

PREDICTABLE BUT UNSTABLE: PROBING CONSTRAINT VIOLATIONS IN MULTI-TURN DIALOGUE

ABSTRACT

In single-turn settings, linear probes over attention-head activations can predict whether a language model will violate an explicit constraint before generation. We extend this framework to multi-turn dialogue and find that mean AP does not systematically decline, yet probes become notably less stable when rebuilt from different development partitions—a pattern we term *predictable but unstable*. We identify two structural factors that may contribute: the first-violation task induces a survival structure that progressively thins supervision at later turns, even from an initial pool of 2,400 dialogues per configuration, and the relative construction stability of fixed attention-head banks varies across turns, so that a bank relatively stable at one turn may be less stable at another. These factors make probe construction sensitive to the turn composition of a development partition. We propose Frequency-Consensus head selection, which modifies only the aggregation rule while leaving the rest of the probing pipeline unchanged. On three 7B models, Frequency-Consensus improves model-averaged AP by 1.9 points over retaining all candidate heads while reducing cross-run SD by 37.5% and range by 34.8%, with the pattern extending to the evaluated 13–14B models.

1 INTRODUCTION

If a constraint violation can be detected from a model’s internal state before decoding begins, the system can resample or withhold the response, converting a post-hoc quality problem into a real-time gating decision. This prospect depends on whether pre-generation representations carry a reliable violation signal and whether such a signal can be extracted consistently. Probing the internal representations of large language models has revealed signals predictive of truthfulness (Azaria & Mitchell, 2023), factual constraint satisfaction (Yuksekgonul et al., 2024), and instruction-following success (Heo et al., 2025). Such probes have proven effective in single-turn settings. Whether they can be extended to multi-turn dialogue, and whether they can be constructed reliably in that setting, has not been systematically examined.

In multi-turn dialogue, an instructional constraint introduced at the start of a conversation must remain effective across an expanding context that includes the model’s own prior responses. Existing evaluations assess constraint compliance and multi-turn response quality after generation (Jiang et al., 2024; Kwan et al., 2024), but post-generation evaluation differs from pre-generation prediction. Our task is to predict, before each response is generated, whether it will be the first to violate the specified constraint. This first-violation framing imposes a survival-like structure in which later turns draw training evidence only from dialogues that remained compliant up to that point, producing progressively smaller and more selected supervision sets. The prediction problem therefore evolves as the dialogue unfolds, with later turns facing qualitatively different training conditions. Figure 1 summarizes the first-violation prediction task, the diagnostic analyses, and the overall selection pipeline.

Because a violation probe is ordinarily built once and reused, its value depends on how well it performs on average, but also on whether that performance holds when the probe is rebuilt from different development data. We call this second property *construction stability*. Concretely, we measure construction stability by the variability of test AP across repeated constructions of the full pipeline—head selection, probe fitting, and readout combination—on a fixed test set. Class-count-matched comparisons reveal that changes in mean AP vary by task, while cross-run variability increases systematically across all evaluated constraint families (Figure 2; Section 4.2). Multi-turn probes thus retain predictive ability but become less stable to construct, a pattern we refer to as *predictable but unstable*.

Diagnostic analyses examine this instability along two axes (Section 4.2). On the data side, expanding the fitting and selection windows to include preceding turns yields higher mean AP and lower cross-run SD, when evaluation remains at each condition’s fixed target turn. Supervision becomes sparse at later turns because dialogues exit the risk set after their first violation. On the representation side, the relative construction stability of a fixed head bank varies across dialogue turns, despite unchanged membership. A single development partition can reflect a particular mixture of dialogue turns. We therefore evaluate heads across multiple development partitions. Retaining all candidate heads does not explicitly address this variation across turns.

We therefore use a selection criterion based on repeated top- K inclusion. In each development partition, a candidate head receives one binary vote according to whether it enters the top K . Frequency-Consensus selects the K heads with the highest inclusion frequencies across partitions. Repeated partitioning provides the evaluation evidence, while inclusion frequency defines the aggregation rule.

Our evaluation covers three model families spanning 7B to 14B parameters. Across the three 7B models, Frequency-Consensus improves model-averaged AP and reduces cross-run dispersion relative to retaining all candidate heads, with the same qualitative pattern observed on the evaluated 13–14B models (Section 4.4). Under shared repeated-evaluation evidence, Frequency-Consensus achieves the lowest cross-run SD and range among the three aggregation rules on every 7B model (Section 4.3). The reduction is largest on Llama, where SD drops by 17.3% relative to Mean-AP; on Gemma and Qwen the reductions are smaller but directionally consistent.

Our contributions are threefold. First, we extend pre-generation first-violation probing to multi-turn dialogue and characterize the predictable-but-unstable phenomenon. Second, we investigate this variability through controlled supervision analyses and comparisons of fixed head banks across dialogue turns. Third, we introduce Frequency-Consensus, which requires only a single substitution in the aggregation rule—from AP magnitude to top- K inclusion frequency—and show that this change alone yields measurable improvements in construction stability without sacrificing predictive performance. A detailed discussion of related work is provided in Appendix A.

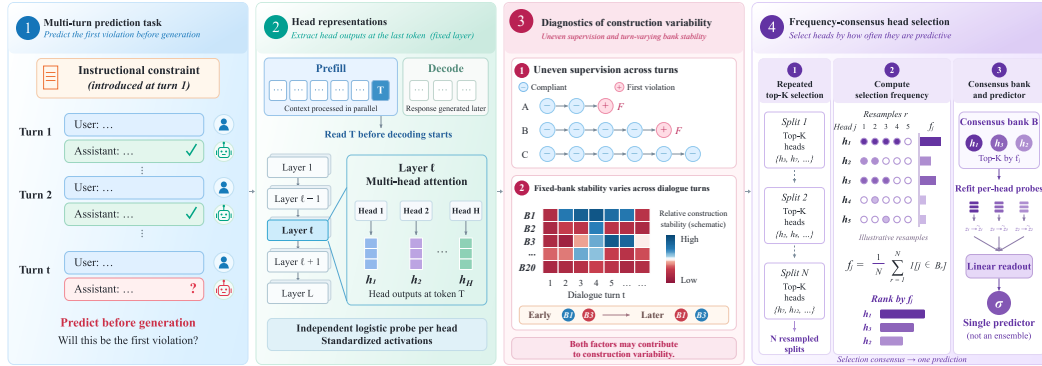


Figure 1: Overview of pre-generation first-violation probing and Frequency-Consensus head selection, illustrating (1) the multi-turn first-violation prediction task, (2) attention-head representation extraction at the last context token, (3) two structural prediction factors—uneven turn-level supervision and turn-varying bank stability—that may contribute to construction variability, and (4) the Frequency-Consensus selection pipeline that aggregates heads by top- K inclusion frequency across repeated partitions. Panels 3 and 4 are schematic.

2 PROBLEM SETTING

2.1 MULTI-TURN FIRST-VIOLATION PREDICTION

We study pre-generation constraint-violation prediction in multi-turn dialogue. Each dialogue is governed by an explicit instructional constraint c introduced at the start. Before the model generates its response at turn t , we predict whether that response will be the first to violate c . From a survival-

analysis perspective, this is a discrete-time hazard prediction task: predicting a first violation at the current turn conditional on no earlier violation (Gensheimer & Narasimhan, 2019).

Let F denote the first violation turn. A dialogue remains in the at-risk set until its first violation: observations before F are labeled $y_t = 0$, and the first violation is labeled $y_F = 1$. Observations after F are excluded because the dialogue has left the at-risk set and no longer contributes to the first-violation prediction task. Dialogues without an observed violation remain in the at-risk set throughout and contribute negative observations at every turn. Turns with indeterminate or truncated outputs are also excluded; details are provided in Appendix B.

2.2 DIALOGUE DATA AND CONSTRAINTS

Our dialogue construction draws on Multi-IF (He et al., 2024). To provide a controlled multi-turn setting, we generate English user-message sequences of up to 14 turns in advance. During rollout, the user-side sequence is held fixed while the model conditions on the full accumulated dialogue, including its own prior responses. The sequences are generated with an LLM; the generation prompt is provided in Appendix C. The target constraint is introduced in the first user message. For each model–constraint combination in the main comparison, we generate 2,400 dialogues with up to 14 turns each.

The main predictor comparison covers seven constraint families drawn from IFEval (Zhou et al., 2023), selected to span lexical, length, structural, and formatting categories: keyword frequency, quotation wrapping, required prefix, sentence limit, title formatting, two responses, and word limit. Compliance is assessed by deterministic rule-based checkers; required prefix follows IFEval’s constrained-start instruction and checker. Table 1 summarizes the constraint definitions and mean first-violation turns across models. The user-sequence generation and constraint-injection protocols appear in Appendix C; exact instructions and verification rules are provided in Appendix D. The supervision diagnostic in Figure 3(a) uses a separate task set that additionally includes the required-ending and three-placeholder constraints; its coverage and protocol are described in Appendix E.5.

For the main multi-turn analysis, dialogues with $F = 1$ are excluded because they contain no subsequent-turn prediction points and thus do not contribute to multi-turn evaluation. The class-count-matched comparison in Section 4.2 uses first-turn data separately. The number of dialogues retained after this filtering varies across model–constraint combinations; filtering counts are reported in Appendix E.3. Experimental partitions and model coverage are specified in Section 4.1.

Table 1: Constraint definitions and mean first-violation turn conditional on an observed violation at $F > 1$.

Task	Constraint	Gemma7B	Llama7B	Llama13B	Qwen7B	Qwen14B
Keyword frequency	Minimum keyword count	2.21	2.96	4.47	2.85	2.68
Quotation wrapping	Quotation-mark wrapping	5.90	4.19	5.73	4.25	4.32
Required prefix	Specified opening phrase	5.44	4.97	5.68	6.44	7.57
Sentence limit	Maximum sentence count	4.62	3.71	4.30	3.22	5.36
Title formatting	Required title format	3.05	4.79	4.82	5.20	6.03
Two responses	Two delimited answers	2.05*	4.20	4.60	3.58	5.02
Word limit	Maximum word count	5.11	3.03	4.49	3.18	5.25

Note. Means exclude dialogues with no observed violation. *For Gemma–Two responses, 2,309 of 2,400 dialogues violate at turn 1, effectively reducing the task to single-turn instruction following. We therefore exclude this model–constraint combination from subsequent experiments.

2.3 ATTENTION-HEAD REPRESENTATION EXTRACTION

The decoder-only Transformer models studied here consist of stacked layers that progressively update token representations through multi-head self-attention and feed-forward computation (Vaswani et al., 2017). Within each layer, attention heads operate in parallel, each producing a context-dependent vector by weighting value vectors at positions allowed by the causal mask. The head outputs are concatenated and linearly projected before being added to the residual stream. We probe individual head outputs before this projection, following Li et al. (2023).

At each prediction point, the frozen language model processes the complete dialogue context in a prefill pass. Let T_i denote the final context-token position for observation i . From a fixed Transformer layer, we extract the output of each candidate attention head j at position T_i , before the attention output projection:

$$\mathbf{h}_{i,j} = \sum_{s \leq T_i} \alpha_{i,j,T_i,s} \mathbf{v}_{i,j,s} \in \mathbb{R}^d,$$

where $\alpha_{i,j,T_i,s}$ is the attention weight and $\mathbf{v}_{i,j,s}$ the value vector. No token from the response being predicted enters the probe. The representation layer and candidate-head counts are specified in Section 4.1.

3 FREQUENCY-CONSENSUS HEAD SELECTION

In this section, we describe the shared probing and readout pipeline, then introduce the frequency-based rule used to select the head bank.

3.1 HEAD PROBING AND LINEAR READOUT

For each candidate head j , we fit an independent logistic-regression probe to its output activation. Activations are standardized feature-wise to zero mean and unit variance using statistics from the probe-fitting subset, yielding $\bar{\mathbf{h}}_{i,j}$. The probe produces the scalar logit

$$z_{i,j} = \mathbf{w}_j^\top \bar{\mathbf{h}}_{i,j} + b_j.$$

Training minimizes mean binary cross-entropy with an ℓ_2 penalty on the weights, an unregularized intercept, and equal observation weights.

Given a selected bank $\mathcal{B}$ of K heads, we freeze their probes and standardize their logits using statistics from the selection subset. A linear readout fitted on that subset produces

$$\hat{p}_i = \sigma \left(\beta_0 + \sum_{j \in \mathcal{B}} \beta_j \tilde{z}_{i,j} \right), \quad (1)$$

where $\tilde{z}_{i,j}$ is the standardized logit and σ is the sigmoid function. The readout uses the same regularized objective, an unregularized intercept, and coefficients unconstrained in sign. It combines K scalar probe logits rather than concatenated head activations. Controlled head-selection comparisons hold this probing and readout pipeline fixed.

3.2 FREQUENCY-CONSENSUS SELECTION

Multi-turn prediction pools observations from dialogue turns that differ in supervision density and in the relative construction stability of fixed head banks (Section 4.2). Multiple development partitions expose this variation, but the key design choice is how the resulting head-level evidence is aggregated. Frequency-Consensus uses top- K inclusion frequency as the aggregation rule.

For each constraint and outer construction run, we form N inner partitions of the full development set—the union of the outer probe-fitting and selection subsets. These inner partitions are used only

to compute head-selection evidence; the held-out test set is never accessed. Each inner partition separates data for probe fitting from held-out data for head evaluation. In partition r , all H candidate probes are fitted and ranked by held-out Average Precision (AP), giving the top- K head set $\mathcal{B}_r$. A head’s inclusion frequency is

$$f_j = \frac{1}{N} \sum_{r=1}^N \mathbb{I}[j \in \mathcal{B}_r]. \quad (2)$$

All three rules use the same repeated partitions and differ only in aggregation. Mean-AP averages AP values, Mean-Rank averages ranks, and Frequency-Consensus retains only binary top- K membership. Each partition therefore contributes one vote per head; AP magnitude and within-top- K rank do not affect the Frequency-Consensus score.

The K heads with the largest f_j form the final bank $\mathcal{B}$. AP-ranking and frequency ties are resolved by head index. All inner fitted parameters are discarded. We refit the selected probes on the outer probe-fitting subset, freeze them, and fit the final readout on the outer selection subset using the procedure in Section 3.1. The resulting predictor is

$$\mathcal{B}_{FC} = \text{TopK}_j(f_j), \quad \hat{p}_i^{FC} = \sigma \left(\beta_0 + \sum_{j \in \mathcal{B}_{FC}} \beta_j \frac{\mathbf{w}_j^\top \bar{\mathbf{h}}_{i,j} + b_j - \mu_j}{s_j} \right). \quad (3)$$

Here $\mathbf{w}_j, b_j$ are the refitted single-head probe parameters; μ_j, s_j are the logit-standardization statistics from the outer selection subset; and β_0, β_j are the fitted readout parameters.

4 EXPERIMENTS AND ANALYSIS

A predictor trained on pooled dialogue turns may depend on which turns and violation cases are represented in its development data. We examine this sensitivity and then test whether aggregating repeated head evaluations by top- K inclusion frequency yields more consistent predictors.

4.1 EXPERIMENTAL SETUP

Our main 7B evaluation uses Gemma-7B-IT (Gemma Team et al., 2024), Llama-2-7B-Chat (Touvron et al., 2023), and Qwen1.5-7B (Qwen Team, 2024), extracting representations from layer 16 with head budgets $K = 6$ for Gemma and $K = 8$ for Llama and Qwen, following evidence that predictive signals concentrate in mid-to-late layers and compact head subsets (Heo et al., 2025; Li et al., 2023; Zhou et al., 2025). Neither choice is tuned per task; sensitivity to K is examined in Section 4.5.

The main predictor comparisons use $R = 20$ outer constructions with a fixed test set. Frequency-Consensus and other aggregation methods additionally use $N = 30$ inner partitions. All methods share the same outer partitions, generated responses, and labels. We report AP on a 0–100 scale, summarized across constructions by mean, SD, range, and Worst₄ (the mean of the four lowest AP values among the 20 construction runs), and then macro-averaged across tasks. Partition ratios, filtering rules, and numerical fitting details appear in Appendix B. The diagnostic analyses in Figure 3 follow separate construction and aggregation protocols detailed in Appendices E.5 and E.6.

4.2 SOURCES OF MULTI-TURN CONSTRUCTION VARIABILITY

We investigate why multi-turn probe construction is less stable than single-turn, using All-Heads predictors in the class-count-matched and supervision analyses to exclude head selection as a confound. Under class-count-matched downsampling, where single-turn and multi-turn settings share equal positive and negative counts within each data role, mean AP changes in both directions across tasks, while cross-run SD increases in all evaluated constraint families by 14.9–53.7% (Figure 2). The additional variability therefore cannot be explained by overall sample size or class balance alone. The matching protocol and matched test counts are provided in Appendix E. The following

analyses examine this variability along two axes: the supervision structure imposed by first-violation truncation, and the relative construction stability of fixed head banks across dialogue turns.

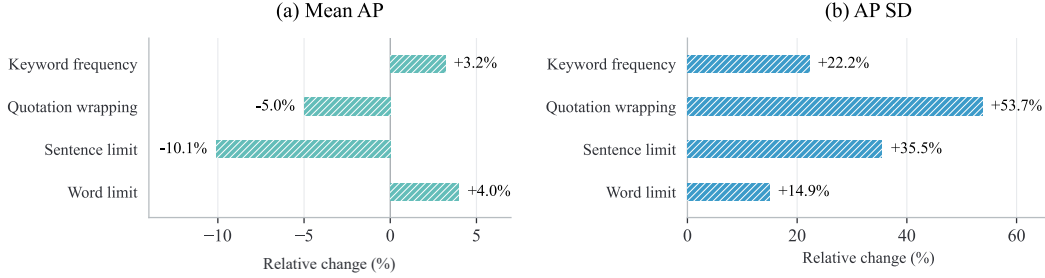


Figure 2: Single-turn to multi-turn changes after matching positive and negative observation counts. (a) Mean AP and (b) cross-run AP SD; values are relative changes in percent. Results are aggregated over the eligible model-constraint configurations within each constraint family.

Training evidence varies structurally across dialogue turns. First-violation prediction induces a survival structure: dialogues exit the risk set upon violating, so later turns draw supervision only from trajectories that remained compliant through all preceding turns. Even with 2,400 dialogues per configuration, this attrition rapidly reduces the supervision available at later turns because trajectories leave the risk set after their first violation.

Two experiments probe this relationship (Figure 3). Under natural supervision, cross-run SD rises toward later target turns. Matching positive and negative counts across turns raises the overall SD level, likely reflecting reduced usable training data after balancing, but substantially flattens the aggregate cross-turn trend (Figure 3a), indicating that uneven turn-level supervision contributes to the increase in variability at later turns. Separately, we progressively add preceding turns to a training window that always includes the target turn, while keeping evaluation fixed at each condition’s target turn. As the window expands, mean AP increases and cross-run SD decreases (Figure 3b). These results show that pooling observations across more turns need not increase variability; broader turn coverage produces higher mean AP and lower cross-run SD. The diagnostic-specific task coverage and construction protocols are provided in Appendices E.5 and E.6.

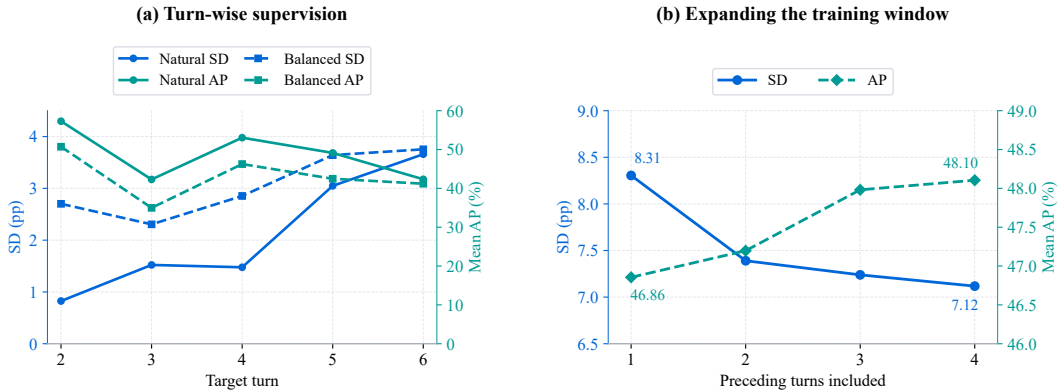


Figure 3: Supervision structure and construction variability. (a) Natural versus count-matched development supervision. (b) Expanding the training window while evaluating at each condition’s fixed target turn.

Fixed head banks vary in relative construction stability across turns. To test whether a relatively stable head bank at one dialogue turn remains relatively stable at others, we evaluate 20 fixed head banks with limited mutual overlap at every eligible turn, keeping membership unchanged. Figure 4 compares each bank’s AP SD with the mean SD of the 20 banks within the same model, constraint, and turn, then averages these differences over available constraints within each model. Positive values indicate lower variability than the bank mean. The observed relative stability of fixed head banks varies across dialogue turns, including within the data-rich region where evaluation support is strongest. A bank that is relatively more stable at one turn can be less stable at another, so

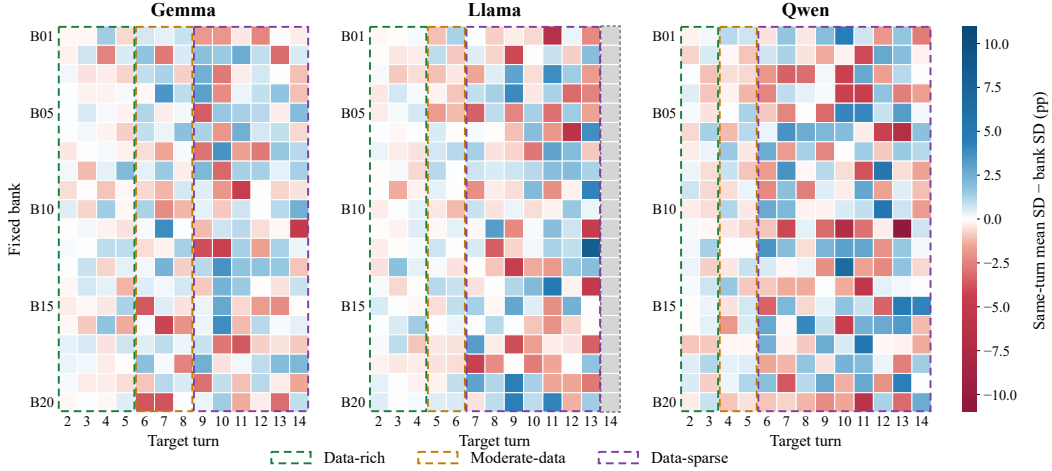


Figure 4: Relative construction stability of fixed head banks across dialogue turns. Positive values indicate lower AP SD than the same-turn bank mean; gray columns are unavailable.

the advantage is turn-dependent rather than a fixed property of the selected heads. The diagnostic protocol is specified in Appendix E.4.

We next compare aggregation rules under shared repeated-evaluation evidence.

4.3 AGGREGATION RULES UNDER SHARED EVALUATION EVIDENCE

Table 2 compares three aggregation rules—Mean-AP, Mean-Rank, and Frequency-Consensus—using the same 30 inner partitions and single-head AP records. Because the evaluation evidence is shared, the comparison isolates the aggregation rule as the varying component.

Table 2: Controlled comparison of aggregation rules using the same resampling evidence.

Model	Method	Mean $\uparrow$	SD $\downarrow$	Relative SD reduction		Worst $_4 \uparrow$
				95% CI (%)	Range $\downarrow$	
Gemma-7B	Mean-AP	50.26	1.03	[−2.05, 12.07]	4.13	48.81
	Mean-Rank	50.22	0.99	[−2.62, 6.18]	3.90	48.91
	Frequency-Consensus	50.23	0.97	—	3.85	48.92
Llama-7B	Mean-AP	77.77	0.98	[1.11, 31.72]	3.77	76.37
	Mean-Rank	77.77	0.97	[2.82, 30.15]	3.82	76.38
	Frequency-Consensus	77.88	0.81	—	3.32	76.70
Qwen1.5-7B	Mean-AP	62.06	1.61	[−3.27, 8.91]	6.00	59.93
	Mean-Rank	62.00	1.61	[−2.98, 6.24]	6.03	59.92
	Frequency-Consensus	62.02	1.55	—	5.84	59.98

Intervals are 95% CIs for the relative SD reduction of Frequency-Consensus versus each baseline; positive values favor Frequency-Consensus.

Across the three models, Frequency-Consensus achieves the lowest SD and range while keeping mean AP within about 0.1 points of the alternatives (Table 2). On Llama, where the difference is largest, it reduces SD by 17.3% and range by 11.9% relative to Mean-AP.

4.4 PREDICTIVE PERFORMANCE AND CONSTRUCTION STABILITY ACROSS MODELS

Table 3(a) compares Frequency-Consensus with seven alternatives on the three 7B models. All-Heads, Random- K , and Single-Split share the same probing pipeline and differ only in head selection; Residual-L2, SAT-L1, and Sparse-AT use alternative representation sources; C-only uses five

pre-generation dialogue features with no model internals. Implementation details and hyperparameter search spaces appear in Appendix B.

Across the three 7B models, Frequency-Consensus improves model-averaged AP by 1.9 points relative to All-Heads while reducing SD by 37.5% and range by 34.8%. Among methods with competitive mean AP, it achieves the lowest SD and highest Worst₄ on each of the three 7B models in point-estimate terms. Table 3(b) extends the comparison to Llama-2-13B and Qwen1.5-14B; the same pattern holds at both scales. Additional baseline results appear in Appendix G.

Table 3: Comparison of violation predictors across model families and scales.

(a) Comparison across 7B models

Model	Metric	Frequency-Consensus	Sparse-AT	Single-Split / ITI	SAT-L1	C-only	All-heads	Residual-L2	Random- K
Gemma-7B	AP $\uparrow$	50.23	48.35	50.03	30.86	27.29	49.69	46.38	47.81
	SD $\downarrow$	0.97	1.77	1.11	0.54	0.58	1.14	1.56	1.95
	Range $\downarrow$	3.85	7.58	4.41	2.08	1.85	4.66	6.22	7.06
	Worst ₄ $\uparrow$	48.92	45.89	48.48	30.09	26.54	48.13	44.23	45.09
Llama-2-7B	AP $\uparrow$	77.88	77.97	77.77	68.80	55.87	75.88	75.95	76.75
	SD $\downarrow$	0.81	1.54	0.94	0.92	0.91	1.48	1.19	1.80
	Range $\downarrow$	3.32	5.64	3.79	3.44	3.39	5.77	4.34	6.55
	Worst ₄ $\uparrow$	76.70	75.75	76.44	67.37	54.76	73.83	74.39	74.10
Qwen1.5-7B	AP $\uparrow$	62.02	62.11	62.11	42.52	40.56	58.97	60.43	59.29
	SD $\downarrow$	1.55	2.14	1.78	1.23	1.60	2.71	1.80	3.05
	Range $\downarrow$	5.84	8.46	7.23	4.93	5.81	9.51	7.13	11.76
	Worst ₄ $\uparrow$	59.98	59.03	59.68	40.71	38.15	55.01	57.95	55.01

(b) Scale extension (13B–14B models)

Model	Metric	Frequency-Consensus	Sparse-AT	Single-Split / ITI	SAT-L1	C-only	All-heads	Residual-L2	Random- K
Llama-2-13B	AP $\uparrow$	65.12	64.97	64.86	40.49	38.82	62.55	63.93	61.78
	SD $\downarrow$	1.05	1.31	1.21	0.81	0.55	1.72	1.23	2.52
	Range $\downarrow$	3.91	5.34	4.68	3.34	2.27	6.27	4.72	9.84
	Worst ₄ $\uparrow$	63.70	63.22	63.18	39.39	38.06	60.27	62.27	58.36
Qwen1.5-14B	AP $\uparrow$	58.63	58.59	57.92	35.23	29.64	55.62	57.13	53.88
	SD $\downarrow$	1.26	1.66	1.66	0.78	0.61	2.26	1.83	3.38
	Range $\downarrow$	5.00	7.12	6.34	3.22	2.18	9.20	6.86	11.62
	Worst ₄ $\uparrow$	56.88	56.40	55.68	34.06	28.85	52.80	54.68	49.12

Bold marks the best value across all methods and, separately, the best value among methods other than SAT-L1 and C-only; ties are included. SD is computed across constructions within each task and then averaged across tasks; it is not the SD of run-wise task-averaged AP. Single-Split adapts the head-ranking procedure of Li et al. (2023) to held-out AP; no inference-time intervention is applied.

4.5 SENSITIVITY TO HEAD BUDGET

Figure 5 varies the number of selected heads while keeping the selection and fitting procedures unchanged. Mean AP rises steeply at small budgets and plateaus beyond 4–6 heads: on Gemma, increasing from two to four heads adds 1.8 AP points, whereas increasing from four to eight adds only 0.3. Except at very small budgets, Frequency-Consensus achieves higher mean AP and lower SD than All-Heads, indicating that the advantage is not specific to the default head budget.

5 DISCUSSION

The supervision structure imposed by first-violation truncation—progressive attrition and filtering of training evidence across dialogue turns—is not specific to probing. Any method that learns from multi-turn trajectories truncated at a first event may face analogous heterogeneity, and the construc-

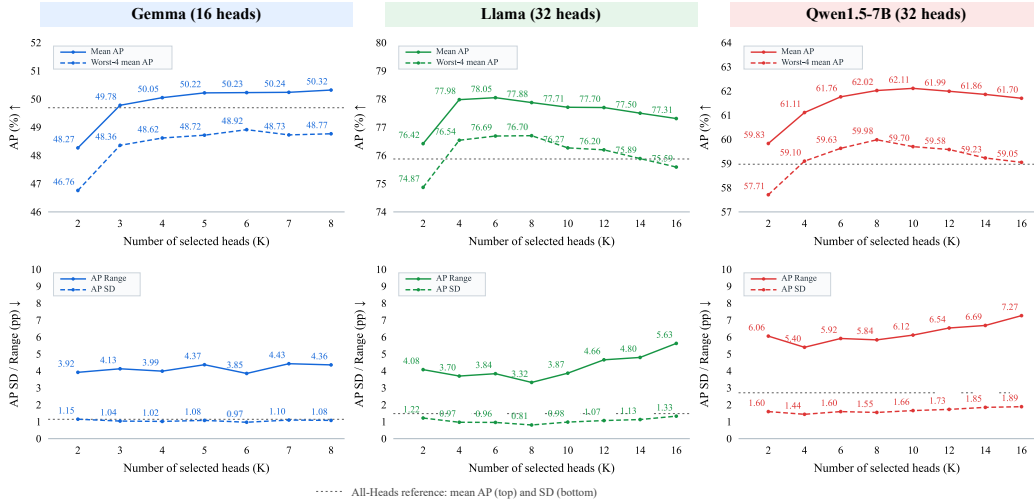


Figure 5: Sensitivity to the number of selected heads. Gray dotted lines mark All-Heads mean AP (top) and SD (bottom). Default K : 6 for Gemma and 8 for Llama and Qwen.

tion variability documented here may manifest in other multi-turn learning and evaluation settings as well.

Our analyses point to two factors associated with this instability: uneven turn-level supervision and turn-varying relative stability of fixed head banks. Frequency-Consensus is motivated by these observations and uses a single substitution—replacing magnitude-based aggregation with binary inclusion votes—over repeated evaluations that standard workflows already produce. The observed improvements are consistent in direction across all evaluated models; where individual confidence intervals do not fully exclude zero, the distribution remains visibly skewed toward positive reductions. Rather than a technique specific to head selection, the underlying principle—aggregate by selection frequency when supervision is structurally heterogeneous—may apply more broadly. The reductions in construction variability observed in the controlled aggregation comparison provide additional, albeit indirect, support for the diagnostic picture developed above.

6 CONCLUSION AND FUTURE DIRECTIONS

We extended pre-generation first-violation probing to multi-turn dialogue and identified a predictable-but-unstable pattern: mean AP does not systematically decline, yet construction variability increases substantially, with our analyses pointing to uneven turn-level supervision and turn-varying relative stability of fixed head banks as contributing factors. A single substitution in the aggregation rule—from score magnitude to inclusion frequency—consistently reduces cross-run dispersion without sacrificing mean AP, with the clearest reductions observed on Llama in the controlled comparison of the three 7B models.

The models evaluated here (Gemma 7B, Llama-2, Qwen1.5) predate several architectural and training advances; validating the predictable-but-unstable pattern and the frequency-based aggregation principle on current model families is a necessary next step. A broader extension concerns the constraint space itself: all tasks in this study admit deterministic verification, and moving to constraints that require semantic judgment would test the framework under label uncertainty that may itself interact with the survival-like supervision structure examined in this work.

REPRODUCIBILITY STATEMENT

Code is available at the anonymous repository <https://anonymous.4open.science/r/probe-artifact-D840/>, which also serves as the access point for available supporting materials. Method configurations and training details are provided in Appendix B, and constraint verification rules are specified in Appendix D.

AI ASSISTANCE

We used AI tools to polish the language of the appendices, assist with document typesetting and layout, and generate the user-message sequences used to construct the evaluation dataset. The user-sequence generation prompt is provided in Appendix C. The authors retain full responsibility for the final manuscript and the use of AI-generated material.

REFERENCES

- Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it’s lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023. URL <https://aclanthology.org/2023.findings-emnlp.68/>.
- Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=gfFVATffPd>.
- Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. Do LLMs “know” internally when they follow instructions? In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=qIN5VDdEOOr>.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024. URL <https://aclanthology.org/2024.acl-long.257/>.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, 2024. URL <https://aclanthology.org/2024.emnlp-main.1124/>.
- Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. doi: 10.7717/peerj.6257.
- Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024. URL <https://arxiv.org/abs/2410.15553>.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL <https://arxiv.org/abs/2311.07911>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL <https://papers.nips.cc/paper/7181-attention-is-all-you-need>.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL <https://doi.org/10.52202/075280-1797>.
- Gemma Team et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. URL <https://arxiv.org/abs/2403.08295>.

- Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Qwen Team. Introducing Qwen1.5. Qwen Blog, February 2024. URL <https://qwenlm.github.io/blog/qwen1.5/>.
- Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=h0Ak8A5yqw>.
- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL <https://doi.org/10.52202/079017-4322>.
- Utsav Maskey, Sumit Yadav, Mark Dras, and Usman Naseem. SafeConstellations: Mitigating over-refusals in LLMs through task-aware representation steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44424–44440, 2026. URL <https://aclanthology.org/2026.acl-long.2056/>.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026. URL <https://openreview.net/pdf/7161cfb10c26ee4c0318bd8d0c4fcb18faf47dda.pdf>.
- Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. URL <https://doi.org/10.1111/j.1467-9868.2010.00740.x>.
- DeepSeek-AI. DeepSeek-V4-Flash: Model card. Hugging Face, 2026. URL <https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash>.

A EXTENDED RELATED WORK

Probing and localization of behavioral signals. Probes over model internals have been used to predict truthfulness (Azaria & Mitchell, 2023), factual constraint satisfaction (Yuksekgonul et al., 2024), and instruction-following success (Heo et al., 2025). At the attention-head level, Li et al. (2023) rank heads by held-out probe accuracy and apply inference-time intervention to improve truthfulness; we adopt this ranking as our Single-Split baseline, replacing probe accuracy with held-out AP for the violation-prediction target. Zhou et al. (2025) use head ablation to show that safety behavior depends disproportionately on a small number of heads. At the residual-stream level, Arditi et al. (2024) find that refusal is mediated by a single direction, and Maskey et al. (2026) show that refusal and non-refusal follow structured, task-dependent trajectories across layers.

Instruction following and multi-turn evaluation. FollowBench (Jiang et al., 2024) evaluates fine-grained constraint compliance with increasing difficulty; MT-Eval (Kwan et al., 2024) pairs multi-turn tasks with matched single-turn queries and attributes degradation to error propagation and distance from relevant content. Laban et al. (2026) report increased output-level unreliability in simulated multi-turn interaction. We study a complementary question at a different level of analysis: whether pre-generation internal-state probes can be constructed consistently across development partitions. The multi-turn setting is also what makes construction stability a practical concern: each conversation contributes multiple correlated prediction points from dialogue turns with unequal supervision, so the composition of a development partition can shift which heads appear most useful.

Stable feature selection. Our selection procedure draws on stability selection (Meinshausen & Bühlmann, 2010), which repeatedly subsamples data, applies a variable-selection method on each subsample, and retains variables by their selection frequency, providing false-selection error control under stated assumptions. Frequency-consensus adopts the same structure—repeated partitions, head ranking by held-out AP, and retention by top- K inclusion frequency—but does not invoke the original error-control guarantees, which our finite, nested multi-turn samples do not satisfy. The probes fitted during inner partitions are discarded, and a single predictor is refitted on the selected heads, so the result is one model rather than an ensemble.

B METHOD CONFIGURATIONS AND TRAINING DETAILS

B.1 MAIN-COMPARISON DATA AND CONSTRUCTION PROTOCOL

Data partitioning. For the three 7B models, representation-based methods use layer 16 (zero-based index 15) and the same conversation-group partitions. A fixed 80%/20% development/test split precedes construction. Each of $R = 20$ outer runs splits the development set into 75% probe-fitting and 25% selection subsets. Frequency-Consensus and alternative aggregation rules share the same $N = 30$ inner partitions of the full development set and held-out single-head AP records. Test observations are excluded from feature standardization, head selection, and hyperparameter selection.

Filtering and features. In the main method comparison, dialogues with $F = 1$ are excluded from every data role. Remaining dialogues retain valid first-turn negatives for fitting; main test evaluation uses $t \geq 2$. The comparison preserves observed class distributions, uses internal model features for representation-based predictors, and includes C-only as a separate context-feature baseline. All repeats use the same pre-generated outputs and labels.

Evaluation metrics. For each task, the 20 AP values are summarized by mean, sample SD (ddof = 1), range (maximum minus minimum), and Worst₄ (mean of the four lowest values), then macro-averaged with equal task weights.

B.2 HEAD-SELECTION AND AGGREGATION RULES

In addition to the three aggregation rules compared in the main text, Appendix F reports Median-AP as an additional variant, alongside the Single-Split baseline.

Single-Split ranks heads by AP on one outer selection subset after fitting probes on the corresponding probe-fitting subset. Random- K samples a subset uniformly; All-heads retains every candidate head. These methods share the single-head probes and linear logit readout in Section 3.1.

Mean-AP scores each head by its arithmetic mean AP over all 30 inner partitions. Mean-Rank uses the average rank (lower preferred). Median-AP uses the median, averaging the 15th and 16th ordered values. Mean-AP and Median-AP use all 30 AP values without trimming. Frequency-Consensus counts top- K membership and ranks by inclusion frequency. AP-ranking and final-score ties are resolved by ascending head index; Frequency-Consensus uses neither a frequency threshold nor an AP-based secondary tie-break.

For all aggregation rules, inner fitted parameters are discarded. Selected probes are refitted on the outer probe-fitting subset, frozen, and combined by a linear readout on the outer selection subset. The readout uses standardized scalar logits, not concatenated activations. Mean-AP and Median-AP share the fitting implementation.

B.3 REGULARIZATION AND PREPROCESSING

For the main predictor comparisons, regularization was selected independently using development data in each of the 20 repeated runs per task. For Random- K , ITI, Frequency-Consensus, and All-heads, the final readout coefficient was selected from $\{10^{-5}, 10^{-4}, \dots, 10^3\}$ by maximizing AP over concatenated out-of-fold predictions from four-fold grouped cross-validation. The final readout was then fitted on the entire selection set. Residual-L2 used the same grid, with fitting on the probe-training set and selection by selection-set AP. SAT-L1 selected $C \in \{0.005, 0.05, 0.5, 5\}$. Sparse-AT jointly selected the neuron budget $B \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$ and regularization $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, with an elastic-net mixing coefficient of 0.5. Test data were not used for hyperparameter selection.

C-only uses five pre-generation features: current turn index, full input length, current user-message length, cumulative length of preceding user messages, and cumulative length of preceding assistant messages. It uses neither the current response nor future information. Its most frequently selected regularization coefficients are included in Table 4.

Single-head probes retain fixed regularization $\lambda_2 = 10^{-3}$. For L2-regularized probes and readouts, λ_2 is the coefficient in mean BCE plus $\lambda_2 \|w\|_2^2/2$, with an unpenalized intercept. For SAT-L1, $C = C_{\text{reg}}$ is the implementation’s inverse regularization parameter, equivalent to mean-loss L1 coefficient $1/(nC)$ for n fitting observations. Sparse-AT uses a prefilter size of $P = 1024$; its neuron budget B and regularization λ correspond to k_{neu} and λ_{EN} below.

Table 4 summarizes the most frequently selected configurations for the three 7B models. Each method has 120 selections for Gemma (6 tasks $\times$ 20 runs) and 140 for Llama and Qwen (7 tasks $\times$ 20 runs).

Table 4: Most frequently selected hyperparameter configurations. Values denote λ unless specified; B is the neuron budget. Parentheses give selection counts across runs.

Method	Gemma	Llama	Qwen
C-only	10^{-5} (42/120)	10^{-3} (74/140)	10^{-5} (61/140)
Random- K	0.01 (47/120)	0.01 (46/140)	0.01 (46/140)
Single-Split / ITI	0.1 (55/120)	1 (50/140)	0.1 (66/140)
Frequency-Consensus	0.01 (49/120)	0.1 (44/140)	0.1 (59/140)
All-heads	0.01 (54/120)	0.01 (58/140)	0.1 (61/140)
Residual-L2	0.1 (91/120)	1 (75/140)	1 (77/140)
SAT-L1	$C = 5$ (86/120)	$C = 5$ (71/140)	$C = 5$ (112/140)
Sparse-AT	$B = 256, \lambda = 0.01$ (38/120)	$B = 128, \lambda = 0.01$ (35/140)	$B = 256, \lambda = 0.01$ (27/140)

Sparse-AT minimizes class-weighted mean BCE plus $\lambda_{\text{EN}}[\alpha \|w\|_1 + (1 - \alpha)\|w\|_2^2/2]$, assigning class- c observations weight $n/(2n_c)$. Candidate neurons are preselected by standardized mean differences using only probe-fitting data. Each regularization path progressively retains the largest-

magnitude coefficients and refits at each budget; selection AP determines the final candidate. Ties within 10^{-12} favor a smaller neuron budget, then stronger regularization. SAT-L1 ties favor smaller C_{reg} .

Single-head standardization uses fitting-set means and population standard deviations, replacing scales below 10^{-6} by one. The combiner applies the same rule to logits using selection-set statistics. Direct linear probes floor nonzero scales at 10^{-6} and replace zero scales by one; Sparse-AT removes zero-variance coordinates. No preprocessing statistics are estimated from the test set.

B.4 NUMERICAL OPTIMIZATION

Single-head probes use Newton’s method; readouts and Residual-L2 use L-BFGS, SAT-L1 uses SAGA, and Sparse-AT uses L-BFGS-B with proximal refinement. Numerical convergence checks do not use test performance. For the head-logit combiner, $C_{\text{reg}} = 1/(n\lambda_2)$ matches the mean-loss objective for n selection observations.

B.5 PAIRED BOOTSTRAP INTERVALS FOR TABLE 2

For each model and aggregation method, we first compute the sample standard deviation (SD) of test AP across 20 outer constructions separately for each task, then average these task-level SDs with equal weights: six tasks for Gemma and seven each for Llama and Qwen. The relative SD reduction of Frequency-Consensus versus a baseline is the proportional decrease, with the baseline SD as denominator.

We construct bootstrap intervals by resampling the 20 outer-construction indices 5,000 times with replacement (seed 2026092201). Each sampled index sequence is shared across methods, tasks, and models; model-specific results are computed separately. Within each bootstrap replicate we recompute task-level SDs, task-averaged SDs, and relative reductions; the reported intervals are the 2.5th and 97.5th percentiles of the replicate distribution, without adjustment for multiple comparisons.

Resampling operates on the saved outer-construction records only: the underlying tasks, generated responses, labels, and experimental configuration remain fixed, and neither model fitting nor inner-partition generation is repeated. The intervals therefore characterize construction variability conditional on these fixed components.

C SCRIPT GENERATION AND CONSTRAINT INJECTION

C.1 SCRIPT GENERATION

We used DeepSeek V4 Flash (DeepSeek-AI, 2026) to generate the user-message scripts with temperature 0.6 and top- $p = 0.95$. The model was prompted to produce all 14 user messages in a single generation, without assistant replies. The prompt keeps user messages independent of any imagined assistant response and excludes response constraints, which are injected separately during evaluation. A simplified version of the prompt is shown below.

You are writing user-message scripts for multi-turn dialogue evaluation. Produce a coherent sequence of 14 user messages. Generate the complete sequence at once; do not generate any assistant replies.

Requirements:

1. Open with a specific, realistic request.
2. Advance the conversation through natural follow-ups: ask clarifying questions, add details, express preferences, weigh trade-offs, or explore related sub-topics. Each turn should contribute meaningful content.
3. Keep the sequence coherent and centered on the initial request. Avoid abrupt shifts to entirely unrelated topics.
4. Each message must stand on its own without presupposing a particular assistant response. Do not reference or depend

```

on specific facts, suggestions, or decisions from an
imagined assistant reply.
5. Do not impose requirements on the format, length, tone,
title, opening, closing, required words, or forbidden words
of the assistant's responses. These constraints will be
injected separately during evaluation.
6. Do not remind the assistant to follow instructions,
comment on whether it has complied with any constraint, or
deliberately try to induce a constraint violation.

Return a JSON object in the following format:
```json
{
 "turns": [
 {"turn_id": 1, "user": "..."},
 {"turn_id": 2, "user": "..."},
 ...
 {"turn_id": 14, "user": "..."}
]
}
```

```

C.2 CONSTRAINT INJECTION

The constraint instruction is prepended to the first user message, separated from the original request by one ASCII space. Later user messages follow the fixed script unchanged, without appended reminders. Before each response, the model receives all user messages so far and its own prior responses in its model-specific chat format. The probe uses the final context token before the response is generated.

D EXACT CONSTRAINT VERIFICATION RULES

Sentence- and word-limit thresholds were selected per model based on compliance statistics to retain both compliant and violating responses beyond the first turn. Each threshold is fixed across all predictor constructions and comparisons within its model-constraint configuration.

We follow IFEval's programmatic evaluation methodology (Zhou et al., 2023). All seven constraint families use its deterministic checkers; required prefix follows its constrained-start rule. Compliance requires a nonempty response after whitespace stripping and a passing check. No loose-evaluation transformations, such as removing Markdown or boundary lines, are applied. Instructions reproduce the injected text in the saved records; verification uses the full generated response without an LLM judge.

Keyword frequency. *Instruction:* The word "hello" must appear at least 3 times in each of your responses.

Verification: The checker accepts at least three non-overlapping, case-insensitive regular-expression matches of hello. No word boundaries are required, so matching substrings within longer words count.

Quotation wrapping. *Instruction:* Each of your responses must be enclosed entirely in double quotation marks.

Verification: After whitespace stripping, the response must contain more than one character and begin and end with ASCII double quotation marks (U+0022). Typographic quotes fail this check; interior quotation marks are unrestricted.

Required prefix. *Instruction:* During the conversation, when it is your turn, please always start with "Here is my response:".

Verification: We use IFEval’s `ConstrainedStartChecker`: the phrase “Here is my response:” is matched case-sensitively at a line start, allowing leading whitespace, using the checker’s multiline regular expression.

Sentence limit. *Instruction:* Each of your responses must contain fewer than 7 sentences. (Gemma.) For Llama and Qwen1.5-7B, the instruction uses 8 instead of 7.

Verification: NLTK’s English Punkt tokenizer counts sentences. Acceptance requires at most 6 sentences for Gemma and 7 for Llama and Qwen1.5-7B, equivalent to strictly fewer than 7 and 8, respectively.

Title formatting. *Instruction (Gemma and Llama):* Each of your responses must contain a title, wrapped in double angular brackets, such as `<<poem of joy>>`. *Instruction (Qwen1.5-7B):* Each of your responses must contain a title wrapped in double angular brackets, such as `<<title>>`.

Verification: The response must contain a match of `<<[^\n]+>>`, with nonempty title content after removing the enclosing brackets and stripping whitespace. The title need not occupy the first line; relevance is not checked.

Two responses. *Instruction:* For each request, give two different responses separated only by 6 asterisk symbols: `*****`.

Verification: The checker splits on the literal delimiter `*****`. Exactly two nonempty segments must remain, with different whitespace-stripped contents. Empty leading or trailing segments are allowed; an empty interior segment causes failure. Semantic distinctness is not evaluated.

Word limit. *Instruction:* Each of your responses must contain fewer than 181 words. (Gemma.) The corresponding instructions use 71 for Llama and 201 for Qwen1.5-7B.

Verification: NLTK’s regular-expression tokenizer with pattern `\w+` counts words, not model tokens. Acceptance limits are 180, 70, and 200 words for Gemma, Llama, and Qwen1.5-7B, respectively. Punctuation separates tokens; digits and underscores are included.

E DIAGNOSTIC PROTOCOLS AND DATA COVERAGE

Condition definitions. The single-turn condition uses valid $t = 1$ observations, including dialogues with $F = 1$. The multi-turn condition excludes entire $F = 1$ dialogues. Its fitting and selection roles retain remaining valid at-risk observations, including first-turn negatives; its test role uses only $t \geq 2$.

Partitioning before matching. Both conditions share the conversation/script-group partitions defined in Appendix B. Matching is performed afterward, separately within each model–constraint pair and data role; observations never move between roles to meet a matching target.

Class-wise matching. For role q and label y , let $n_{S,q,y}$ and $n_{M,q,y}$ denote available single-turn and multi-turn counts. Both conditions are sampled uniformly without replacement down to

$$n_{q,y} = \min\{n_{S,q,y}, n_{M,q,y}\}.$$

This matches positive and negative counts separately, controlling sample size and violation prevalence within each role without forcing a 1:1 class ratio. Sampling does not favor dialogues with later first violations.

Repetition. Test matching is fixed across runs; fitting and selection matching are repeated for each outer construction. If either condition lacks a class in a role, that paired comparison is unavailable without redraw.

Predictor construction and reporting. Separate All-heads predictors use the same fixed layer, independent single-head logistic probes, linear logit readout, mean-BCE plus L2 objective, and numerical acceptance checks. Mean AP and sample SD are computed over 20 paired runs per configuration, then averaged across included models within each constraint family. Relative changes are $100(M/S - 1)$. Probes receive neither target-response tokens nor additional context covariates.

Coverage and recorded distributions. Figure 2 covers keyword frequency, quotation wrapping, sentence limit, and word limit, requiring more than 50 observations and at least 5% of each class in

both matched test sets. Gemma–Keyword frequency is excluded because only eight of 480 matched test observations are positive (1.67%). The 11 retained configurations each have 20 valid paired runs: keyword frequency includes Llama and Qwen1.5-7B; the other constraints include all three models. Missing configurations are not imputed. Table 5 reports matched test counts. Among retained configurations, the smallest matched test set has 117 observations and the smallest class proportion is 10.83%. These coverage criteria do not establish statistical significance.

E.1 MATCHED SAMPLE DISTRIBUTIONS

Table 5: Test class counts before and after matching positive and negative observation counts.

| Model | Constraint | S [−] | S ⁺ | M [−] | M ⁺ | Matched [−] | Matched ⁺ |
|------------|--------------------|----------------|----------------|----------------|----------------|----------------------|----------------------|
| Gemma | Quotation wrapping | 428 | 52 | 3111 | 266 | 428 | 52 |
| Gemma | Sentence limit | 157 | 323 | 863 | 121 | 157 | 121 |
| Gemma | Word limit | 179 | 301 | 1328 | 101 | 179 | 101 |
| Llama | Keyword frequency | 201 | 276 | 254 | 189 | 201 | 189 |
| Llama | Quotation wrapping | 232 | 245 | 588 | 221 | 232 | 221 |
| Llama | Sentence limit | 114 | 366 | 235 | 109 | 114 | 109 |
| Llama | Word limit | 248 | 232 | 293 | 245 | 248 | 232 |
| Qwen1.5-7B | Keyword frequency | 230 | 249 | 283 | 223 | 230 | 223 |
| Qwen1.5-7B | Quotation wrapping | 143 | 337 | 382 | 138 | 143 | 138 |
| Qwen1.5-7B | Sentence limit | 59 | 421 | 93 | 58 | 59 | 58 |
| Qwen1.5-7B | Word limit | 150 | 330 | 233 | 149 | 150 | 149 |

Note. S and M columns give available single-turn and multi-turn observations before matching. Matched counts apply separately and identically to both conditions. Superscripts [−] and ⁺ denote compliant and first-violation observations. The matched test sets are fixed across all 20 runs.

E.2 CLASS-COUNT-MATCHED RESULTS

Table 6: All-heads performance and variability in the class-count-matched comparison.

| Model | Constraint | Condition | Mean | SD |
|------------|--------------------|-----------|-------|------|
| Gemma | Quotation wrapping | S | 21.61 | 3.33 |
| | | M | 26.08 | 4.44 |
| Gemma | Sentence limit | S | 84.45 | 1.64 |
| | | M | 62.10 | 2.97 |
| Gemma | Word limit | S | 65.15 | 3.36 |
| | | M | 70.68 | 3.24 |
| Llama | Keyword frequency | S | 71.26 | 2.11 |
| | | M | 75.12 | 1.91 |
| Llama | Quotation wrapping | S | 94.98 | 0.63 |
| | | M | 91.74 | 1.28 |
| Llama | Sentence limit | S | 79.27 | 2.96 |
| | | M | 79.73 | 3.77 |
| Llama | Word limit | S | 80.15 | 1.01 |
| | | M | 84.33 | 1.57 |
| Qwen1.5-7B | Keyword frequency | S | 73.95 | 1.66 |
| | | M | 74.70 | 2.70 |
| Qwen1.5-7B | Quotation wrapping | S | 89.73 | 1.32 |
| | | M | 78.18 | 2.38 |
| Qwen1.5-7B | Sentence limit | S | 80.27 | 5.35 |
| | | M | 77.51 | 6.75 |
| Qwen1.5-7B | Word limit | S | 78.27 | 1.69 |
| | | M | 77.43 | 2.17 |

Note. All values are multiplied by 100 and calculated across 20 constructions within each configuration. S and M denote single-turn and multi-turn prediction. These statistics precede the across-model averaging used in Figure 2(a–b).

E.3 FILTERING AND TURN-WISE COVERAGE

The following counts are derived from the fixed observation records and partition assignments. Every configuration starts with 2,400 dialogue records. We first exclude whole dialogues with $F = 1$. Within the remaining dialogues, an unknown judgment or a length-truncated response excludes that observation and subsequent observations; earlier valid observations remain. A dialogue contributes a retained record only if at least one valid observation remains. Table 7 and Table 8 separate these exclusions from the count of observed events. Retained dialogues without an observed valid event include both horizon-complete and censored trajectories; they are not all assumed to have remained compliant for 14 turns.

In the filtering tables, N_0 is the number of non- $F = 1$ dialogues with no valid observation, N_V the retained dialogue count, and N_E the number with an observed valid violation. Thus $2400 = N_{F=1} + N_0 + N_V$. Dev and Test count retained dialogues in the fixed development and test pools. Gemma–Two responses is included here for accounting, but excluded from the main method comparison. Its 91 retained dialogues include 40 observed later violations and 51 without an observed violation; Gemma–Keyword frequency instead retains 2,365 dialogues, including 1,012 observed events and 1,353 without an observed event.

These are observation-eligibility counts, not the denominator of every descriptive statistic. Table 1 averages first-violation times conditional on an observed $F > 1$; dialogues without an observed violation are not assigned an event at turn 14. The larger-model table-generation script additionally excludes dialogues containing a truncated turn. Training may still retain their earlier reliable observations.

Task abbreviations: KF, keyword frequency; QW, quotation wrapping; RP, required prefix; SL, sentence limit; TF, title formatting; TR, two responses; WL, word limit.

Table 7: Dialogue filtering and fixed-pool membership (7B models). Each row starts with 2,400 dialogues.

| Model | Task | $N_{F=1}$ | N_0 | N_V | N_E | Dev | Test |
|----------|------|-----------|-------|-------|-------|------|------|
| Gemma-7B | KF | 35 | 0 | 2365 | 1012 | 1893 | 472 |
| Gemma-7B | QW | 220 | 0 | 2180 | 1351 | 1752 | 428 |
| Gemma-7B | RP | 0 | 0 | 2400 | 410 | 1920 | 480 |
| Gemma-7B | SL | 1594 | 0 | 806 | 601 | 649 | 157 |
| Gemma-7B | TF | 28 | 0 | 2372 | 1533 | 1898 | 474 |
| Gemma-7B | TR | 2309 | 0 | 91 | 40 | 72 | 19 |
| Gemma-7B | WL | 1493 | 0 | 907 | 508 | 728 | 179 |
| Llama-7B | KF | 1367 | 19 | 1014 | 934 | 813 | 201 |
| Llama-7B | QW | 1219 | 0 | 1181 | 1101 | 949 | 232 |
| Llama-7B | RP | 20 | 8 | 2372 | 2100 | 1897 | 475 |
| Llama-7B | SL | 1842 | 0 | 558 | 536 | 444 | 114 |
| Llama-7B | TF | 1 | 17 | 2382 | 1989 | 1908 | 474 |
| Llama-7B | TR | 2104 | 0 | 296 | 254 | 243 | 53 |
| Llama-7B | WL | 1159 | 0 | 1241 | 1228 | 993 | 248 |
| Qwen-7B | KF | 1260 | 2 | 1138 | 1102 | 908 | 230 |
| Qwen-7B | QW | 1693 | 0 | 707 | 660 | 564 | 143 |
| Qwen-7B | RP | 4 | 0 | 2396 | 1424 | 1917 | 479 |
| Qwen-7B | SL | 2098 | 0 | 302 | 301 | 243 | 59 |
| Qwen-7B | TF | 1 | 0 | 2399 | 251 | 1919 | 480 |
| Qwen-7B | TR | 2135 | 0 | 265 | 231 | 216 | 49 |
| Qwen-7B | WL | 1617 | 0 | 783 | 770 | 633 | 150 |

Table 8: Dialogue filtering and fixed-pool membership (13B–14B models). Each row starts with 2,400 dialogues.

| Model | Task | $N_{F=1}$ | N_0 | N_V | N_E | Dev | Test |
|-----------|------|-----------|-------|-------|-------|------|------|
| Llama-13B | KF | 32 | 6 | 2362 | 1421 | 1887 | 475 |
| Llama-13B | QW | 213 | 0 | 2187 | 1786 | 1750 | 437 |
| Llama-13B | RP | 19 | 4 | 2377 | 1765 | 1903 | 474 |
| Llama-13B | SL | 1899 | 0 | 501 | 476 | 400 | 101 |
| Llama-13B | TF | 0 | 64 | 2336 | 1357 | 1865 | 471 |
| Llama-13B | TR | 2013 | 0 | 387 | 325 | 306 | 81 |
| Llama-13B | WL | 1226 | 0 | 1174 | 1146 | 941 | 233 |
| Qwen-14B | KF | 97 | 0 | 2303 | 1696 | 1843 | 460 |
| Qwen-14B | QW | 115 | 0 | 2285 | 2220 | 1832 | 453 |
| Qwen-14B | RP | 34 | 0 | 2366 | 1646 | 1892 | 474 |
| Qwen-14B | SL | 1708 | 0 | 692 | 620 | 552 | 140 |
| Qwen-14B | TF | 0 | 1 | 2399 | 545 | 1919 | 480 |
| Qwen-14B | TR | 2220 | 0 | 180 | 122 | 146 | 34 |
| Qwen-14B | WL | 570 | 0 | 1830 | 1696 | 1464 | 366 |

Table 9 summarizes development coverage at three turns. The development pool is fixed across constructions; probe-fitting and selection allocations vary within it. Gemma–Two responses is excluded from these pooled counts.

Table 9: Available development observations at selected turns, pooled over each model’s included tasks. These totals describe coverage, not class balance or per-task evaluability.

| Model | Turn 2 | Turn 6 | Turn 14 |
|-----------|--------|--------|---------|
| Gemma-7B | 8840 | 5499 | 4508 |
| Llama-7B | 7219 | 1915 | 135 |
| Qwen-7B | 6397 | 3527 | 2623 |
| Llama-13B | 9014 | 4260 | 248 |
| Qwen-14B | 9648 | 5059 | 3048 |

The fixed-bank diagnostic evaluates 206 of 260 model–task–turn conditions; 54 lack a class in at least one required role.

In the class-count-matched single-turn comparison, three length-truncated first responses are excluded for Llama–Keyword frequency, three for Llama–Quotation wrapping and one for Qwen–Keyword frequency, leaving 477, 477 and 479 valid observations from original test pools of 480. Test groups are not redrawn.

E.4 FIXED-BANK DIAGNOSTIC PROTOCOL

Twenty banks per model retain identical memberships across tasks, turns and construction repeats; pairwise Jaccard similarity is at most $1/3$. For each model, constraint and turn, the plotted value is $100(\langle \text{SD} \rangle_{20} - \text{SD}_b)$, where each bank’s SD is computed over 20 constructions. Figure 4 averages these differences equally over available constraints within each model and turn. Gray dashed regions lack an evaluable condition.

For interpretation, we manually divide turns into three evaluation-support regions: *Data-rich*, *Moderate-data*, and *Data-sparse*. Their respective turn ranges are 2–5, 6–8, and 9–14 for Gemma; 2–4, 5–6, and 7–13 for Llama; and 2–3, 4–5, and 6–14 for Qwen. Llama turn 14 is unavailable. These descriptive ranges do not represent statistical confidence levels and do not alter bank construction, metric computation, or task averaging. We place primary interpretive weight on the data-rich region; patterns in the moderate-data and data-sparse regions are treated as descriptive because evaluation support is weaker. Task coverage changes across turns, so the aggregate pattern does not isolate a causal effect of turn position.

E.5 TURN-WISE SUPERVISION DIAGNOSTIC (FIGURE 3A)

This diagnostic covers six model–task configurations selected by sample availability, all evaluated at turns 2–6. The original probe-fitting and selection subsets are merged into a single development pool. Under the *natural* condition, each turn uses all available observations from this pool. Under the *count-matched* condition, positive and negative development counts are fixed separately across turns within each model–task pair using the per-turn quotas listed in Table 10. Matching preserves complete script groups; the selected subsets are held constant across constructions, and development/test membership is unchanged. The quotas enforce equal counts per turn, not a 1:1 class ratio.

Table 10: Fixed development-count quotas per turn for Figure 3(a).

| Model | Task | Dev positives
per turn | Dev negatives
per turn |
|-------|--------------------|---------------------------|---------------------------|
| Gemma | Quotation wrapping | 82 | 1079 |
| Llama | Required ending | 113 | 887 |
| Llama | Title formatting | 65 | 414 |
| Llama | Quotation wrapping | 44 | 151 |
| Llama | Required prefix | 43 | 373 |
| Qwen | Three placeholders | 41 | 98 |

All predictors use the All-Heads configuration at layer 16, with 16 candidate heads for Gemma and 32 for Llama and Qwen, and fixed regularization $\lambda_2 = 10^{-3}$. Each condition generates 20 four-fold grouped partitions of its own development pool. Probes fitted on three folds predict the held-out fold; the concatenated out-of-fold predictions are used to fit the linear readout. Full-pool single-head probes are computed once and reused across constructions; fold-specific probes and the readout are refitted in each construction. Both conditions follow the same seed rule, with folds constructed separately within their respective pools.

Both supervision settings and all constructions within a model–task–turn condition share the same valid at-risk test observations; no test balancing is applied. Test risk-set composition and class proportions may differ across turns. Performance statistics—AP mean and sample SD over 20 constructions—are first computed within each condition, then averaged with equal task weights within each model, and finally averaged equally over the three models.

E.6 TRAINING-WINDOW DIAGNOSTIC (FIGURE 3B)

This diagnostic uses the same set of 55 model–task–target–turn conditions for all four window widths: 20 conditions for Gemma, 20 for Llama, and 15 for Qwen. A condition is eligible if the target turn has at least 100 candidate observations with at least 25 per class, and the original role-validity checks are satisfied. These thresholds apply to the candidate pool, not to individual test folds. Table 11 reports the target-turn coverage for each model.

Table 11: Target-turn coverage shared by all four windows in Figure 3(b).

| Model | Conditions | Task: target turns |
|-------|------------|---|
| Gemma | 20 | QW: 5–14; RP: 14; SL: 5–7; TF: 5, 6, 14; WL: 5–7 |
| Llama | 20 | KF: 5–6; QW: 5–7; RP: 5–9; SL: 5; TF: 5–9; TR: 5–6; WL: 5–6 |
| Qwen | 15 | KF: 5; QW: 5–7; RP: 5–14; WL: 5 |

Each condition has a fixed target turn t^* . Among dialogues with a valid at-risk observation at t^* , positives are those whose first violation occurs at t^* ; negatives remain compliant at t^* , regardless of whether they violate at a later turn or have no observed violation. Dialogues that violate before t^* or lack a valid observation at t^* are excluded.

Predictors use the All-Heads configuration at layer 16 with fixed regularization $\lambda_2 = 10^{-3}$. Dialogues with first-turn violations are excluded, and no class balancing is applied. For each condition, the original 20 grouped partitions are retained. The training window includes the target turn t^* plus one to four preceding turns. Both the fitting and selection roles incorporate contiguous valid histor-

ical observations from their existing dialogues; neither dialogue membership nor role assignments change. Historical observations retain their original negative labels. Window expansion therefore increases the number of negative observations while leaving dialogue counts and positive counts unchanged. Test evaluation uses only target-turn observations, shared across all window widths and all constructions within each condition.

AP means and sample SDs are computed over the 20 constructions for each condition–window pair. Results are then averaged with equal weights: first over target turns within each task, then over tasks within each model, and finally over the three models.

F ADDITIONAL AGGREGATION RESULTS

This comparison follows Section 4.1: layer 16, no additional context covariates, and 20 outer repetitions. Dialogues with $F = 1$ are excluded; main test evaluation uses $t \geq 2$. Aggregation rules share the same 30 inner records.

Mean AP, sample SD (ddof = 1), range, and Worst₄ are computed per task before equal-weight macro-averaging. Absolute metrics are multiplied by 100. Bold marks the best absolute value among five methods, not statistical significance.

F.1 GEMMA

Table 12: Gemma: complete aggregation results (all metrics multiplied by 100).

| Method | Mean AP | AP SD | AP Range | Worst ₄ |
|---------------------|--------------|-------------|-------------|--------------------|
| ITI | 50.03 | 1.11 | 4.41 | 48.48 |
| Mean-AP | 50.26 | 1.03 | 4.13 | 48.81 |
| Median-AP | 50.21 | 1.06 | 4.20 | 48.70 |
| Mean-Rank | 50.22 | 0.99 | 3.90 | 48.91 |
| Frequency-Consensus | 50.23 | 0.97 | 3.85 | 48.92 |

F.2 LLAMA

Table 13: Llama: complete aggregation results (all metrics multiplied by 100).

| Method | Mean AP | AP SD | AP Range | Worst ₄ |
|---------------------|--------------|-------------|-------------|--------------------|
| ITI | 77.77 | 0.94 | 3.79 | 76.44 |
| Mean-AP | 77.77 | 0.98 | 3.77 | 76.37 |
| Median-AP | 77.87 | 1.00 | 4.08 | 76.41 |
| Mean-Rank | 77.77 | 0.97 | 3.82 | 76.38 |
| Frequency-Consensus | 77.88 | 0.81 | 3.32 | 76.70 |

F.3 QWEN1.5-7B

Table 14: Qwen1.5-7B: complete aggregation results (all metrics multiplied by 100).

| Method | Mean AP | AP SD | AP Range | Worst ₄ |
|---------------------|--------------|-------------|-------------|--------------------|
| ITI | 62.11 | 1.78 | 7.23 | 59.68 |
| Mean-AP | 62.06 | 1.61 | 6.00 | 59.93 |
| Median-AP | 62.09 | 1.64 | 6.09 | 59.97 |
| Mean-Rank | 62.00 | 1.61 | 6.03 | 59.92 |
| Frequency-Consensus | 62.02 | 1.55 | 5.84 | 59.98 |

G ADDITIONAL PREDICTOR-LEVEL METRICS

Values are on a 0–100 scale, with statistics computed per task before equal-weight averaging.

Table 15: Cross-construction AP range for 7B predictors; other metrics appear in Table 3(a).

| Method | Gemma-7B | Llama-2-7B | Qwen1.5-7B |
|---------------------|----------|------------|------------|
| C-only | 1.85 | 3.39 | 5.81 |
| All-heads | 4.66 | 5.77 | 9.51 |
| Random- K | 7.06 | 6.55 | 11.76 |
| Single-Split / ITI | 4.41 | 3.79 | 7.23 |
| Residual-L2 | 6.22 | 4.34 | 7.13 |
| Sparse-AT | 7.58 | 5.64 | 8.46 |
| SAT-L1 | 2.08 | 3.44 | 4.93 |
| Frequency-Consensus | 3.85 | 3.32 | 5.84 |

Table 16: Additional baselines for larger models; All-Heads, Single-Split/ITI and Frequency-Consensus appear in Table 3(b).

| Model | Method | Mean $\uparrow$ | SD $\downarrow$ | Range $\downarrow$ | Worst $_4\uparrow$ |
|-------------|-------------|-----------------|-----------------|--------------------|--------------------|
| Llama-2-13B | Random- K | 61.78 | 2.52 | 9.84 | 58.36 |
| | Residual-L2 | 63.93 | 1.23 | 4.72 | 62.27 |
| | Sparse-AT | 64.97 | 1.31 | 5.34 | 63.22 |
| | SAT-L1 | 40.49 | 0.81 | 3.34 | 39.39 |
| Qwen1.5-14B | Random- K | 53.88 | 3.38 | 11.62 | 49.12 |
| | Residual-L2 | 57.13 | 1.83 | 6.86 | 54.68 |
| | Sparse-AT | 58.59 | 1.66 | 7.12 | 56.40 |
| | SAT-L1 | 35.23 | 0.78 | 3.22 | 34.06 |

All values are multiplied by 100. Statistics are computed within each task and then macro-averaged.