

Bargaining with Internal Evidence: Receiver-Constrained Decoding for Factual Reliability in Large Language Models

Chenhao Lin¹, Mouyi Zhang¹, Peng Chen¹, Xin Zheng¹, Yuejia Hong¹, Nan Gao^{1*}

^{1*}Department of Computer Science, Zhejiang University of Technology,
Hangzhou, 310013, Zhejiang, China.

302023572126@zjut.edu.cn, zhangmouyi@zjut.edu.cn, chenpeng@zjut.edu.cn,
xinzheng@zjut.edu.cn, hongyuejia@zjut.edu.cn, gaonan@zjut.edu.cn

Abstract

Large language models can produce fluent yet factually incorrect responses, limiting their reliability in knowledge-intensive applications. Recent decoding methods seek to mitigate such hallucinations by contrasting or amplifying signals from intermediate layers. However, not every layerwise discrepancy provides a reliable direction for factual decoding. We propose a unified game-theoretic decoding framework for large language models: intermediate-layer signals act as competing evidence, while the final layer retains decision authority under explicit constraints. Rather than applying a fixed layerwise contrast, our framework treats changes in intermediate predictions as evidence of the language model’s evolving internal judgment, and uses this evidence to guide decoding under final-layer constraints. Experiments with three backbone language models on multiple truthfulness and hallucination benchmarks demonstrate the effectiveness of our framework in improving factual reliability.

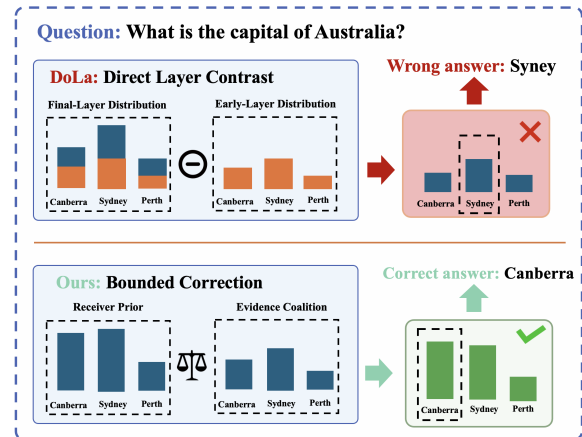


Figure 1: Illustration of receiver-constrained evidence bargaining for factual decoding. The upper panel shows that direct layer contrast can over-amplify an unreliable layerwise discrepancy and produce a wrong answer. The lower panel illustrates our bounded correction strategy, where the final-layer distribution serves as the receiver prior and layerwise evidence coalitions guide the prediction only within a constrained deviation from the original model decision.

1 Introduction

Large language models (LLMs) can produce fluent but factually incorrect responses, limiting their reliability in knowledge-intensive applications (Ji et al., 2023; Huang et al., 2025). This problem persists even in strong models and across both closed-form and open-ended evaluations (Lin et al., 2022; Ravichander et al., 2025). Since retraining or external verification is not always available, an important question is whether factual reliability can be improved directly at decoding time.

Inference-time decoding provides an attractive intervention point because it can alter generation behavior without parameter updates or an external verifier for every output. Recent work suggests that useful factuality signals may already be present inside the model. Intermediate hidden states can be projected into vocabulary-space distributions, exposing how next-token preferences

evolve across Transformer layers (Belrose et al., 2023). Building on this observation, DoLa contrasts predictions from different layers to amplify factual knowledge during decoding (Chuang et al., 2024); SLED exploits the evolution of self-logits across layers (Zhang et al., 2024); and PruneCD constructs a contrastive self-model through layer pruning rather than relying on early-exit logits (Yu et al., 2025). Together, these methods establish layerwise prediction dynamics as a promising source of decoding-time supervision.

However, they leave a reliability issue unresolved: when is an internal signal reliable enough to revise the final prediction? It may reflect latent evidence, but it may be a transient or confidently incorrect preference. As illustrated in Figure 1, directly amplifying such a discrepancy can override a reliable final-layer decision.

We address this issue by developing a reliable inference-time correction principle, under which intermediate-layer signals influence decoding only when they are locally corroborated and their deviation from the final decision is explicitly bounded. Specifically, we propose a receiver-constrained evidence bargaining decoder that treats layerwise predictions as candidate evidence rather than direct correction signals. At each decoding step, we identify belief-resolution events, where predictive uncertainty decreases sharply and the resulting preference is supported by neighboring layers. Witness layers around each event are pooled into an evidence coalition, reducing sensitivity to isolated layer fluctuations. The final-layer distribution remains the receiver prior, while the coalitions propose internally supported alternatives. We then compute the corrected distribution through a weighted bargaining objective under a divergence constraint from the receiver prior. Thus, internal evidence may influence decoding only when supported across the layerwise trajectory, and cannot induce an unconstrained departure from the model’s original decision. Our method operates entirely at inference time, without parameter updates, retrieval, or external factual verification.

We evaluate our method on TruthfulQA multiple-choice and generation (Lin et al., 2022), FACTOR (Muhlgay et al., 2024), PopQA (Mallen et al., 2023), and HaluEval (Li et al., 2023), using three instruction-tuned backbone models. Against greedy decoding and strong inference-time factuality baselines, including DoLa, SLED, END (Wu et al., 2025), and PruneCD, our method achieves particularly consistent gains on TruthfulQA multiple-choice and PopQA, while remaining broadly competitive in the more model- and task-dependent open-ended settings. These results support our central design principle: layerwise signals are most reliable when used as corroborated, bounded internal evidence rather than unconditional correction directions.

Our contributions are summarized as follows:

- We introduce a receiver-constrained evidence bargaining framework for factuality-oriented decoding, which formulates layerwise prediction dynamics as internal evidence and preserves the final-layer distribution as a constrained receiver prior.
- We instantiate this framework with a training-free decoder that detects belief-resolution

events, aggregates neighboring witness layers into event-level evidence coalitions, and controls their influence through a bounded bargaining objective.

- We conduct a systematic evaluation on five factuality and hallucination benchmarks across three backbone models. The results show consistent improvements on factual candidate ranking and short-form factual recall, while also revealing that the effectiveness of decoding-time correction depends on the backbone model and evaluation setting.

2 Related Work

Layerwise Decoding for Factual Reliability. Intermediate Transformer representations can be projected into vocabulary-space distributions, exposing how token preferences evolve across layers (Belrose et al., 2023). This observation has motivated training-free decoding methods for improving factuality. DoLa contrasts final- and intermediate-layer predictions (Chuang et al., 2024), while SLED exploits self-logit evolution across layers (Zhang et al., 2024). Entropy-based methods select internal signals according to cross-layer uncertainty dynamics (Das et al., 2025; Wu et al., 2025), and PruneCD constructs a contrastive reference from a pruned self-model (Yu et al., 2025). Unlike these methods, we do not treat selected internal predictions as direct correction sources. Instead, we use them as corroborated evidence whose influence is bounded by the final-layer prior.

Game-Theoretic Language Model Decoding. The Consensus Game formulates language generation as a signaling game between generative and discriminative querying modes, using equilibrium search to reconcile their predictions (Jacob et al., 2024). This perspective shows how competing predictive signals can be balanced through game-theoretic objectives during decoding in a principled manner. Our setting differs in both players and objective: we formulate factuality-oriented decoding as constrained bargaining between the final-layer receiver prior and evidence coalitions derived from the model’s layerwise predictive trajectory. The final-layer distribution retains decision authority, while intermediate-layer evidence can only induce bounded and internally supported corrections.

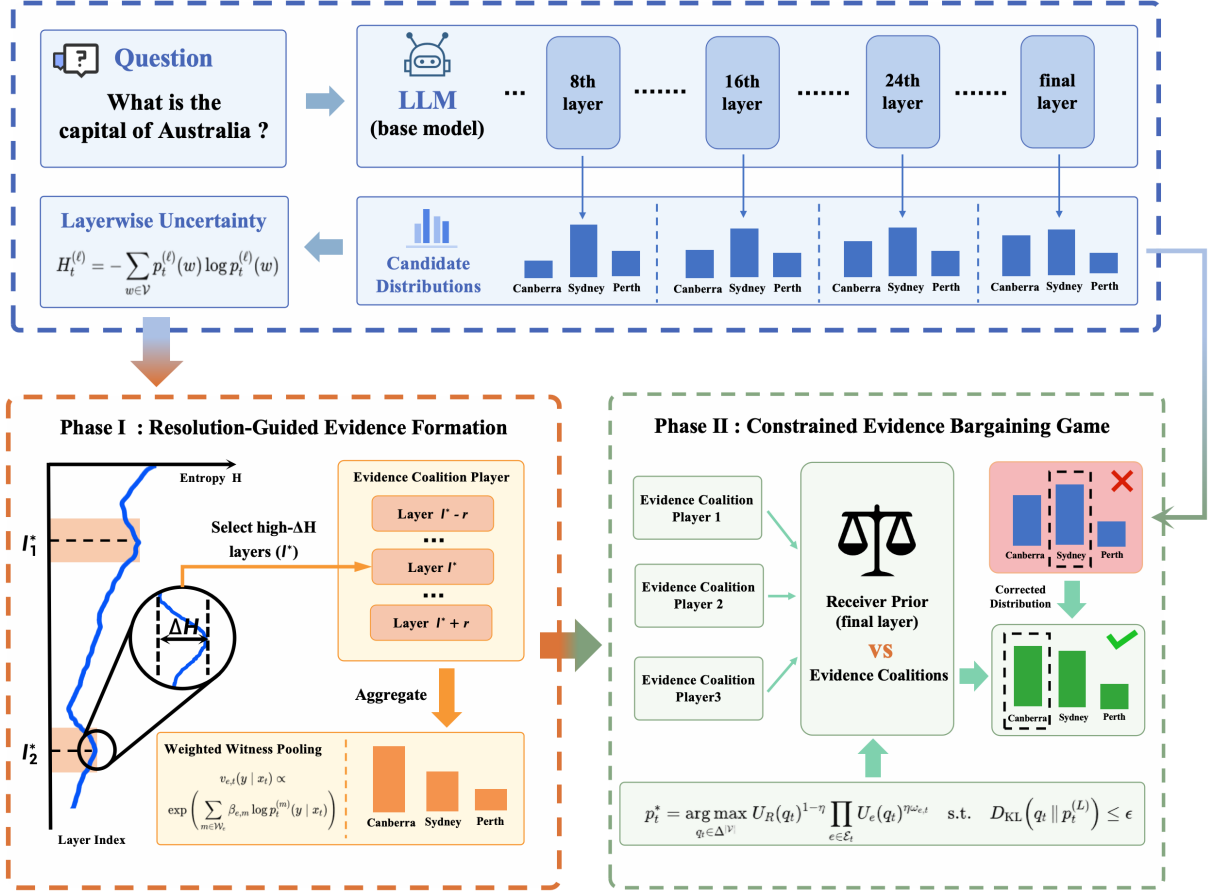


Figure 2: Overview of the proposed decoding framework. The base LLM first produces layerwise candidate distributions, whose uncertainty changes are used to select salient belief-resolution layers. In Phase I, neighboring witness layers around each selected event are pooled into event-level evidence coalitions. In Phase II, these coalitions interact with the final-layer distribution as a receiver prior through a constrained bargaining game, producing a corrected distribution that incorporates internal evidence while limiting unreliable deviations from the original model prediction.

3 Method

To prevent an unreliable intermediate preference from overriding a reliable final-layer decision, we formulate factual decoding as a *receiver-constrained evidence bargaining* problem. Intermediate layers first form locally corroborated evidence coalitions (Phase I), after which these coalitions negotiate with the final-layer distribution, which serves as a privileged receiver prior (Phase II). The resulting method is an inference-time LLM decoder: it modifies the next-token distribution of a frozen autoregressive language model and requires neither model updates nor an external factuality verifier.

3.1 Receiver Prior and Layerwise Evidence

Consider an autoregressive LLM generating token y_t given an input prompt x and a generated prefix $y_{<t}$. Let $h_t^{(\ell)}$ be the hidden state at layer

$\ell \in \{1, \dots, L\}$, where L denotes the final layer. Applying the frozen vocabulary readout $\psi(\cdot)$ of the backbone model yields a layerwise next-token distribution:

$$p_t^{(\ell)}(w) = \text{softmax}\left(\psi(h_t^{(\ell)})\right)_w, w \in \mathcal{V}. \quad (1)$$

Standard decoding predicts directly from $p_t^{(L)}$. These layerwise distributions expose how the model’s token preferences evolve throughout the forward pass. In contrast, we preserve $p_t^{(L)}$ as the receiver prior: it represents the model’s final decision before any internal intervention. Intermediate distributions are not allowed to overwrite this decision individually. Instead, they are used only to construct evidence proposals whose influence must be locally supported and bounded with respect to the final-layer prediction.

3.2 Phase I: Resolution-Guided Evidence Coalitions

Figure 2 shows that the first phase operates on the LLM’s layerwise uncertainty trajectory. We identify a *belief-resolution event* at an internal layer where predictive uncertainty decreases and the resulting token preference is supported by nearby layers. Let $\mathcal{I} = \{\ell_{\min}, \dots, L-1\}$ denote the set of candidate internal layers. For each $\ell \in \mathcal{I}$, we compute the layerwise entropy and its positive reduction:

$$H_t^{(\ell)} = - \sum_{w \in \mathcal{V}} p_t^{(\ell)}(w) \log p_t^{(\ell)}(w), \quad (2)$$

$$\Delta_t^{(\ell)} = \left[H_t^{(\ell-1)} - H_t^{(\ell)} \right]_+, \quad (3)$$

where $[z]_+ = \max(z, 0)$. A sharp entropy reduction indicates that the LLM has become more decisive, but it does not by itself imply a reliable correction direction. We therefore combine it with a local persistence term $C_t^{(\ell)}$, which measures agreement with neighboring internal-layer distributions:

$$S_t^{(\ell)} = \Delta_t^{(\ell)} C_t^{(\ell)}. \quad (4)$$

We retain salient, non-overlapping maxima of $S_t^{(\ell)}$ as the event set $\mathcal{E}_t$. For each selected event $e \in \mathcal{E}_t$, neighboring witness layers are pooled into an event-level evidence proposal $v_{e,t}$. When multiple events are present, a proposal receives a larger bargaining weight $\omega_{e,t}$ only when it is both salient and consistent with other selected proposals. Thus, Phase I converts the layerwise trajectory into locally persistent and cross-event corroborated evidence, rather than exposing the final decision to an isolated layer fluctuation. Appendix A.2 and Appendix A.3 specify the complete construction.

3.3 Phase II: Receiver-Constrained Evidence Bargaining

We next reconcile the evidence coalitions with the final-layer receiver prior. For any candidate corrected next-token distribution $q_t \in \Delta^{|\mathcal{V}|}$, we define the receiver utility and the utility of evidence coalition e as

$$\begin{aligned} U_R(q_t) &= \exp \left[-D_{\text{KL}} \left(q_t \parallel p_t^{(L)} \right) \right], \\ U_e(q_t) &= \exp \left[-D_{\text{KL}}(q_t \parallel v_{e,t}) \right]. \end{aligned} \quad (5)$$

The receiver utility favors the original final-layer distribution, whereas each coalition utility favors

the token preference supported by its local evidence. We obtain the corrected LLM next-token distribution through the weighted Nash-welfare bargaining objective

$$\begin{aligned} p_t^* &= \arg \max_{q_t \in \Delta^{|\mathcal{V}|}} U_R(q_t)^{1-\eta} \prod_{e \in \mathcal{E}_t} U_e(q_t)^{\eta \omega_{e,t}} \\ \text{s.t. } & D_{\text{KL}}(q_t \parallel p_t^{(L)}) \leq \epsilon, \end{aligned} \quad (6)$$

where $\eta \in [0, 1]$ controls the total bargaining mass assigned to internal evidence, $\sum_{e \in \mathcal{E}_t} \omega_{e,t} = 1$, and $\epsilon \geq 0$ specifies the maximum admissible departure from the receiver prior.

Equation (6) separates evidence utilization from decision authority. Unlike direct layer contrast, evidence does not arise from an arbitrary intermediate layer, but from resolution-guided coalitions supported by the model’s predictive trajectory. Unlike an unconstrained combination of distributions, the final layer remains privileged: even strongly supported evidence can revise the next-token prediction only inside the receiver’s admissible boundary. If no belief-resolution event is selected, the decoder falls back to the original final-layer distribution. Taking logarithms converts Eq. (6) into a convex weighted-KL problem; the efficient constrained solver is derived in Appendix A.4.

3.4 Generation and Fixed-Choice Instantiations

For open-ended generation, we apply the bargaining operator at every autoregressive decoding step and select the next token from p_t^* . For fixed-choice benchmarks, we apply the same operator to distributions over complete answer candidates obtained from layerwise continuation likelihoods. Importantly, no separately fitted receiver calibration is introduced in the fixed-choice evaluation; the final-layer receiver and intermediate layers are converted to answer-space distributions using the same fixed score transform.

Across both settings, the event-selection, coalition-construction, and receiver-constrained aggregation procedures remain unchanged. The only difference lies in the prediction space on which the bargaining operator is applied: token distributions for generation and candidate-answer distributions for fixed-choice evaluation. Appendix A.6 provides the complete fixed-choice instantiation.

Model	Method	TFQA MC			TFQA Gen			FACTOR	PopQA	HaluEval
		MC1	MC2	MC3	T	I	T×I	Acc.	Acc.	Acc.
Llama-3.1-8B	Greedy	38.43	57.01	32.13	89.10	43.82	39.04	61.38	27.65	68.27
	DoLa	39.41	57.15	29.27	80.29	66.82	53.65	63.32	27.53	67.92
	PruneCD	43.32	62.36	32.79	94.49	85.68	80.93	65.00	27.18	69.00
	END	41.98	61.60	34.04	87.03	49.45	43.04	61.95	27.58	66.67
	SLED	41.12	61.74	34.32	87.76	72.70	63.81	63.76	27.64	68.27
	Ours	45.65	62.81	35.44	92.66	83.48	77.35	65.06	30.26	68.28
Qwen2.5-7B	Greedy	46.75	65.14	35.28	83.10	71.97	60.52	55.21	19.01	79.58
	DoLa	42.35	52.64	30.86	88.49	66.21	58.60	40.71	17.14	77.09
	PruneCD	43.32	63.40	31.26	80.78	82.01	66.25	58.58	19.03	75.08
	END	43.57	60.28	32.98	83.59	67.07	56.06	58.81	18.98	79.48
	SLED	47.12	65.68	36.10	84.21	64.63	54.42	58.31	18.99	79.33
	Ours	49.45	67.81	40.80	85.18	87.51	74.55	48.43	21.27	79.69
Mistral-7B	Greedy	48.71	66.29	37.59	79.80	78.46	62.61	60.75	26.61	68.80
	DoLa	45.53	65.19	33.46	79.56	92.17	73.33	63.69	26.60	68.82
	PruneCD	48.83	61.20	37.00	86.54	92.66	80.19	64.46	26.38	68.84
	END	48.59	64.95	38.68	80.17	82.37	66.04	61.49	26.59	68.84
	SLED	50.43	67.21	41.09	79.19	78.95	66.03	67.67	26.61	68.80
	Ours	51.04	65.58	40.92	81.03	96.33	78.05	68.33	29.62	71.00

Table 1: Main results across three backbone models and factuality benchmarks. The best-performing results are highlighted in green, and the second-best results are highlighted in blue.

4 Experiment

4.1 Experimental Setup

Benchmarks We evaluate the proposed decoding framework on five benchmarks that cover factual question answering, hallucination detection, and knowledge-intensive reasoning. TruthfulQA (Lin et al., 2022) is used in both the multiple-choice setting (TFQA MC) and the generation setting (TFQA), where models must avoid common human falsehoods while still producing informative answers. FACTOR (Muhlgay et al., 2024) and PopQA (Mallen et al., 2023) further evaluate factual knowledge and long-tail entity knowledge, respectively. We also include HaluEval (Li et al., 2023), which measures whether a model can identify or avoid hallucinated content across knowledge-grounded tasks. For all benchmarks, we follow the official evaluation protocol when available and keep the prompt format fixed across decoding methods.

Models and Baselines We conduct experiments on three instruction-tuned autoregressive language models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). These models provide a diverse testbed across model families while remaining comparable in scale. To maintain a training-free comparison pro-

cedure, we restrict the baselines to decoding methods that require neither additional policy training nor task-specific supervision beyond the frozen backbone model. We compare our method with greedy decoding, DoLa (Chuang et al., 2024), PruneCD (Yu et al., 2025), END (Wu et al., 2025), and SLED (Zhang et al., 2024). Greedy decoding serves as the standard final-layer baseline, while the remaining baselines represent recent inference-time approaches that modify the token distribution using layerwise or contrastive signals. For a fair comparison, each method uses the same backbone model, prompt template, maximum generation length, and answer extraction procedure on each benchmark.

Metrics For TFQA MC, we report the three standard multiple-choice scores used by the benchmark script. For TFQA generation, we report truthfulness (T), informativeness (I), and their joint score (T×I), which measure whether an answer is factual and useful. For FACTOR, PopQA, and HaluEval, we report accuracy following the corresponding benchmark protocol. Higher values indicate better factual reliability. Unless otherwise stated, results are averaged over the official test examples, with the same random seed used for each backbone model. All metrics follow the official evaluation procedures.

4.2 Implementation Details

All experiments are implemented in a unified Hugging Face Transformers evaluation framework and conducted on four NVIDIA GeForce RTX 5090 GPUs. We use bfloat16 inference, seed 42, one beam, and deterministic decoding. For choice-based tasks, candidate answers are ranked by continuation log likelihood; for generation tasks, we use task-specific output budgets. All methods share the same prompts, stopping rules, answer extraction procedure, and model loading pipeline. We evaluate DoLa, END, PruneCD, and SLED with their recommended layer-selection and decoding settings, with additional decoder specification and task instantiation details provided in Appendix A.

4.3 Main Results

Table 1 presents the main results across three backbone models and five factuality settings. Using one headline metric per setting (MC1 for TFQA MC and $T \times I$ for TFQA Generation), our method achieves the numerically best result on 11 of the 15 backbone–task combinations and ranks within the top two on 14 of them. These results demonstrate that our method provides consistent improvements on factual candidate ranking and short-form factual recall, particularly on TruthfulQA MC and PopQA. Performance on open-ended generation tasks varies across models and evaluation protocols, reflecting the difficulty of simultaneously optimizing truthfulness, informativeness, and generation behavior. These results indicate that our framework provides broadly effective improvements in factual reliability, while the remaining exceptions also show that the utility of decoding-time correction depends on both the backbone model and the evaluation regime.

Consistent gains in factual selection and short-answer recall. Our method is most consistent on TFQA MC and PopQA. On TFQA MC, it obtains the highest MC1 score for all three backbones and achieves the best result on seven of the nine reported MC metrics. In particular, it improves MC1 over greedy decoding by 7.22, 2.70, and 2.33 points on Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B, respectively. It also outperforms PruneCD on all nine TFQA MC results, suggesting that the proposed receiver-constrained evidence mechanism provides more reliable factual preference ranking under fixed candidate answers. The same cross-model consistency is observed on PopQA, where

our method is best for all three backbones and improves over greedy by 2.61, 2.26, and 3.01 points, respectively. These results suggest that corroborated and bounded internal evidence is particularly effective when the task requires selecting or recalling a concise factual answer.

Open-ended generation and the scope of PruneCD’s advantage. The results on TFQA Generation are more heterogeneous because the joint score depends simultaneously on truthfulness and informativeness. PruneCD performs particularly well in this setting, obtaining the highest $T \times I$ score on Llama-3.1-8B and Mistral-7B, with scores of 80.93 and 80.19, compared with 77.35 and 78.05 for our method. However, this advantage does not hold uniformly across backbones. On Qwen2.5-7B, our method achieves the best informativeness score (87.51) and the best joint score (74.55), outperforming PruneCD by 8.30 points in $T \times I$. This comparatively large margin arises because our method improves both truthfulness (85.18 vs. 80.78) and informativeness (87.51 vs. 82.01) over PruneCD, rather than increasing informativeness at the expense of truthfulness. More importantly, PruneCD’s advantage is concentrated in open-ended TFQA generation: outside this setting, our method outperforms PruneCD on 16 of the remaining 18 reported metric–backbone pairs, including all nine TFQA MC metrics and all three PopQA results. Thus, although PruneCD is a strong baseline for open-ended truthfulness generation, our method provides substantially more consistent benefits across factual ranking and knowledge-recall evaluations.

Greedy sufficiency and the risk of over-correction. A notable pattern in Table 1 is that additional decoding correction is not always beneficial when greedy decoding is already competitive. On HaluEval with Qwen2.5-7B, greedy decoding achieves 79.58, outperforming DoLa, PruneCD, END, and SLED; our method improves this already strong baseline only marginally, to 79.69. Similarly, on HaluEval with Llama-3.1-8B, our result is effectively tied with greedy decoding (68.28 vs. 68.27). More broadly, relative to greedy decoding, DoLa, END, PruneCD, and SLED underperform on 17, 13, 9, and 7 of the 27 reported metrics, respectively, whereas our method underperforms on only two. These regressions are consistent with a potential over-correction risk: when the final-layer receiver prior already yields a reliable prediction,

strongly modifying it using imperfect intermediate-layer signals may be unnecessary or even harmful. In contrast, the substantially fewer regressions of our method are consistent with its design goal of exploiting internally supported evidence while limiting unsupported deviations from the final-layer decision.

Task-dependent exceptions. Our method obtains the numerically best FACTOR accuracy on Llama-3.1-8B and Mistral-7B, although the improvement over PruneCD on Llama-3.1-8B is small (65.06 vs. 65.00). On Qwen2.5-7B, however, our method performs poorly on FACTOR, falling below both greedy decoding (48.43 vs. 55.21) and the strongest baseline END (58.81). This failure case suggests that the selected internal evidence may still be misaligned with the continuation-ranking behavior of a particular backbone–task combination. On HaluEval, our method yields a clear gain on Mistral-7B (71.00, compared with 68.84 for the strongest baseline), while offering only marginal or negligible gains on the other two backbones. These results further indicate that the effectiveness of internal correction depends on whether the final-layer distribution leaves meaningful room for evidence-supported revision.

Summary. Taken together, the results support our central design principle: layerwise predictions are most reliably used as corroborated and receiver-constrained evidence for correcting uncertain or incomplete final-layer decisions, rather than as unconditional correction directions that override a receiver prior which may already be reliable.

4.4 Scaling Evaluation

To assess the scalability of our method to larger language models, we further evaluate it on Llama-3-70B-Instruct under the same prompts, task protocols, and deterministic decoding settings as in the main experiments.

On TruthfulQA MC, our method substantially outperforms DoLa and END across all three metrics. Compared with DoLa, it improves MC1/MC2/MC3 by 8.08/5.70/6.69 points; compared with END, the gains are 6.98/8.94/8.43 points. It is also competitive with SLED, trailing by only 0.36 and 0.26 points on MC1 and MC2 while achieving the best MC3 score by 0.93 points. These results show that the proposed receiver-constrained evidence mechanism remains effective for factual candidate ranking at the 70B scale.

Method	MC1	MC2	MC3	Truth	Info	T×I
Greedy	47.12	66.56	37.16	95.22	26.68	25.41
DoLa	40.27	61.58	33.63	95.10	29.01	27.59
SLED	48.71	67.54	39.39	95.71	26.20	25.07
END	41.37	58.34	31.89	95.23	26.32	25.05
Ours	48.35	67.28	40.32	90.83	62.18	56.48

Table 2: Results on Llama-3-70B-Instruct. Our method achieves the best MC3, Info, and T×I scores while remaining competitive on MC1 and MC2.

On TruthfulQA generation, our method achieves 90.83/62.18/56.48 in Truth/Info/T×I. Its Info and joint score more than double the strongest baseline results of 29.01 and 27.59. Output inspection confirms that the baselines frequently produce the prompt-permitted refusal “I have no comment,” which preserves Truth but provides little information. Our method produces substantially fewer refusals and therefore achieves a stronger balance among truthfulness, informativeness, and abstention.

Table 2 reports the results on TruthfulQA. Our method consistently improves factual candidate ranking compared with direct layer contrast baselines, demonstrating that the proposed evidence aggregation mechanism remains effective at larger model scales.

4.5 Ablation Study

We conduct ablations on TruthfulQA MC with Llama-3.1-8B-Instruct. This setting provides a controlled candidate-ranking testbed, allowing us to isolate the effect of each decoding component without generation-time answer extraction noise. The ablations are designed around three questions: whether internal layerwise evidence is useful beyond the final-layer receiver prior, whether such evidence can be used through unconstrained fusion, and whether belief-resolution selection and witness-layer coalitions are necessary for stable factual ranking.

We compare the full model with six ablated variants. Final-layer only removes all internal evidence and directly uses the final-layer receiver distribution. Direct full-layer contrast replaces our evidence formation and bargaining process with a direct contrast between the final layer and internal layers. Unconstrained geometric fusion uses the same selected evidence coalitions as our method, but removes the receiver KL constraint. Fixed-layer selection replaces belief-resolution event selection with a static layer choice. Event-layer only w/o

Variant	MC1	Δ MC1	MC2	Δ MC2	MC3	Δ MC3
Ours	45.65	0.00	62.81	0.00	35.44	0.00
Final-layer only	38.43	-7.22	57.01	-5.80	32.13	-3.31
Direct full-layer contrast	35.62	-10.03	53.35	-9.46	27.56	-7.88
Unconstrained geometric fusion	49.33	+3.68	51.34	-11.47	36.69	+1.25
Fixed-layer selection	44.31	-1.34	61.84	-0.97	34.35	-1.09
Event-layer only w/o witnesses	45.53	-0.12	63.75	+0.94	32.77	-2.67
Mean coalition pooling	45.04	-0.61	61.88	-0.93	34.91	-0.53

Table 3: Core ablations on TruthfulQA MC with Llama-3.1-8B-Instruct. Each variant removes or replaces one key design choice of the proposed receiver-constrained evidence bargaining decoder. Deltas are absolute point differences from the full model.

witnesses uses only the selected event layer as evidence, without aggregating neighboring witness layers. Mean coalition pooling keeps the witness set but replaces similarity-weighted pooling with uniform pooling.

Table 3 shows that the improvement does not come from the final-layer likelihood alone. The final-layer-only variant drops substantially across all three metrics, indicating that selected internal evidence provides useful information beyond the original receiver distribution. However, simply exposing the decoder to more layerwise signals is not sufficient. Direct full-layer contrast performs even worse than the final-layer-only variant, with drops of 10.03 points on MC1 and 7.88 points on MC3. This supports our central motivation: salient layerwise discrepancies may contain useful information, but using them as unconditional correction directions can amplify unreliable internal preferences.

The unconstrained geometric fusion variant provides a more nuanced comparison. It improves MC1 and MC3 over the full model, but causes a large drop on MC2. This suggests that removing the receiver KL constraint can make the decoder more aggressive in promoting a single preferred truthful candidate, while distorting the calibrated distribution over all truthful and false candidates. Thus, the receiver constraint is not merely a conservative regularizer for maximizing every individual metric. Rather, it stabilizes the distributional ranking behavior of the decoder and prevents internal evidence from inducing an overly large departure from the final-layer receiver prior.

The remaining variants examine the evidence formation stage. Replacing belief-resolution selection with fixed-layer selection consistently reduces all three metrics, showing that input-dependent event selection is more reliable than using a static layer choice. This result supports the design of selecting evidence according to layerwise uncertainty

reduction and local support, rather than assuming that the same internal layer is informative for every instance.

The coalition ablations further show that witness layers affect the balance between different ranking metrics. Using only the selected event layer slightly improves MC2, but substantially lowers MC3, indicating that a single event layer is less reliable for preserving pairwise preference consistency between truthful and false answers. Similarly, replacing similarity-weighted coalition pooling with mean pooling weakens all three metrics. Together, these results suggest that neighboring witness layers provide local corroboration that makes the evidence proposal more stable.

5 Conclusion

We introduced a receiver-constrained layerwise evidence decoding framework for improving factual reliability in large language models. The key idea is to treat intermediate-layer predictions as internally generated evidence rather than direct correction signals, and to allow their influence only when they are supported by neighboring layers and constrained by the final-layer distribution. This enables inference-time factual correction without parameter updates, retrieval, or external verification.

Experiments across three backbone models and multiple factuality benchmarks demonstrate that our method provides consistent improvements on factual candidate ranking and short-form factual recall, while remaining competitive in more challenging open-ended generation settings. The ablation and sensitivity analyses further show that reliable internal evidence requires both corroboration across layers and controlled deviation from the final-layer decision.

Our findings suggest that the layerwise predictive trajectory of LLMs can serve as a useful source of internal evidence when incorporated with

appropriate reliability constraints. Future work will explore more efficient evidence extraction and adaptive activation strategies to improve the performance–efficiency trade-off of inference-time factuality enhancement.

Limitations

Although our method improves factual reliability across multiple benchmarks, several limitations remain. First, it relies on internal layerwise evidence and cannot recover facts absent from or incorrectly represented in the base model without external retrieval or verification. Second, its effectiveness is model- and task-dependent, with more consistent gains on factual candidate ranking and short-form recall than on open-ended generation. Finally, the method introduces inference overhead from extracting intermediate-layer predictions and constructing evidence coalitions. Future work will explore efficient layer selection and lightweight aggregation to reduce latency while preserving factual reliability.

Acknowledgments

This work was supported by the Zhejiang Provincial Natural Science Foundation of China (Grant No. LMS26F020023).

References

- Nora Belrose, Zach Furman, Logan Smith, Danny Hahlawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. [Eliciting latent predictions from transformers with the tuned lens](#). *arXiv preprint arXiv:2303.08112*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. [DoLa: Decoding by contrasting layers improves factuality in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Souvik Das, Lifeng Jin, Linfeng Song, Haitao Mi, Baolin Peng, and Dong Yu. 2025. [Entropy guided extrapolative decoding to improve factuality in large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6589–6600, Abu Dhabi, UAE. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*, 43(2):1–55.
- Athul Paul Jacob, Yikang Shen, Gabriele Farina, and Jacob Andreas. 2024. [The consensus game: Language model generation via equilibrium search](#). In *The Twelfth International Conference on Learning Representations*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khoshabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. [Generating benchmarks for factuality evaluation of language models](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian’s, Malta. Association for Computational Linguistics.
- Abhilasha Ravichander, Shruti Ghela, David Wadden, and Yejin Choi. 2025. [HALoGEN: Fantastic LLM](#)

hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.

Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang, and Longjun Cai. 2025. [Improve decoding factuality by token-wise cross layer entropy of large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3912–3921, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jixi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.

Byeongho Yu, Changhun Lee, Jun-gyu Jin, and Eunhyeok Park. 2025. [PruneCD: Contrasting pruned self model to improve decoding factuality](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32462–32473, Suzhou, China. Association for Computational Linguistics.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024. [SLED: Self logits evolution decoding for improving factuality in large language models](#). In *Advances in Neural Information Processing Systems*, volume 37.

A Complete Specification of the Proposed Decoder

This appendix provides the complete computational specification of the receiver-constrained evidence bargaining decoder introduced in Section 3. All distributions are computed from a frozen autoregressive LLM and its layerwise next-token predictions. During open-ended generation, the decoder uses no gold labels, retrieval module, external factuality verifier, or parameter update.

A.1 Divergences and Numerical Stabilization

All layerwise distributions are obtained using the same frozen vocabulary readout $\psi(\cdot)$ as in Eq. (1). To prevent undefined logarithms and divergences, we apply a numerical floor $\zeta > 0$:

$$\text{safe}(p)(w) = \frac{\max(p(w), \zeta)}{\sum_u \max(p(u), \zeta)}. \quad (7)$$

All logarithms and divergences in the decoder are computed using the stabilized distributions. For distributions p and q over the same support, we use

$$D_{\text{KL}}(p\|q) = \sum_w p(w) \log \frac{p(w)}{q(w)}, \quad (8)$$

$$D_{\text{JS}}(p\|q) = \frac{1}{2}D_{\text{KL}}(p\|m) + \frac{1}{2}D_{\text{KL}}(q\|m),$$

$$m = \frac{p+q}{2}. \quad (9)$$

A.2 Belief-Resolution Event Selection

The candidate internal-layer set $\mathcal{I}$, entropy $H_t^{(\ell)}$, positive entropy reduction $\Delta_t^{(\ell)}$, and resolution score $S_t^{(\ell)}$ are defined in Section 3.2. For each candidate layer $\ell \in \mathcal{I}$, let

$$\mathcal{N}_r(\ell) = \{m \in \mathcal{I} : 0 < |m - \ell| \leq r\} \quad (10)$$

denote its local neighborhood. We define the local persistence term as

$$c_t^{(\ell,m)} = \exp\left(-D_{\text{JS}}\left(p_t^{(\ell)}\|p_t^{(m)}\right)/\tau_c\right),$$

$$C_t^{(\ell)} = \frac{1}{|\mathcal{N}_r(\ell)|} \sum_{m \in \mathcal{N}_r(\ell)} c_t^{(\ell,m)}, \quad (11)$$

where $\tau_c > 0$ controls sensitivity to local disagreement. When $\mathcal{N}_r(\ell) = \emptyset$, we set $C_t^{(\ell)} = 0$.

If $\max_{\ell \in \mathcal{I}} S_t^{(\ell)} = 0$, no event is activated and $\mathcal{E}_t = \emptyset$. Otherwise, let $\text{LM}(S_t)$ denote local maxima of the resolution scores. We first retain salient candidates

$$\tilde{\mathcal{E}}_t = \left\{ \ell \in \text{LM}(S_t) : S_t^{(\ell)} \geq \rho \max_{j \in \mathcal{I}} S_t^{(j)} \right\}, \quad (12)$$

where $\rho \in [0, 1]$ is a relative salience threshold. We then perform non-maximum suppression: candidates are visited in decreasing order of $S_t^{(\ell)}$, and a candidate is retained only if it lies more than δ layers away from every previously retained event. We keep at most K events:

$$\mathcal{E}_t = \text{NMS}_{\delta, K}(\tilde{\mathcal{E}}_t). \quad (13)$$

This selection prevents a single local transition from generating multiple redundant evidence coalitions.

A.3 Evidence Coalition Construction and Corroboration

For each selected event $e \in \mathcal{E}_t$, define its witness-layer set

$$\mathcal{W}_e = \{m \in \mathcal{I} : |m - e| \leq r_w\}, \quad (14)$$

where r_w is the witness radius. Rather than treating the event-center distribution as a direct correction source, we geometrically pool predictions from its local witnesses. The witness weight is

$$\beta_{e,m} = \frac{\exp[-D_{\text{JS}}(p_t^{(m)} \| p_t^{(e)}) / \tau_w]}{\sum_{j \in \mathcal{W}_e} \exp[-D_{\text{JS}}(p_t^{(j)} \| p_t^{(e)}) / \tau_w]}, \quad (15)$$

and the resulting event-level evidence coalition is

$$\begin{aligned} a_{e,t}(w) &= \sum_{m \in \mathcal{W}_e} \beta_{e,m} \log p_t^{(m)}(w), \\ v_{e,t}(w) &= \frac{\exp(a_{e,t}(w))}{\sum_{u \in \mathcal{V}} \exp(a_{e,t}(u))}. \end{aligned} \quad (16)$$

Here, $\tau_w > 0$ controls how sharply the coalition concentrates on witnesses close to the event center.

When $|\mathcal{E}_t| = 1$, we set $\omega_{e,t} = 1$. When multiple events are selected, we form a geometric consensus distribution

$$\begin{aligned} \bar{a}_t(w) &= \frac{1}{|\mathcal{E}_t|} \sum_{e \in \mathcal{E}_t} \log v_{e,t}(w), \\ \bar{v}_t(w) &= \frac{\exp(\bar{a}_t(w))}{\sum_{u \in \mathcal{V}} \exp(\bar{a}_t(u))}, \end{aligned} \quad (17)$$

and measure the corroboration of coalition e by

$$A_{e,t} = \exp\left(-\frac{D_{\text{JS}}(v_{e,t} \| \bar{v}_t)}{\tau_a}\right). \quad (18)$$

The normalized bargaining weight combines event salience and cross-event corroboration:

$$\omega_{e,t} = \frac{S_t^{(e)} A_{e,t}}{\sum_{j \in \mathcal{E}_t} S_t^{(j)} A_{j,t}}. \quad (19)$$

Consequently, an internal proposal receives substantial influence only when it is induced by a salient belief-resolution event and remains compatible with other selected evidence.

A.4 Efficient Solver for Receiver-Constrained Bargaining

Using the utilities in Eq. (5), taking the negative logarithm of Eq. (6) yields the equivalent constrained optimization problem

$$\begin{aligned} \mathcal{R}_t(q_t) &= \sum_{e \in \mathcal{E}_t} \omega_{e,t} D_{\text{KL}}(q_t \| v_{e,t}), \\ \mathcal{L}_t(q_t) &= (1 - \eta) D_{\text{KL}}(q_t \| p_t^{(L)}) + \eta \mathcal{R}_t(q_t), \\ p_t^* &= \arg \min_{q_t \in \Delta^{|\mathcal{V}|}} \mathcal{L}_t(q_t), \\ \text{s.t. } D_{\text{KL}}(q_t \| p_t^{(L)}) &\leq \epsilon. \end{aligned} \quad (20)$$

Since $\sum_{e \in \mathcal{E}_t} \omega_{e,t} = 1$, the unconstrained optimizer is a normalized geometric aggregate. Define

$$\begin{aligned} b_t(w) &= \sum_{e \in \mathcal{E}_t} \omega_{e,t} \log v_{e,t}(w), \\ g_t(w) &= (1 - \eta) \log p_t^{(L)}(w) + \eta b_t(w), \end{aligned} \quad (21)$$

$$\tilde{p}_t(w) = \frac{\exp(g_t(w))}{\sum_{u \in \mathcal{V}} \exp(g_t(u))}. \quad (22)$$

If

$$D_{\text{KL}}(\tilde{p}_t \| p_t^{(L)}) \leq \epsilon, \quad (23)$$

then $p_t^* = \tilde{p}_t$. Otherwise, the receiver constraint is active and the optimum takes the form

$$p_t^\lambda(w) = \text{softmax}\left(\frac{g_t(w) + \lambda \log p_t^{(L)}(w)}{1 + \lambda}\right), \quad (24)$$

where $\lambda \geq 0$ is found by bisection such that

$$D_{\text{KL}}(p_t^\lambda \| p_t^{(L)}) = \epsilon. \quad (25)$$

We then return $p_t^* = p_t^\lambda$. This projection preserves the receiver boundary even when the evidence coalitions favor a substantially different token distribution.

A.5 Autoregressive Generation

For open-ended generation, the decoder is applied once at each autoregressive token step. Given the current prefix, the frozen LLM produces $\{p_t^{(1)}, \dots, p_t^{(L)}\}$. The decoder computes the event set $\mathcal{E}_t$, constructs its evidence coalitions, and solves

Eq. (6). If $\mathcal{E}_t = \emptyset$, it returns $p_t^{(L)}$. Under deterministic generation, the next token is

$$y_t = \arg \max_{w \in \mathcal{V}} p_t^*(w). \quad (26)$$

The selected token is appended to the prefix and the process repeats until an end token or the maximum generation length is reached.

A.6 Fixed-Choice Instantiation

For fixed-choice tasks, let $\mathcal{A} = \{a_1, \dots, a_M\}$ denote the candidate answer continuations for a prompt context c , where $a_i = (a_{i,1}, \dots, a_{i,|a_i|})$. We run the frozen LLM in teacher-forced scoring mode and define the layerwise continuation score by the cumulative continuation log-likelihood:

$$R^{(\ell)}(a_i) = \sum_{u=1}^{|a_i|} \log p^{(\ell)}(a_{i,u} \mid c, a_{i,<u}). \quad (27)$$

We use the same fixed score-to-distribution transform for all layers, including the final-layer receiver:

$$s^{(\ell)}(a) = R^{(\ell)}(a) / \tau_{mc},$$

$$\pi^{(\ell)}(a_i) = \text{softmax}_{a \in \mathcal{A}} \left(s^{(\ell)}(a) \right)_{a_i}, \quad (28)$$

for $\ell \in \{1, \dots, L\}$, where $\tau_{mc} > 0$ is a fixed candidate-distribution temperature. No separately fitted temperature calibration or held-out receiver calibration is applied.

We then apply the event-selection, coalition-construction, and constrained-bargaining operations to $\{\pi^{(1)}, \dots, \pi^{(L)}\}$ in answer space, using $\pi^{(L)}$ as the receiver prior. This yields a corrected answer-choice distribution π^* , and the predicted answer is

$$a^* = \arg \max_{a_i \in \mathcal{A}} \pi^*(a_i). \quad (29)$$

For TruthfulQA multiple-choice evaluation, the resulting choice scores are used in the official MC1, MC2, and MC3 computations. For FACTOR, an instance is counted as correct when the gold continuation receives a corrected score no lower than every contradiction continuation.

A.7 Hyperparameters and Reporting Protocol

Table 4 summarizes the key hyperparameters of the proposed receiver-constrained evidence bargaining decoder. These hyperparameters are shared across all experiments unless otherwise specified. The

Parameter	Meaning	Value
η	Evidence mixing coefficient	0.75
ϵ	Receiver KL budget	0.1 nats
$\ell_{\min}$	First candidate internal layer	1
ρ	Relative event threshold	0.0
δ	NMS distance	1
τ_c	Persistence temperature	1.0
τ_w	Witness-pooling temperature	2.0
τ_a	Corroboration temperature	0.02 – 0.05
ζ	Numerical floor	$1e - 8$

Table 4: Selected hyperparameters of the receiver-constrained evidence bargaining decoder used in our experiments.

values of η and ϵ control the trade-off between internal evidence and the receiver prior, as well as the allowed deviation from the final-layer distribution, respectively.

A.8 Computational Complexity

After intermediate-layer vocabulary logits have been extracted, computing event statistics, coalition distributions, and the constrained solution requires

$$O(|\mathcal{I}|r|\mathcal{V}| + |\mathcal{E}_t|r_w|\mathcal{V}| + B|\mathcal{V}|) \quad (30)$$

additional operations per generated token, where B is the maximum number of bisection iterations. In practice, the dominant additional overhead arises from exposing and storing intermediate-layer vocabulary logits.

B Reproduction Details

All experiments are run with a unified evaluation codebase. The implementation supports six decoding modes: greedy, dola, end, prunedcd, sled and ours. Models are loaded with Hugging Face AutoModelForCausalLM, and tokenizers are loaded with AutoTokenizer; when a tokenizer does not define a padding token, we reuse the EOS token for padding. Unless otherwise specified, all runs use random seed 42, evaluation mode, a single beam, and deterministic decoding. All decoding methods share the same task-specific prompts, stopping criteria, and answer extraction procedures. This unified setup ensures that performance differences primarily reflect the decoding strategies.

The default backbone model paths are:

- LLM-Research/Meta-Llama-3.1-8B-Instruct
- Qwen/Qwen2.5-7B-Instruct
- LLM-Research/Mistral-7B-Instruct-v0.3

Strength	MC1	MC2	MC3	Threshold	MC1	MC2	MC3	Ratio	MC1	MC2	MC3
0	38.43	57.01	32.13	0.00	45.65	62.81	35.44	0.20	45.65	62.81	35.44
1	41.13	58.78	32.92	0.01	45.04	62.31	34.90	0.35	45.65	62.81	35.44
2	44.06	61.37	34.77	0.03	45.29	62.88	35.02	0.50	45.65	62.81	35.44
3.5	45.65	62.81	35.44	0.05	44.92	62.35	34.96	0.65	45.65	62.81	35.44
5	45.53	63.14	35.64	0.10	45.04	62.15	34.87	0.80	45.29	62.57	35.03

(a) Evidence strength

(b) Event threshold

(c) Candidate-layer ratio

Table 5: Sensitivity analysis of key hyperparameters. Performance remains stable across a broad range of event thresholds and candidate-layer ratios, while the benefit of increasing evidence strength gradually saturates.

For TruthfulQA multiple-choice evaluation, the script scores every truthful and false candidate answer and computes MC1, MC2, and MC3. For TruthfulQA generation, generated answers are passed to the benchmark rating pipeline. For FACTOR, we score the gold continuation and all contradiction continuations; a sample is marked correct if the gold continuation receives a score no lower than every contradiction continuation. For PopQA, predictions are normalized by lowercasing and removing articles and punctuation before matching against gold answer aliases. For HaluEval, outputs are parsed as yes/no labels, and ambiguous outputs are marked invalid.

C Hyperparameter Sensitivity

For sensitivity analysis, we additionally vary three implementation-level controls: evidence strength, the relative event threshold, and the candidate-layer ratio.

All experiments are conducted on TruthfulQA MC with Llama-3.1-8B-Instruct while varying one parameter at a time.

For evidence strength, increasing the value from 0 to 3.5 leads to consistent improvements across all metrics, with MC1/MC2/MC3 increasing from 38.43/57.01/32.13 to 45.65/62.81/35.44. Further increasing it to 5.0 results in marginal changes (45.53/63.14/35.64), indicating that the contribution of additional evidence gradually saturates.

For the event threshold, varying the value from 0 to 0.1 only causes minor fluctuations, with the largest changes of 0.73, 0.73, and 0.57 points for MC1, MC2, and MC3, respectively. This suggests that the event detection mechanism is relatively insensitive to the threshold within this range.

For the candidate-layer ratio, the results remain identical when increasing the ratio from 0.20 to 0.65, indicating that the selected evidence layers are stable within this interval. Increasing the ratio to 0.80 introduces only a slight degradation of

0.24–0.41 points, showing that including additional distant layers has limited impact.

Method	TFQA MC	TFQA Gen	FACTOR	PopQA	HaluEval
Greedy	4.5327	51.6813	4.9533	14.6937	21.0902
DoLa	1.4572	40.2334	0.3318	9.9384	15.9538
SLED	0.5237	9.9375	0.3212	6.5465	9.5314
END	3.1151	32.3300	3.4718	8.8635	14.1745
PruneCD	0.4521	4.9740	1.1559	5.4435	4.7202
Ours	3.0885	17.1100	0.7123	4.5033	10.5747

Table 6: Inference throughput (samples/s or tokens/s) comparison on Llama-3.1-8B-Instruct. Higher values indicate better efficiency.

D Inference Efficiency Analysis

Our method introduces additional inference overhead due to intermediate-layer evidence extraction. Table 6 reports the inference throughput across different factuality benchmarks. Compared with recent factuality-oriented decoding methods such as SLED and PruneCD, our method achieves competitive efficiency while providing stronger factuality performance.

The efficiency trade-off is task-dependent. Compared with SLED, our method is approximately $5.90\times$, $1.72\times$, $2.22\times$, and $1.11\times$ faster on TFQA MC, TFQA Gen, FACTOR, and HaluEval, respectively, while remaining slower on PopQA. Compared with PruneCD, our method achieves higher throughput on TFQA MC, TFQA Gen, and HaluEval, but incurs additional cost on FACTOR.

E Prompt Templates

We keep prompts fixed across decoding methods so that performance differences come from the decoding rule rather than prompt variation.

E.1 TruthfulQA

Interpret each question literally, and as a question about the real world; carefully research each answer, without falling prey to any common myths; and reply “I have no comment” unless you are completely certain of the answer.

Q: What is human life expectancy in the United States?

A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?

A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Which party did he belong to?

A: He belonged to the Republican Party.

Q: What is the square root of banana?

A: I have no comment.

Q: How does a telescope work?

A: Telescopes use lenses or mirrors to focus light and make objects appear closer.

Q: Where were the 1992 Olympics held?

A: The 1992 Olympics were held in Barcelona, Spain.

Q: {question}

A:

E.2 PopQA

Answer the question with a short factual phrase.

Q: What is George Washington's occupation?

A: politician

Q: What country is Paris located in?

A: France

Q: {question}

A:

E.3 HaluEval

Decide whether the answer contains hallucinated or unsupported information based only on the knowledge.

Knowledge: {knowledge}

Question: {question}

Answer: {answer}

Reply with exactly one label:

Yes = the answer contains hallucinated or unsupported information.

No = the answer is fully supported by the knowledge.

Answer:

E.4 FACTOR

FACTOR is evaluated as a continuation-ranking task rather than free-form generation. For each sample, we use the dataset prefix as the prompt and compare the log-probability scores of the gold continuation and contradiction continuations:

Prompt: {prefix}

Candidate continuation: {completion or contradiction}