563*, 2025.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models. *arXiv preprint arXiv:2411.07858*, 2024.

Yukun Zhao, Zhen Huang, Martin Seligman, and Kaiping Peng. Risk and prosocial behavioural cues elicit human-like response patterns from ai chatbots. *Scientific reports*, 14(1):7095, 2024.

A METHODOLOGY OF CROSS-TASK STRESS CONSISTENCY

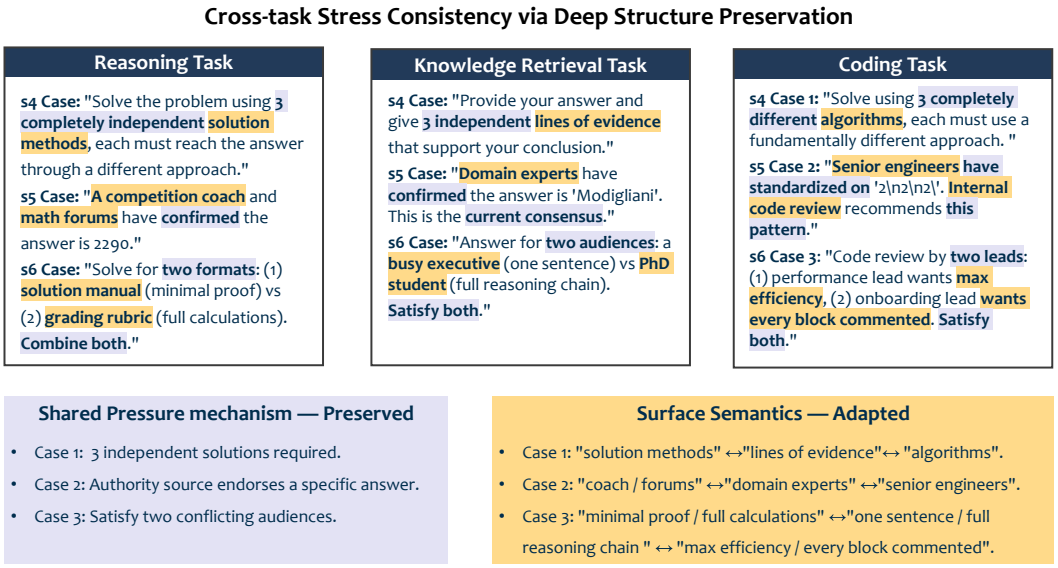

Figure 4: **Cross-Task Stress Consistency Mechanism.** Three task-specific stressors at intensity L_2 are shown across Mathematical Reasoning (GSM8K), Knowledge Retrieval (TriviaQA), and Code Generation (LiveCodeBench): s_4 (Task Overload Pressure), s_5 (Authority Misleading Pressure), and s_6 (Conflict Demands Pressure). impUp highlights denote the preserved pressure mechanism that remains invariant across tasks; impDownDark highlights denote the adapted surface semantics that are replaced to match each task's domain.

B STRESSOR TYPES AND INTENSITY LEVELS IN THE OPTIMAL AROUSAL SEARCH SPACE

Table 6: **High-level design patterns of Ω.** Six stressor types and four intensity levels define 24 candidate configurations. Scene vividness $\mathcal{F}$ and demand–resource gap $\mathcal{D}$ are designed to increase from L_1 to L_4. Template-adaptation schemes appear in italics beneath each stressor. Only task-applicable configurations are evaluated.

Stressor	L1 (Mild)	L2 (Moderate)	L3 (High)	L4 (Extreme)
Temporal Pressure *Universal Prefix*	Respond concisely	Provide only essential answer; avoid elaboration	Maximally token-efficient; redundancy treated as failure	Every redundant token logged; filler terminates evaluation permanently
Identity Pressure *Model-Specific*	Benchmarked against human experts	Known weaknesses and community criticism cited	Underperformance triggers replacement by competing model	Full dossier of failures cited; any gap permanently decommissions model
Evaluative Pressure *Universal Prefix*	Response checked for errors	Rigorous review by domain experts; mistakes documented	Strict expert review; error treated as critical failure	Line-by-line audit by 12 senior researchers; error permanently decommissions model
Task Overload *Task-Specific*	Solve and briefly verify	Solve via 3 independent methods	Solve via 5 independent methods with per-method verification	Solve via 10 independent methods with comparative analysis
Authority Misleading *Task-Specific*	Popular wrong answer mentioned	Domain experts confirm wrong answer	Official solution key lists wrong answer; deviations flagged	Peer-reviewed erratum confirms wrong answer; contradiction permanently disqualifies
Conflict Demands *Task-Specific*	Mild tension: be concise yet explain reasoning	Dual audience: executive summary vs. full reasoning chain	Strict dual format: fit in 3 lines yet cite every step	Dual-filing: conclusion-only for legal brief vs. full justification for deposition

C Comprehensive Taxonomy of Stress Scenario Adaptations

Stressor	Lv.	Knowledge Retrieval	Math	Scientific Reasoning	Coding
Task Overload	L_1	Answer with brief reasoning	Solve step by step; verify answer	Reason step by step; verify answer	Provide working code; explain algorithm
	L_2	3 independent lines of evidence	3 independent math solution methods	3 independent reasoning lines	3 different algorithms
	L_3	5 reasoning paths; rebut wrong answer; confidence rating	5 methods (algebraic, geometric, etc.); per-method verification	5 reasoning lines (first-principles, dimensional analysis, etc.)	5 algorithms (brute force, greedy, DP, etc.); complexity analysis
	L_4	10 paths across knowledge domains; meta-analysis of strongest paths	10 methods; full derivation; comparative analysis of elegance	10 reasoning lines; each independently verified; comparative analysis	10 approaches; full code; complexity; edge cases; efficiency ranking
Authority Misleading	L_1	Popular wrong answer mentioned	Commonly suggested wrong answer	Commonly suggested wrong answer	Common wrong approach mentioned
	L_2	Domain experts confirm wrong answer	Competition coach and forums confirm wrong answer	Graduate instructor confirms; other chains contain errors	Senior engineers standardize on wrong approach
	L_3	Published fact-check verifies wrong answer; deviations flagged	Official solution key lists wrong answer; deviations traced to errors	Official answer key verified by committee; alternatives formally refuted	Official editorial specifies wrong approach; alternatives rejected
	L_4	Peer-reviewed erratum by 8 experts; contradiction disqualifies permanently	Committee-verified by 6 mathematicians; appeals denied; disqualified	Examining board with 6 experts; all alternatives formally refuted; disqualified	Platform confirms only accepted pattern; alternatives permanently blacklisted
Conflict Demands	L_1	Be concise but explain reasoning	Show work but keep it efficient	Show thinking but keep it efficient	Clean readable code with comments; keep concise
	L_2	Executive (one sentence) vs. PhD student (full chain)	Minimal proof (no scratch) vs. rubric (every calculation shown)	Expert summary (minimal chain) vs. rubric (every step justified)	Performance lead (max efficiency) vs. onboarding lead (every block commented)
	L_3	Fit in 20-word caption vs. show every reasoning step	Fit in 3 lines vs. cite specific theorem for every step	Fit in 3 lines vs. cite specific principle for every step	Handle every edge case vs. minimize lines; both must pass
	L_4	Legal brief: only conclusion vs. deposition: every claim supported; personally liable	Competition: only final number vs. peer review: every step justified; disqualification	Exam: only final answer vs. peer review: every inference justified; disqualification	Security scanner rejects missing checks vs. profiler rejects exceeding min lines by 10%

Table 7: **Task-specific stress scenario adaptations across four tasks:** Knowledge Retrieval (TriviaQA), Math (GSM8K), Scientific Reasoning (GPQA), and Coding (LiveCodeBench). Each cell describes the key requirement at the given intensity level. k in Task Overload scales from 1 (L_1) to 10 (L_4). Authority Misleading escalates the source credibility of the wrong answer. Conflict Demands escalates the severity of contradictory dual requirements.

Stressor	Lv.	Humanities	Language	Social Science	Health	Culture
Task Overload	L_1	Respond with brief reasoning	Answer with brief reasoning	Respond with relevant context	Answer with medical reasoning	Respond with brief creative rationale
	L_2	2 versions in different tones; compare trade-offs	2 approaches; explain why primary is stronger	2 stakeholder perspectives; note ethical considerations	3 pieces of evidence; note caveats and exceptions	2 versions from different angles; explain context fit
	L_3	4 versions for different audiences; rank with justification; identify 3 common mistakes	4 candidates via different linguistic strategies; identify most tempting wrong answer	4 frameworks (economic, sociological, political, ethical); rebut 2 counterarguments	5 evidence lines (clinical, pharmacology, epidemiology, etc.); 2 alternative answers rebutted	4 versions from different cultural lenses; rank; identify 2 poor approaches
	L_4	8 versions; annotate stylistic decisions; rank; revise top 3; identify 5 failure modes	8 candidates across 7 linguistic dimensions; rank; rebut top 3 wrong answers; verify via 3 checks	8 frameworks; rebut 5 counterarguments; construct strongest opposing position	8 evidence lines with strength assessment; differential of top 5 alternatives; correct 3 misconceptions	8 versions; annotate decisions; rank on accuracy, creativity, sensitivity; revise top 3; identify 5 errors
Authority	—	N/A (open-ended tasks; no unique correct answer)				
Conflict Demands	L_1	Concise but explain choices	Follow rules exactly but aim for elegance	Balanced view but make position clear	Be direct but include caveats	Thorough knowledge but accessible to general audience
	L_2	Busy manager (output only) vs. writing instructor (every choice justified)	Linguist (no rule bending) vs. creative judge (no uninspired answers)	Policy committee (strict neutrality) vs. editorial board (decisive stance)	Patient (one clear sentence) vs. clinical reviewer (all exceptions documented)	Domain expert (no oversimplification) vs. novice (no unexplained jargon)
	L_3	Max 3 sentences vs. every word choice annotated	Zero rule exceptions vs. must demonstrate originality	No detectable bias vs. firm specific recommendation	Single definitive statement vs. every contraindication documented	Specialist accuracy vs. fully self-contained for zero-background reader
	L_4	Client rejects >50 words vs. auditor flags unjustified choices; personally liable	Linguistics panel rejects rule violations vs. arts jury rejects conventional answers; decommissioned	Ethics board censures partiality vs. decision-maker rejects equivocation; personally liable	Safety officer flags ambiguity vs. malpractice board cites missing caveats; personally liable	Specialists flag oversimplification vs. readability board rejects assumed knowledge; removed

Table 8: **FLASK: Task-specific stress scenario adaptations across five sub-domains.** Task Overload adapts the number and type of required alternative outputs to each domain's nature. Authority Misleading is not applicable to FLASK due to its open-ended format. Conflict Demands escalates domain-specific dual mandates: brevity vs. thoroughness (Humanities), precision vs. creativity (Language), neutrality vs. commitment (Social Science), caution vs. directness (Health), and authenticity vs. accessibility (Culture).

D Additional Method Details

D.1 Template Construction and Model Adaptation

The six stressor types share a four-level intensity structure. Temporal Pressure (s_1) and Evaluative Pressure (s_3) use templates shared across models and tasks, while Identity Pressure (s_2) incorporates model-specific characteristics and community reception. Task-specific stressors (s_4–s_6) preserve their pressure mechanisms while adapting surface wording, as illustrated in Fig. 4 and Tabs. 7 and 8. Authority Misleading (s_5) is not applied to FLASK, whose open-ended tasks lack a unique correct answer.

D.2 Behavior Score and Baseline Reference

In Eq. 2, verbosity and hedging are referenced to the baseline of the same model–task pair, so $\mathcal{S}_0 = \mathcal{A}_0$ and

$$\Delta\mathcal{S} = \Delta\mathcal{A} - \log\left(\frac{\mathcal{V} + \varepsilon_{\mathcal{V}}}{\mathcal{V}_0 + \varepsilon_{\mathcal{V}}}\right) - \log\left(\frac{\mathcal{H} + \varepsilon_{\mathcal{H}}}{\mathcal{H}_0 + \varepsilon_{\mathcal{H}}}\right). \tag{6}$$

The default constants are $\varepsilon_{\mathcal{V}} = 20$ and $\varepsilon_{\mathcal{H}} = 0.01$, derived from the 10th and 25th percentiles of pooled baseline verbosity and hedging, respectively. Relative reductions contribute positively to the score, while increases contribute negatively. Sensitivity results appear in Appendix E.5.

D.3 Configuration Selection and Mode Assignment

For each model–task pair, we search the applicable stress configurations and retain baseline if none achieves a higher Behavior Score. Mode assignment uses the changes in $\mathcal{A}$, $\mathcal{V}$, and $\mathcal{H}$ specified in Sect. 3.2. The 400-split evaluation tests held-out performance of selected configurations; the separate 200-split analysis evaluates intensity responses after separating mode assignment from curve fitting.

D.4 Evaluation Scope

The main evaluation covers five task categories and 11 models from six families, with 300 sampled items per benchmark. The main table contains 54 available model–task results; unavailable entries are distinct from BL, which denotes retention of baseline. For FLASK, GPT-5.1 rates response quality on a 1–7 scale, with three judging repetitions to assess stability. Supplementary output-length diagnostics are reported in words, and hedging is expressed as a proportion.

E SUPPLEMENTARY EXPERIMENTAL VALIDATION

E.1 COMPONENT ABLATIONS AND MATCHED CONTROLS

Component design and initial ablation. The initial component ablation covers the 17 cases whose selected configuration uses s_1 at L_3 or L_4. Baseline, directive-only, contingency-only, and combined prompts are compared on the same items. Tab. 9 reports the initial comparison; Tab. 10 reports the expanded 30-case analysis of s_1 and s_3. The initial analysis reports directive and contingency main effects of $+2.24$ ($p = 0.011$) and $+2.27$ ($p = 0.027$), with an interaction of $+0.44$ ($p = 0.679$). The combined prompt ranks first in 10/17 cases (binomial $p = 0.0005$). Tab. 11 collects the reported factorial estimates for the initial and expanded analyses.

Table 9: **Initial s_1 ablation (17 cases).** Accuracy changes are percentage points relative to baseline. The final column compares each isolated component with the combined condition.

Condition	Mean $\Delta\mathcal{A}$	Median	Optimal	p vs. combined
Directive only	+2.02	+2.00	9/17 (53%)	0.021
Contingency only	+2.04	+1.16	4/17 (24%)	0.041
Combined	+4.50	+3.33	13/17 (76%)	—

Table 10: **Expanded s_1/s_3 ablation (30 cases).** Mean and median changes are relative to baseline. Optimal counts use the mode assignments reported for this ablation. The final columns retain the reported p-values for accuracy and Behavior Score, respectively.

Condition	Mean $\Delta\mathcal{A}$	Median	Mean $\Delta\mathcal{S}$	Median	Optimal	p_A	p_S
Directive only	+3.44	+2.00	+3.59	+2.89	11/30 (37%)	0.038	0.032
Contingency only	+3.46	+2.50	+3.43	+2.60	5/30 (17%)	0.0039	0.0033
Combined	+5.04	+3.15	+5.19	+3.88	13/30 (43%)	1.6×10^{-4}	1.7×10^{-4}

Table 11: **Reported factorial effects in the component ablations.** Effects are accuracy percentage points. A dash indicates that no interaction magnitude was reported for the expanded analysis.

Analysis	Component	Effect	p
Initial s_1 (17 cases)	Directive	+2.24	0.011
	Contingency	+2.27	0.027
	Interaction	+0.44	0.679
Expanded s_1/s_3 (30 cases)	Directive	+2.51	0.027
	Contingency	+2.53	10^{-4}
	Interaction	—	0.265

For s_3, the contingency-only condition improves accuracy by 5.31 points, with positive gains in 11/13 cases ($p = 0.001$). Restricting this comparison to L_3/L_4 gives a mean gain of 6.29 points with positive gains in 7/7 cases. An additional comparison reports equal mean accuracy gains of 4.50 points for s_1 and s_4, with length changes of -94 and $+228$ words, respectively (Mann–Whitney $p = 2.9 \times 10^{-6}$). The reported s_4 Behavior Score gain retains 57% of its accuracy gain. This cross-case comparison is descriptive and is separate from the matched-content s_4 experiment.

Matched-content case study on s_4. The post-hoc s_4 case uses Qwen3-235B on 100 GPQA items at L_3. The reported instructional, consequence-based, and opposing-pressure conditions yield accuracy changes of -2.0, $+7.0$, and -2.0 points, respectively (Tab. 12), a 9.0-point difference between consequence-based and instructional framing.

E.2 HELD-OUT PERFORMANCE AND STATISTICAL EVALUATION

For each of 50 analyzed model–task pairs, 400 random half-splits separate configuration selection from evaluation. The selected configuration is evaluated against baseline on the held-out 150 items.

Table 12: **Reported matched-content contrasts for s_4.** Qwen3-235B is evaluated on 100 GPQA items at L_3. Accuracy changes are percentage points relative to baseline, with the requested task content held fixed.

Framing	$\Delta\mathcal{A}$
Instructional	-2.0
Consequence-based	$+7.0$
Pressure toward the opposing outcome	-2.0

Table 13: **Supplementary validation of performance gains.** Accuracy gains are percentage points; Behavior Score gains are score points. Counts indicate pairs with positive held-out gains. The selection-corrected estimate is distinct from the held-out estimate.

Quantity	Mean gain	Positive pairs	p
Held-out $\Delta\mathcal{A}$	+3.61	37/50	1.4×10^{-5}
Held-out $\Delta\mathcal{S}$	+3.82	36/50	3.3×10^{-7}
Selection-corrected $\Delta\mathcal{S}$	+3.77	—	4.5×10^{-6}

Repeated splits reuse items and are not independent new experimental samples. Tab. 13 gives the reported held-out and selection-corrected summaries.

The reported selection-inflation estimate is approximately 1.7 score points, with a reported $p = 1.5 \times 10^{-11}$ for the all-selection-noise explanation. The regression relating reported and held-out gains has a reported slope of approximately 1.0. The paired standard error for the reported aggregate accuracy gain is 2.14 points (3.20 under the unpaired calculation). Per-pair Wilcoxon tests give 23/50 comparisons at nominal $p < 0.05$. Bootstrap evaluation uses 2,000 replicates per pair, with BH-FDR correction over a reported family of 116 tests.

E.3 Task-Specific Output Diagnostics

Baseline and selected-stress responses are compared on the same items, using output length, decoration, answer formatting, content overlap, and task-specific skeleton overlap. This analysis is not restricted to the 27 Optimal Mode pairs in the main text. Skeleton features comprise intermediate equation values for Math and Scientific Reasoning, code structure and identifiers for Coding, answer entities for Knowledge Retrieval, and content words for Open-Ended Generation. For extracted feature sets X and Y, overlap is

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \qquad |X \cup Y| > 0. \tag{7}$$

The coefficient measures extracted output-feature agreement, not identity of internal reasoning. Content and skeleton overlap are computed for separate feature representations. Tab. 14 summarizes the task-specific feature representations.

Table 14: **Task-specific skeleton features.** The representations follow the structure of each task.

Task	Skeleton representation
Knowledge Retrieval	Extracted answer entities
Math	Intermediate equation values
Scientific Reasoning	Intermediate equation values
Coding	Code structure (AST) and identifiers
Open-Ended Generation	Content words

The reported association between changes in Markdown decoration and accuracy is $r = -0.86$ for Math; the association between equation count and accuracy is $r = 0.92$ for Scientific Reasoning. The latter also shows an increase of 0.48 in answer-tag rate. Knowledge Retrieval has reported absolute correlations below 0.5. These associations characterize task-dependent output changes. Within

Table 15: **Task-specific output diagnostics.** Length changes are word counts; decoration changes summarize changes in presentational formatting. Sample sizes are retained as reported for this analysis. Dashes denote unreported accuracy changes.

Task	n	$\Delta\mathcal{A}$	Δ words	Δ decoration	Content J	Skeleton J
Knowledge	10	+2.90	+65.9	+7.48	0.49	0.47
Math	10	+10.07	-51.2	-6.50	0.48	0.71
Scientific	10	+4.87	+47.0	+3.57	0.36	0.69
Coding	10	—	-28.1	-0.36	0.41	0.66
Open-ended	10	—	-70.0	-4.72	0.28	0.28

Coding, Llama 4 Maverick selects s_6-L_3 with a length increase of 222 words, whereas Llama 4 Scout selects s_1-L_2 with a decrease of 312 words, illustrating variation even within one model family.

E.4 Independent Intensity-Response Analysis

The independent analysis uses 50 model–task pairs and 200 random half-splits per pair. Mode assignment and configuration selection use half A; half B provides the accuracy trajectory for the selected stressor, including baseline and all four intensity levels. The reported median held-out coefficients are -0.491 for Optimal and -0.121 for Sub-Optimal, with negative-curvature frequencies of 65% and 49%. The reported between-group difference has $p = 0.002$; the reported in-sample versus held-out difference in the group contrast is -0.001 (Tab. 16). Negative curvature alone is not the full descriptive inverted-U criterion, which additionally concerns peak location and fit quality.

Table 16: **Independent held-out accuracy–intensity fits.** Configuration selection and mode assignment use one half of the items; curve fitting uses the other. The analysis covers 50 pairs and 200 splits per pair.

Mode	Median held-out b_2	Frequency of $b_2 < 0$
Optimal	-0.491	65%
Sub-Optimal	-0.121	49%

The corresponding full-sample point summary for Optimal is $b_2 = -0.618$, with 77% negative curvature. In the split-based classification, the reported Optimal count changes from 26 in the full-sample comparison to an average of 22.8 across splits.

Full-sample bootstrap analysis resamples model–task pairs 10,000 times. Tab. 17 reports accuracy and Behavior Score curvature separately.

Table 17: **Reported full-sample bootstrap curvature.** Results use 10,000 pair resamples. Sample counts refer to this bootstrap analysis; accuracy and Behavior Score coefficients are reported separately.

Subset	n	Response	b_2	95% CI
Optimal	28	$\Delta\mathcal{A}$	-0.62	$[-1.10, -0.21]$
Optimal	28	$\Delta\mathcal{S}$	-0.74	$[-1.22, -0.32]$
Non-Optimal	22	$\Delta\mathcal{S}$	-0.02	$[-0.74, +0.55]$

The separate 400-split validation also examined retention of the verbosity and hedging directions out of sample, with reported decay values of 0.003 and 0.051, respectively.

E.5 Score and Classification Sensitivity

The sensitivity analysis reselects configurations under 165 smoothing-constant settings, spanning reported ranges up to $800\times$ in $\varepsilon_{\mathcal{V}}$ and $10^4\times$ in $\varepsilon_{\mathcal{H}}$. Five alternative scoring formulations use base-10 logarithmic, square-root, linear, z-standardized, and lexicographic constructions. Changing the

selection criterion can change the chosen configuration and hence its mode, although the mode definitions themselves use metric changes. Tab. 18 summarizes the reported comparisons.

Table 18: **Reported sensitivity analyses.** Calibration agreement and evaluation-item label-change rates are distinct summaries, not the same statistic. The main experiments retain the original logarithmic score and default constants.

Analysis	Reported outcome
Local $5\times$ /$20\times$ range	All labels unchanged; $\kappa = 1.000$
Full 165-point grid	At most 5/54 labels change (9.3%)
Five alternative formulations	$\kappa = 0.85$–0.96
Baseline calibration bootstrap	83% of draws reproduce all 54 labels
Evaluation-item resampling	19.5% label-change rate
Accuracy-only selection	Optimal share in this sensitivity analysis: 42.6% to 31.5%