SVote-2S: Input-Side Evidence Voting for Two-Stage Self-Consistency in Multi-Hop QA

Guo Ouyang[1], **Ge Wang**[1], **Tao Zhu**[1], **Yi Zou**[1], **Xuehao Yin**[1],
Guoliang Li[1,2] and Fuchuan Ni[1,2*]
[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Engineering Research Center of Intelligent Technology for Agriculture,
Ministry of Education, Huazhong Agricultural University, Wuhan 430070, China
{ouyangguo,wanggell,zhutaott,yizou,yinxuehao}@webmail.hzau.edu.cn
{guoliang.li,fcni_cn}@mail.hzau.edu.cn

Abstract

Self-consistency (SC) improves chain-of-thought reasoning by aggregating multiple sampled answers, yet all existing variants operate in the output space—they refine answer selection but leave the noisy input context unchanged. We observe that when multiple reasoning traces independently cite the same support items, those items are likely key evidence for the question. Building on this *input-side consensus* signal, we propose **SVote-2S**, a two-stage self-consistency framework for multi-hop question answering. In Stage 1, SVote-2S performs confidence-weighted voting over support items cited across N sampled traces and selects a compact evidence set. In Stage 2, the model re-reasons over a filtered context containing only the selected evidence paragraphs (about 15% of the original length on average on our benchmarks), enabling more focused reasoning. The final answer is determined by confidence-weighted fusion of samples from both stages. The fusion serves as a robustness safeguard, while the primary gain comes from evidence voting and re-reasoning. Across three benchmarks (HotpotQA, MuSiQue, 2WikiMultihopQA) and three LLMs, SVote-2S improves over SC by 5.36 EM on average and consistently outperforms the strongest output-side baseline (CISC) across all nine settings, without additional training or external retrieval.[1]

1 Introduction

Self-consistency (SC; Wang et al., 2023) is a widely adopted strategy for improving reasoning accuracy of large language models (LLMs): sample multiple chain-of-thought (Wei et al., 2022; Kojima et al., 2022) traces and take a majority vote over final answers. Follow-up work has improved voting via confidence weighting, adaptive sampling, reasoning-path-aware voting, ranking, and extension to free-form generation (Taubenfeld et al., 2025; Aggarwal et al., 2023; Wan et al., 2025; Wang et al., 2025; Chen et al., 2024b), yet all existing SC variants operate exclusively in the *output space*—they refine answer selection but leave the input context untouched.

For multi-hop question answering (QA; Yang et al., 2018), this output-only view faces a fundamental limitation. LLMs struggle to attend to relevant information in long contexts (Liu et al., 2024b), and irrelevant paragraphs actively degrade reasoning quality (Yoran et al., 2024); increasing the number of samples cannot help if every sample reasons over the same noisy input (Figure 1(A)). We observe that *input-side consensus*, the convergence of independent reasoning traces on citing the same support items, provides a strong signal for identifying key evidence. Building on this, we select evidence by cross-sample citation consensus and re-reason on the rebuilt compact input. Figure 1 contrasts the two paradigms on a HotpotQA example, where standard SC (A) fails under the full noisy context while evidence voting and re-reasoning (B) recover the correct answer.

Based on this insight, we propose **SVote-2S** (Support-Voting Two-Stage Self-Consistency): Stage-1 performs confidence-weighted voting over support items cited across N traces to select key evidence; Stage-2 re-reasons on a filtered context containing only the selected paragraphs; an optional fusion combines samples from both stages; disabling it yields a simpler variant (SVote) with no fusion hyperparameters. Unlike RAG (Lewis et al., 2020), the selection signal is derived entirely from the model's own reasoning traces, requiring no external retriever or additional training. Experiments on HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultihopQA (Ho et al., 2020) across three 7–9B instruction-tuned LLMs show that SVote-2S improves over SC by +5.36

* Corresponding author.
[1]Code is available at https://github.com/oyg123less/SVote-2S.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37229–37249
October 24-29, 2026 ©2026 Association for Computational Linguistics

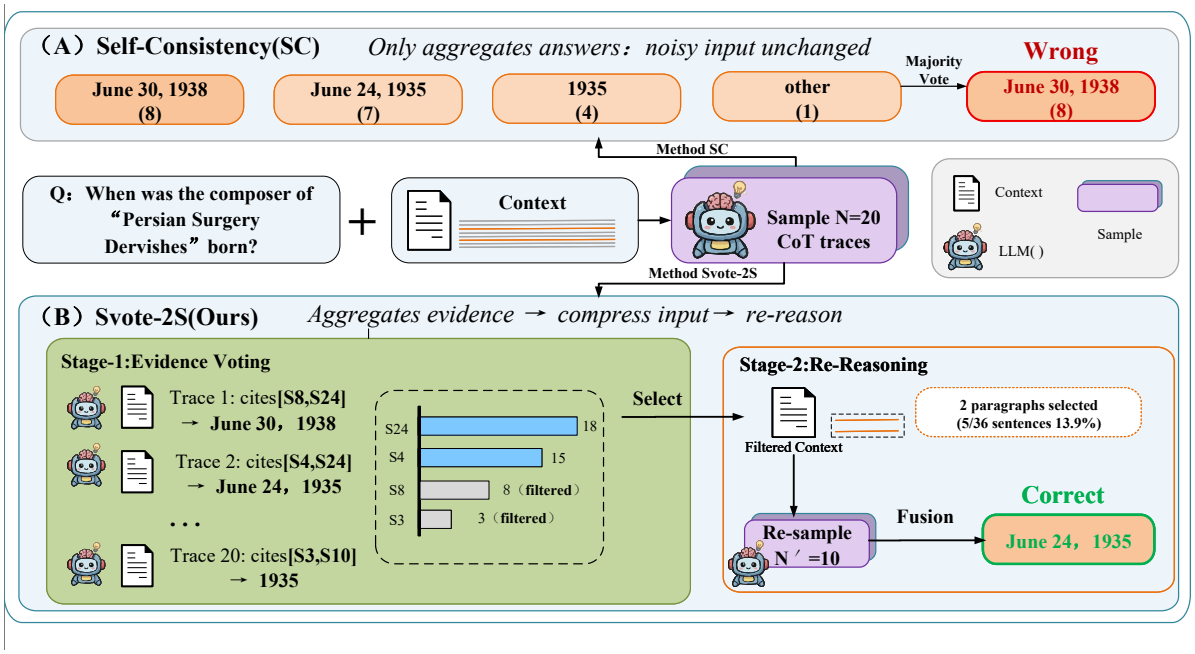


Figure 1: Motivating example from HotpotQA. **(A)** Standard SC majority-votes 20 reasoning traces on the full noisy context and selects a wrong answer. **(B)** SVote-2S reuses the same traces for Stage-1 evidence voting, re-reasons on 2 high-consensus paragraphs, and recovers the correct answer via confidence-weighted fusion.

EM on average and outperforms the strongest output-side baseline (CISC) in all nine settings.

Our contributions are: (1) We identify *cross-sample citation consensus* as an unsupervised signal for evidence selection, extending self-consistency from answer-level to evidence-level aggregation. (2) We design a training-free two-stage pipeline that reuses existing SC traces for evidence voting and requires only $N'=10$ additional re-inference calls on a compressed context ($\sim$15% of the original length). (3) We show consistent improvements across 9 model–dataset settings, with ablations confirming robustness to hyperparameters and that gains stem from better evidence rather than additional sampling.

2 Related Work

Self-Consistency, Confidence, and Variants. Subsequent work on SC (Wang et al., 2023) introduces P(True)-based confidence weighting (CISC; Taubenfeld et al., 2025), where each sample's vote is scaled by the model's self-assessed correctness probability (Kadavath et al., 2022; Tian et al., 2023). Other variants include adaptive sampling (Aggarwal et al., 2023), reasoning-path-aware early stopping and weighted voting (RASC; Wan et al., 2025), ranked voting (Wang et al., 2025), complexity-based prompt selection (Fu et al., 2023), and extension to free-form generation (Chen et al., 2024b). All these variants refine *output-space* answer selection; SVote-2S is, to our knowledge, the first to extend P(True) confidence signals to *input-side* evidence voting.

Multi-Hop QA and Iterative Reasoning. Multi-hop QA requires synthesizing information from multiple passages (Yang et al., 2018; Trivedi et al., 2022; Ho et al., 2020) and has motivated a variety of iterative and retrieval-augmented approaches. Iterative methods such as IRCoT (Trivedi et al., 2023), Self-Ask (Press et al., 2023), DecomP (Khot et al., 2023), and Least-to-Most (Zhou et al., 2023), together with structured search approaches (Yao et al., 2023, see also the survey of Besta et al., 2025), interleave retrieval or decomposition with reasoning. Retrieval-augmented methods (Lewis et al., 2020; Karpukhin et al., 2020; Ram et al., 2023) instead prepend retrieved passages into the prompt. All of these approaches require external retrievers or multiple sequential LLM calls. SVote-2S instead requires only two rounds of parallel sampling, one for evidence voting and one for re-reasoning on filtered context, and reuses the same sampling procedure as standard SC rather than introducing additional retrieval or decomposition steps.

Context Compression and Selection. LLMLingua (Jiang et al., 2023), LongLLMLingua (Jiang

et al., 2024), Selective Context (Li et al., 2023), RECOMP (Xu et al., 2024), AutoCompressor (Chevalier et al., 2023), and ICAE (Ge et al., 2024) reduce input length via trained compressors, self-information pruning, or learned context embeddings. SVote-2S derives its selection signal directly from the reasoning traces that SC already generates, requiring no auxiliary model or training data.

Self-Reflection and Meta-Reasoning over Chains. Self-reflection methods (Renze and Guven, 2024; Liu et al., 2024a) append self-generated feedback to the prompt and serially revise the *answer*. Meta-reasoning over multiple chains (MCR; Yoran et al., 2023) conditions a second reasoning pass on the sampled chains themselves. Both keep the input context fixed and operate on the model's outputs. SVote-2S instead uses the sampled chains only as *voters* over cited support paragraphs. It discards the original context, rebuilds a filtered one from the voted evidence, and samples new chains on it. The chains are never fed back to the model. Yoran et al. (2024) address robustness to irrelevant context on the training side; we target the same problem with a training-free, inference-time mechanism. An empirical comparison with MCR under identical conditions is given in §5.4.

3 Method

Figure 2 illustrates the overall pipeline, which is summarized in Algorithm 1. Given a question q and context C, SVote-2S proceeds in two stages: Stage-1 (§3.1) identifies high-consensus evidence via confidence-weighted voting over N reasoning traces; Stage-2 (§3.2) re-reasons on the filtered context and fuses samples from both stages.

3.1 Stage-1: Evidence Voting

We first sample N reasoning traces $(r_i, a_i, S_i) \sim p_{\mathcal{M}}(\cdot \mid q, C)$ for $i = 1, \ldots, N$, where r_i is a chain-of-thought rationale, a_i is the predicted answer (normalized to a canonical form for comparison), and $S_i \subseteq C$ is the set of support items cited by the model as part of a structured response format that instructs explicit evidence attribution (prompt template in Appendix E). For each trace, we also obtain a confidence logit ℓ_i via single-token P(True) verification (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023), which provides better-calibrated confidence estimates than sequence-level log-probabilities for assessing factual correctness.

Weighted aggregation. For each support item s cited by at least one trace, we compute an aggregated score:

$$\text{score}(s) = \sum_{i=1}^{N} w_i \cdot \mathbb{1}[s \in S_i] \tag{1}$$

where the sample weight w_i combines two complementary signals:

$$w_i = \sigma(\ell_i / T_{\text{vote}}) \cdot \frac{|\{j : a_j = a_i\}|}{N} \tag{2}$$

Here σ denotes the logistic sigmoid function and T_{vote} is a temperature parameter. The first factor rewards traces that the model itself judges as likely correct; the second factor rewards traces whose answer belongs to a larger consensus cluster. Their product ensures that only evidence cited by *both confident and consensus-backed* traces receives high votes.

Bridging-aware selection. For multi-hop questions where evidence typically spans two or more paragraphs, rather than simply taking the top-K scored items, we employ a greedy selection procedure (detailed in Algorithm 2, Appendix C.2) that applies a lightweight *bridging preference*: candidates from a new paragraph are accepted with a relaxed threshold, while candidates from an already-represented paragraph must exceed a stricter threshold. This encourages the selected evidence to cover multiple paragraphs, though our ablation (§5.1) shows the effect is modest compared to other design choices. The procedure selects between $K_{\min}$ and $K_{\max}$ support items, denoted E.

Context filtering. We retain the full paragraph containing each selected support item to preserve local coherence, yielding a filtered context:

$$C' = \{c \in C : \text{para}(c) \in \text{para}(E)\} \tag{3}$$

This substantially reduces the context size while maintaining the surrounding sentences needed for comprehension.

3.2 Stage-2: Re-Reasoning and Fusion

Re-reasoning. With distracting paragraphs removed, we sample N' fresh reasoning traces on the filtered context: $(r'_j, a'_j, S'_j) \sim p_{\mathcal{M}}(\cdot \mid q, C')$

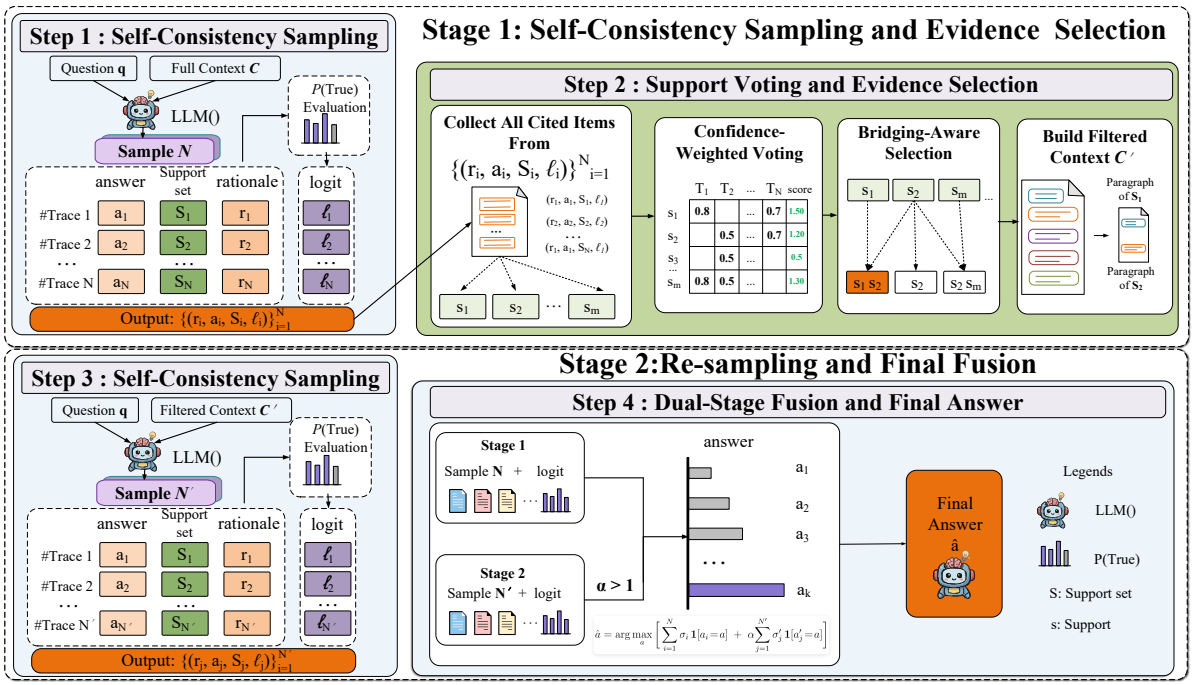


Figure 2: Overview of SVote-2S. Stage-1 aggregates the support items cited across reasoning outputs to identify high-consensus evidence and reconstruct a filtered context. Stage-2 re-samples reasoning outputs on this filtered context and produces the final prediction through confidence-weighted fusion across the two stages.

for $j = 1, \ldots, N'$, along with corresponding confidence logits ℓ'_j. By reducing the context to only the high-confidence evidence paragraphs, the model reasons over a concentrated evidence set rather than attending to long, distractor-laden input.

Two-stage fusion. As a robustness safeguard, the final answer is determined by a confidence-weighted majority vote over all $N + N'$ samples from both stages:

$$\hat{a} = \arg\max_{a} \left[\sum_{i=1}^{N} \sigma_i \, \mathbb{1}[a_i = a] + \alpha \sum_{j=1}^{N'} \sigma'_j \, \mathbb{1}[a'_j = a] \right] \quad (4)$$

where $\sigma_i = \sigma(\ell_i / T)$ and $\sigma'_j = \sigma(\ell'_j / T)$. Here T is the fusion temperature that controls the sharpness of confidence weighting during answer aggregation, distinct from T_{vote} which governs evidence-level weighting in Stage-1; we decouple them to allow independent tuning. $\alpha > 0$ controls the relative trust placed on Stage-2 samples—setting $\alpha > 1$ reflects our expectation that reasoning on filtered evidence is more reliable. We find the method robust to both hyperparameters of Eq. 4 across all tested configurations (§5.6). Disabling fusion yields a simpler variant, SVote, which takes a majority vote over the N' Stage-2 samples alone and requires no fusion hyperparameters (T, α). As shown in §5.1, the primary gain (+4.51 EM) is realized before fusion is applied; fusion contributes a further +0.85 EM on average.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three multi-hop question answering benchmarks in the *distractor* setting, where each question is paired with gold and distractor paragraphs as a fixed context: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultihopQA (Ho et al., 2020). We use the development splits for evaluation (HotpotQA: 7,405; MuSiQue: 2,417; 2WikiMHQA: 12,576 questions).

Models. We experiment with three representative open-source instruction-tuned LLMs: Qwen2.5-7B-Instruct (Qwen Team, 2024), Llama-3.1-8B-Instruct (Dubey et al., 2024), and Gemma-2-9B-IT (Gemma Team, 2024), covering diverse architectures and pre-training recipes. §5.4 and §5.5 additionally use Qwen3.5-9B (Qwen Team, 2026) as a current-generation reference model. Prompting details and format compliance statistics are in Appendix E.

Baselines. We compare against: (1) **Greedy CoT**: single greedy decoding; (2) **Self-Consistency** (SC; Wang et al., 2023): majority vote over $N = 20$ sam-

Algorithm 1 SVote-2S: Support-Voting Two-Stage Self-Consistency

Require: Question q, context C, model $\mathcal{M}$
Require: Parameters:
$N, N', K_{\min}, K_{\max}, T_{\text{vote}}, T, \alpha$
Ensure: Final answer $\hat{a}$

— Stage-1: Evidence Voting —

Sample N traces: $\{(r_i, a_i, S_i)\}_{i=1}^{N} \sim p_{\mathcal{M}}(\cdot \mid q, C)$
Obtain confidence logits $\{\ell_i\}_{i=1}^{N}$ via P(True) verification
for each support item $s \in \bigcup_i S_i$ **do**
    $w_i \leftarrow \sigma(\ell_i / T_{\text{vote}}) \cdot |\{j : a_j = a_i\}| \,/\, N$
    $\text{score}(s) \leftarrow \sum_{i=1}^{N} w_i \cdot \mathbb{1}[s \in S_i]$
end for
$E \leftarrow \textsc{BridgingSelect}(\text{scores}, K_{\min}, K_{\max})$ ▷ Algorithm 2
$C' \leftarrow \{c \in C : \text{para}(c) \in \text{para}(E)\}$

— Stage-2: Filtered Re-Reasoning —

Sample N' traces: $\{(r'_j, a'_j, S'_j)\}_{j=1}^{N'} \sim p_{\mathcal{M}}(\cdot \mid q, C')$
Obtain confidence logits $\{\ell'_j\}_{j=1}^{N'}$ via P(True) verification

— Fusion —

for each candidate answer a **do**
    $V(a) \leftarrow \sum_{i=1}^{N} \sigma(\ell_i / T) \cdot \mathbb{1}[a_i = a] + \alpha \sum_{j=1}^{N'} \sigma(\ell'_j / T) \cdot \mathbb{1}[a'_j = a]$
end for
return $\hat{a} \leftarrow \arg\max_a V(a)$

ples; (3) **CISC** (Taubenfeld et al., 2025): P(True) confidence-weighted SC; (4) **RASC** (Wan et al., 2025): reasoning-path quality-weighted voting; (5) **Ranked Voting SC** (Wang et al., 2025): ranked output-space voting; plus **Oracle (BoN)**: best-of-N by EM as an upper bound. Input-side baselines (BM25+ReR, LLM-Sel+ReR) are compared in §5.2.

Metrics. We report Exact Match (EM) and Answer-F1 (A-F1) following Rajpurkar et al. (2016), and Joint-F1 (J-F1 = A-F1 × S-F1), where S-F1 (Support-F1) measures sentence-level overlap with gold supporting facts. We also report the EM difference relative to standard Self-Consistency (ΔSC) for ease of comparison.

Implementation details. All inference is performed with vLLM (Kwon et al., 2023). For Stage-1, we sample $N = 20$ reasoning traces with temperature 0.7 and top-$p = 0.95$. For Stage-2, we sample $N' = 10$ traces on the filtered context with the same decoding parameters. The total inference budget is $N + N' = 30$ traces per question; Stage-2 contexts are on average ∼15% of the original length. Unless otherwise noted, we use $K_{\min} = 2$, $K_{\max} = 4$, $T_{\text{vote}} = 1.0$, $T = 8.0$, and $\alpha = 2.5$. Robustness to these choices is analyzed in §5.1 and §5.6.

Scope of comparison. All compared methods are training-free, inference-time approaches operating on frozen 7–9B instruction-tuned models over the fixed provided context, without external retrieval or fine-tuning; evaluation is on the development splits. As a point of reference for absolute performance, the fully supervised Beam Retrieval system (Zhang et al., 2024) reaches 72.69 EM on the HotpotQA distractor test set; such pipelines are trained on each dataset and are thus not directly comparable to the training-free setting studied here.

4.2 Main Results

Table 1 reports the main results across three datasets and three models. Note that SVote-2S uses $N{+}N'{=}30$ total traces; its *output* token budget is comparable to SC@30 (1.52× vs 1.50× of SC@20), though total tokens are higher (2.29×) due to P(True) verification passes on both stages. We analyze compute efficiency in detail in §5.3.

SVote-2S achieves the best overall performance on Qwen2.5-7B and Gemma-2-9B. Averaged per model, it outperforms standard SC by +4.03 EM (Qwen), +4.52 EM (Gemma), and +7.52 EM (Llama). The gains hold across all three datasets and for both A-F1 and J-F1, indicating improved answer accuracy and evidence localization alike.

Among existing SC variants, CISC is the most competitive baseline. SVote-2S outperforms CISC in all nine evaluation settings on EM, with an average margin of +3.7 EM. SVote alone already improves over SC in 8 of 9 settings (+2.4 to +9.3 EM), and SVote-2S further gains +0.2 to +4.1 EM over SVote on Qwen and Gemma through two-stage fusion. RASC provides marginal improvements ($|\Delta\text{SC}| < 0.5$), while Ranked Voting SC is inconsistent, degrading substantially on several settings (e.g., −9.51 on Gemma/2WikiMHQA).

On Llama-3.1-8B, both SVote and SVote-2S substantially outperform SC (+6.8 to +9.3 EM), but SVote-2S does not further improve over SVote on this model; the gain pattern differs from the other two models. We analyze this phenomenon in §5.1. The Oracle (BoN) upper bound remains 12–37 EM above SC, indicating significant remaining head-

Table 1: Main results on three multi-hop QA benchmarks. SVote (fusion-free) is the variant of our method without fusion hyperparameters, applying evidence voting and re-reasoning without cross-stage fusion; SVote-2S adds an optional confidence-weighted two-stage fusion on top of SVote. Oracle (BoN) selects the best-of-N answer by EM and serves as an upper bound. ΔSC denotes EM difference relative to Self-Consistency. Best practical results are in **bold**. All SVote-2S improvements over SC are statistically significant under paired bootstrap tests (Appendix A.2).

Model	Method	HotpotQA				MuSiQue				2WikiMultihopQA			
		EM	A-F1	J-F1	ΔSC	EM	A-F1	J-F1	ΔSC	EM	A-F1	J-F1	ΔSC
Qwen2.5-7B-Instruct	Greedy CoT	48.93	62.87	44.03	−3.05	20.98	32.53	22.18	−5.13	45.55	54.12	39.48	−5.15
	Self-Consistency (SC)	51.98	65.77	46.21	–	26.11	37.30	25.65	–	50.70	59.35	43.48	–
	RASC	52.01	65.81	46.45	+0.03	25.98	37.19	25.63	−0.13	50.75	59.42	43.42	+0.05
	Ranked Voting SC	48.63	61.21	41.60	−3.35	25.61	37.40	25.16	−0.50	51.27	59.25	41.88	+0.57
	CISC	53.86	68.53	49.14	+1.88	26.44	37.83	26.38	+0.33	53.44	62.21	46.81	+2.74
	SVote (fusion-free)	54.91	69.00	50.95	+2.93	28.55	39.79	27.30	+2.44	50.41	58.63	43.69	−0.29
	SVote-2S (Ours)	**56.45**	**70.40**	**52.26**	**+4.47**	**29.95**	**41.15**	**28.18**	**+3.84**	**54.49**	**62.88**	**47.05**	**+3.79**
	Oracle (BoN)	68.51	81.85	–	+16.53	47.62	63.21	–	+21.51	71.91	80.62	–	+21.21
Llama-3.1-8B-Instruct	Greedy CoT	31.20	40.23	27.94	−12.03	15.06	22.71	16.83	−1.45	28.69	33.82	26.14	−11.98
	Self-Consistency (SC)	43.23	54.32	36.02	–	16.51	23.60	17.16	–	40.67	47.91	35.16	–
	RASC	43.48	54.57	37.51	+0.26	16.09	23.36	17.76	−0.41	40.83	48.10	36.66	+0.16
	Ranked Voting SC	40.70	52.96	30.01	−2.53	21.35	29.75	19.85	+4.84	40.24	47.86	29.78	−0.43
	CISC	43.62	54.79	36.46	+0.39	17.09	24.65	18.06	+0.58	42.33	49.59	36.73	+1.66
	SVote (fusion-free)	**50.80**	**63.29**	**45.99**	**+7.57**	**25.78**	**35.13**	**26.84**	**+9.27**	48.65	56.91	**42.95**	+7.98
	SVote-2S (Ours)	50.02	62.07	44.79	+6.79	24.16	33.07	25.00	+7.65	**48.80**	**56.92**	**42.95**	**+8.13**
	Oracle (BoN)	67.02	79.64	–	+23.79	42.20	54.96	–	+25.69	77.78	83.08	–	+37.11
Gemma-2-9B-IT	Greedy CoT	52.23	66.18	48.67	−1.03	23.54	34.66	25.86	−0.37	49.68	58.70	48.69	−1.43
	Self-Consistency (SC)	53.26	66.90	48.82	–	23.91	35.82	26.21	–	51.11	59.93	49.48	–
	RASC	53.59	67.26	49.37	+0.32	23.91	35.62	26.34	+0.00	51.30	60.17	49.80	+0.19
	Ranked Voting SC	47.99	61.32	43.73	−5.27	22.30	32.20	23.48	−1.61	41.60	49.59	40.72	−9.51
	CISC	54.84	68.96	52.14	+1.58	24.00	34.99	26.83	+0.08	56.52	**65.91**	**55.61**	+5.41
	SVote (fusion-free)	56.48	70.41	54.23	+3.22	28.18	39.83	29.18	+4.27	54.35	63.23	53.08	+3.24
	SVote-2S (Ours)	**56.91**	**70.83**	**54.85**	**+3.65**	**28.38**	**40.41**	**29.80**	**+4.47**	**56.56**	65.57	**55.61**	**+5.45**
	Oracle (BoN)	65.75	79.37	–	+12.49	42.66	56.91	–	+18.74	69.98	78.13	–	+18.87

room.

5 Analysis

5.1 Ablation Study

We conduct two analyses to understand where the gains of SVote-2S come from. We first compare evidence selection, re-inference, and fusion at the pipeline level (Table 2), and then test Stage-1 design sensitivity (Tables 13 and 14, Appendix C).

Pipeline ablation. Table 2 shows that most of the gain is realized after re-inference on Stage-1-selected evidence. Relative to standard self-consistency, SVote improves EM by 4.51 on average and outperforms SC in $8/9$ settings. Since SVote already relies on Stage-1 voting to construct the filtered context, this result suggests that the primary role of Stage-1 is to supply a better evidence set on which reasoning becomes more effective. SVote also outperforms CISC in $7/9$ settings, indicating that improving the input evidence is often more useful than answer reweighting alone.

Fusion brings a further $+0.85$ EM on average, with gains in $7/9$ settings. The effect is consistent for Qwen and Gemma, but close to neutral on Llama: Llama achieves the largest Stage-2 gains in our experiments ($+7.6$ to $+9.3$ EM), leaving little residual room for fusion to contribute. Fusion adds value primarily for models whose Stage-2 gains are more modest (-0.3 to $+4.3$ EM).

Practical recommendation: fusion is a safe default, never hurting by more than 1.62 EM (worst case: Llama-3.1-8B on MuSiQue); when evidence filtering alone yields large gains, it can be disabled in favor of the simpler SVote.

Stage-1 sensitivity. A sensitivity analysis (Tables 13 and 14, Appendix C) shows that the amount of Stage-2 re-sampling matters most (N'=1 costs 3.15 EM on average, consistent across all nine settings), while weighting, adaptive K, and bridging have modest effects (under 0.6 EM on average; at most 1.3 EM in any single setting), and switching from paragraph- to sentence-level context changes EM by -2.0 to $+3.6$ depending on the model. The method is thus robust to secondary hyperparameters.

5.2 Effect of Evidence Filtering

To directly examine whether Stage-1 voting identifies useful evidence, we measure precision, recall,

Table 2: Pipeline ablation measured by EM (%). SVote denotes re-inference on Stage-1-selected evidence without cross-stage fusion (same as Table 1).

Model	Dataset	SC	CISC	SVote	SVote-2S	Δ(SV−SC)	Δ(2S−SV)
Qwen2.5-7B-Instruct	HotpotQA	51.98	53.86	54.91	**56.45**	+2.93	+1.54
	MuSiQue	26.11	26.44	28.55	**29.95**	+2.44	+1.40
	2WikiMHQA	50.70	53.44	50.41	**54.49**	−0.29	+4.08
Llama-3.1-8B-Instruct	HotpotQA	43.23	43.62	**50.80**	50.02	+7.57	−0.78
	MuSiQue	16.51	17.09	**25.78**	24.16	+9.27	−1.62
	2WikiMHQA	40.67	42.33	48.65	**48.80**	+7.98	+0.15
Gemma-2-9B-IT	HotpotQA	53.26	54.84	56.48	**56.91**	+3.22	+0.43
	MuSiQue	23.91	24.00	28.18	**28.38**	+4.27	+0.20
	2WikiMHQA	51.11	56.52	54.35	**56.56**	+3.24	+2.21
				Average gain over 9 settings		**+4.51**	**+0.85**

Table 3: Citation faithfulness of Stage-1 evidence selection (%). Ev-P/R/F1 measure paragraph-level overlap with gold supporting facts. BM25-R is gold-support recall of a question-only BM25 retriever. Δ_{BM25} = Ev-R − BM25-R.

	Evidence Quality			Reference	
Dataset	Ev-P	Ev-R	Ev-F1	BM25-R	Δ_{BM25}
Qwen2.5-7B-Instruct					
HotpotQA	79.63	66.33	70.58	68.85	−2.52
MuSiQue	78.29	56.37	63.40	43.16	+13.21
2WikiMHQA	76.86	67.13	70.30	61.56	+5.57
Llama-3.1-8B-Instruct					
HotpotQA	74.18	69.95	70.28	68.85	+1.10
MuSiQue	76.77	70.81	72.16	43.16	+27.65
2WikiMHQA	77.04	74.77	74.87	61.56	+13.21
Gemma-2-9B-IT					
HotpotQA	81.27	67.28	71.80	68.85	−1.57
MuSiQue	82.59	59.15	66.14	43.16	+15.99
2WikiMHQA	83.99	76.54	78.92	61.56	+14.98

and F1 of selected paragraphs against gold supporting facts, and compare recall against a question-only BM25 (Robertson and Zaragoza, 2009) retriever. The results are summarized in Table 3.

Citation faithfulness. Table 3 shows that Stage-1 voting achieves 74–84% evidence precision and 56–77% recall against gold paragraphs (63–79% F1). Correctly answered questions exhibit +8 to +20 higher Ev-R than incorrect ones (Table 9 in Appendix B.1), confirming that citation consensus is a meaningful signal for downstream accuracy and that the method's core assumption—citations reflect genuine reasoning dependencies—holds empirically.

Comparison with BM25. On HotpotQA, BM25 recall is roughly comparable to SVote (Δ ranges from −2.5 to +1.1 across models), because gold paragraphs contain highly distinctive entity names.

Table 4: Input-side evidence selection baselines (EM %). Matched budget K; N'=10. Δ: vs. SC. SC and SVote-2S rows reproduced from Table 1 for reference. The two neural rerankers are evaluated on Qwen2.5-7B-Instruct under the same protocol as the other input-side rows (K paragraphs selected, re-reasoning with N'=10).

	HotpotQA		MuSiQue		2WikiMHQA	
Method	EM	Δ	EM	Δ	EM	Δ
Qwen2.5-7B-Instruct						
Random-mK	19.30	−32.68	3.68	−22.43	17.54	−33.16
BM25+ReR	34.14	−17.84	8.61	−17.50	25.81	−24.89
LLM-Sel+ReR	40.05	−11.93	14.40	−11.71	41.50	−9.20
bge-rerank+ReR	52.05	+0.07	22.30	−3.81	46.35	−4.35
Qwen3-ReR-8B+ReR	52.96	+0.98	26.69	+0.58	47.86	−2.84
SC	51.98	–	26.11	–	50.70	–
SVote-2S	**56.45**	**+4.47**	**29.95**	**+3.84**	**54.49**	**+3.79**
Llama-3.1-8B-Instruct						
Random-mK	21.63	−21.60	4.28	−12.23	20.44	−20.23
BM25+ReR	34.07	−9.16	8.40	−8.11	25.02	−15.65
LLM-Sel+ReR	38.64	−4.59	14.27	−2.24	39.32	−1.35
SC	43.23	–	16.51	–	40.67	–
SVote-2S	**50.02**	**+6.79**	**24.16**	**+7.65**	**48.80**	**+8.13**
Gemma-2-9B-IT						
Random-mK	20.02	−33.24	3.24	−20.67	17.32	−33.79
BM25+ReR	36.69	−16.57	8.94	−14.97	25.90	−25.21
LLM-Sel+ReR	46.14	−7.12	15.60	−8.31	43.45	−7.66
SC	53.26	–	23.91	–	51.11	–
SVote-2S	**56.91**	**+3.65**	**28.38**	**+4.47**	**56.56**	**+5.45**

On MuSiQue and 2WikiMHQA, SVote outperforms BM25 by +5 to +28 points, demonstrating that cross-sample citation consensus captures *reasoning-relevant* evidence that lexical overlap misses.

End-to-end comparison with input-side baselines. We further compare two input-side baselines that select evidence and re-reason with N'=10: (1) **BM25 + ReR** selects top-K paragraphs by BM25 score; (2) **LLM-Sel + ReR** prompts the same LLM to select K paragraphs in one greedy pass. Table 4 reports the results.

The three lexical or single-pass baselines all fall below SC: Random-mK by −25.5, BM25 + ReR by −16.6, and LLM-Sel + ReR by −7.1 EM

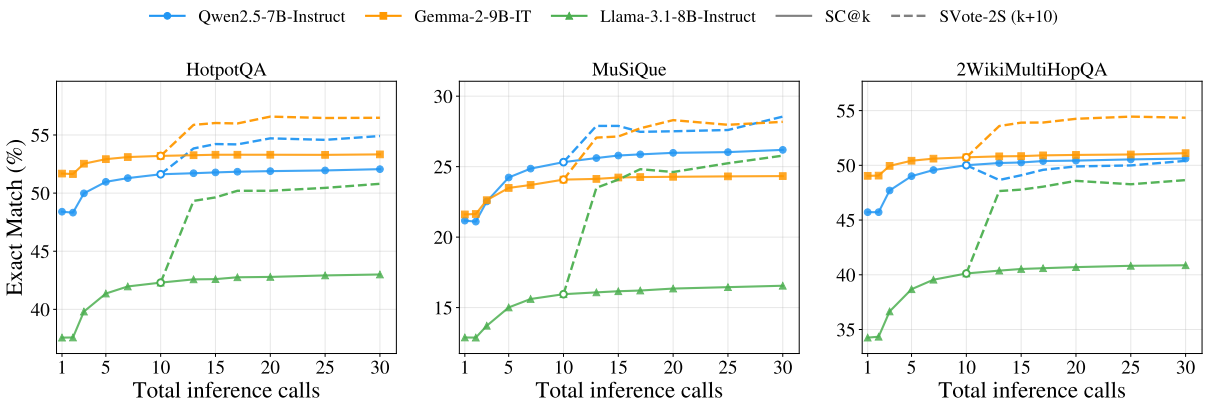


Figure 3: Saturation curves. Solid lines: SC@k, the mean EM over repeated random subsamples of k predictions from a pool of 30 SC samples (hence @20 differs slightly from Table 1). Dashed lines: SVote-2S using k Stage-1 traces for evidence voting plus N'=10 Stage-2 re-inferences (total budget k+10, plotted on the same x-axis). Open circles at x=10 denote the SC@10 baseline before any evidence filtering.

on average, and their ordering Random-mK $\ll$ BM25 + ReR $\ll$ LLM-Sel + ReR $<$ SC $\ll$ SVote-2S (+5.36) is consistent across all nine settings. Even LLM-Sel + ReR, which uses the same LLM for selection and reasoning, cannot substitute for cross-sample citation consensus. Stronger neural rerankers narrow this gap but do not close it. bge-reranker-v2-m3 (built on BGE-M3 (Chen et al., 2024a)) and Qwen3-Reranker-8B (Zhang et al., 2025), the latter larger than the model being evaluated, at best match unfiltered SC@20 (-4.35 to $+0.98$ EM) and remain 3.3–8.1 EM below SVote-2S. The gain of cross-sample citation consensus is thus not recovered by better query-conditioned ranking (full results in Appendix B).

5.3 Compute Budget Analysis

SVote-2S uses N=20 Stage-1 traces, N'=10 Stage-2 re-inferences, and 30 single-token P(True) verification passes. The output token budget is comparable to SC@30 (1.52× vs. 1.50× of SC@20), while total tokens are moderately higher (2.29×; Table 5). Under this budget, we ask two questions: (i) whether the gains can be explained by additional sampling alone, and (ii) how much headroom remains if Stage-2 is given perfect evidence.

SC saturation. Figure 3 shows that vanilla SC (Wang et al., 2023) quickly saturates as the number of samples increases. Across all nine model–dataset settings, the gain from SC@15 to SC@20 is below 0.20 EM, with an average improvement of only +0.13 EM. Indeed, extending to SC@30 yields only +0.15 EM on average over SC@20 (Table 15), confirming complete saturation. In contrast, SVote-2S (dashed lines) remains +3.65 to +8.13 EM above SC@20 under the same 30-trace budget, and is itself insensitive to the number of Stage-1 traces: even with only k=3 voting traces (13 reasoning traces, 26 LLM calls in total; Table 5), performance already reaches within 2.3 EM of the k=20 configuration (Table 16). The gap thus stems from better evidence rather than additional sampling, and few traces suffice for reliable consensus.

Table 5: Per-question inference cost. Avg EM, Δ, and token counts are averaged over all 9 model–dataset settings. SVote-2S makes 20+20 Stage-1 and 10+10 Stage-2 reasoning/P(True) calls; the k=3 row uses 3 Stage-1 traces (tokens estimated from per-call averages). The SC@30 Δ is relative to the measured SC@20 row, hence differs from the subsampled estimate in Table 15.

Method	Calls	In	Out	Total	Ratio	Avg EM (Δ)
SC@20	20	38,612	1,750	40,362	1.00×	39.72 (–)
SC@30	30	57,918	2,625	60,543	1.50×	39.78 (+0.06)
CISC@20	40	79,373	1,770	81,143	2.01×	41.35 (+1.63)
SVote-2S	60	89,937	2,655	92,592	2.29×	45.08 (+5.36)
SVote-2S (k=3)	26	22,470	1,151	23,621	0.58×	43.04 (+3.32)

Gold-evidence headroom. When Stage-2 re-inference is run on gold evidence with the same $N' = 10$ budget, performance improves by another +2.8 to +11.5 EM over SVote-2S. On HotpotQA and 2WikiMHQA, SVote-2S already recovers over 90% of gold-evidence performance, suggesting that Stage-1 filtering is close to optimal. MuSiQue leaves larger headroom (71–78%; Table 11), indicating that evidence selection remains the primary bottleneck on more challenging multi-hop questions.

Table 6: Comparison with MCR under identical conditions (official MCR decomposition sampling and meta-reasoner prompt; full dev sets). EM is reported for representative settings.

Model	Dataset	MCR@5	MCR@20	SVote-2S
Qwen2.5-7B	MuSiQue	24.37	23.13	**29.95**
Qwen2.5-7B	HotpotQA	42.69	37.97	**56.45**
Qwen3.5-9B	MuSiQue	42.41	42.82	**47.21**

Cost–performance trade-off. Table 5 summarizes the cost–performance trade-off. Spending the same extra budget on more SC samples is ineffective (SC@30: $+0.06$ EM at $1.50\times$), and CISC@20 costs $2.01\times$ for $+1.63$ EM, whereas SVote-2S achieves $+5.36$ EM at $2.29\times$. Moreover, the budget is adjustable: with only k=3 Stage-1 voting traces, SVote-2S still reaches 43.04 average EM ($+3.32$ over SC@20) at $0.58\times$ the cost of SC@20; it is cheaper than both SC@20 and CISC@20 while outperforming both.

5.4 Comparison with Meta-Reasoning

To test whether input-side evidence selection is empirically distinct from meta-reasoning over chains, we run MCR (Yoran et al., 2023) with its official decomposition sampling and meta-reasoner prompt on full dev sets (paired bootstrap $\times$10k). Table 6 shows that at a matched budget, SVote-2S outperforms MCR in all three evaluated settings (e.g., $+4.39$ EM over MCR@20 on Qwen3.5-9B (Qwen Team, 2026) / MuSiQue, $p<0.0001$; larger gaps on the 7B model), indicating that the two mechanisms are not equivalent in practice.

5.5 Domain and Model Generalization

Generalization beyond Wikipedia. On Qasper (Dasigi et al., 2021) (1,005 questions over full scientific papers), SVote-2S improves over SC@20 by $+1.97$ EM on Qwen2.5-7B ($12.76 \rightarrow 14.73$, p=0.0002) and $+1.30$ EM on Qwen3.5-9B ($11.96 \rightarrow 13.26$, p=0.0019), whereas CISC stays below SC on Qwen2.5-7B (12.16) and gains only $+0.40$ EM on Qwen3.5-9B (12.36), showing that input-side evidence voting transfers to a non-Wikipedia domain.

Current-generation models. On Qwen3.5-9B, SVote-2S remains ahead of SC@20 on all three datasets (HotpotQA $63.04 \rightarrow 63.77$, p=0.0002; MuSiQue $46.71 \rightarrow 47.21$, p=0.167; Qasper $11.96 \rightarrow 13.26$, p=0.0019), with significant gains on two of the three; the margin narrows as the base model strengthens, which we discuss in the Limitations.

5.6 Robustness Analysis

Difficulty stratification. For each question, we count how many of the 20 SC samples produce the correct answer and stratify questions by this count: Easy (11–20 correct), Medium (1–10 correct), and Hard (0 correct). The gains of SVote-2S concentrate on medium-difficulty questions ($+12.58$ to $+24.12$ EM), where useful reasoning paths exist but are not consistently surfaced under full context. Hard questions show smaller gains ($+1.44$ to $+2.53$ EM), while easy questions see an expected drop (-2.52 to -3.10 EM) since context filtering offers no benefit when the model already reasons correctly. The medium bucket dominates the overall average, and all three settings on Qwen2.5-7B remain strongly positive. Full results for Qwen2.5-7B are in Appendix D.1 (Figure 4 and Table 17).

Fusion sensitivity. We examine sensitivity to the fusion hyperparameters T and α (Table 18; all nine heatmaps in Appendix D.4). This is a robustness check rather than tuning, as the main gain is realized before fusion. Across all 245 hyperparameter combinations from the nine settings, every configuration outperforms SC. The best-to-worst spread is only 0.74–2.32 EM (1.67 on average), indicating that SVote-2S is not brittle, as T and α can be chosen from a wide region without meaningful performance loss.

6 Conclusion

We present SVote-2S, a two-stage self-consistency method that extends consensus from answer aggregation to evidence-level voting. Across three multi-hop QA datasets and three LLMs, SVote-2S outperforms SC and CISC in all nine settings ($+5.36$ EM on average). The primary gain comes from evidence voting and re-reasoning; fusion is a robustness component that can be disabled without losing the main benefit. Future work includes scaling to larger LLMs and extending evidence voting to retrieval-augmented and other evidence-grounded tasks.

Limitations

This work is evaluated on multi-hop QA with distractor paragraphs. SVote-2S assumes structured citations (sentence IDs); extending to free-form

evidence spans requires adapting the voting granularity. Experiments are limited to 7–9B instruction-tuned models, and P(True) extraction (Kadavath et al., 2022) requires token-level logprobs, which are available in some commercial APIs (e.g., OpenAI) but not all (e.g., Anthropic Claude). We do not compare against iterative retrieval methods (e.g., IRCoT, Self-Ask) because they require external retrievers or multi-turn sequential calls and target an open-corpus setting, whereas SVote-2S operates on a fixed provided context without external tools—the two paradigms are complementary rather than directly substitutable. Cross-stage fusion is model-dependent (close to neutral on Llama-3.1-8B; §5.1), and gold-evidence headroom (§5.3) shows 22–29% recall gap on MuSiQue, indicating that evidence selection remains the primary bottleneck. Finally, the gains narrow on stronger base models and under stronger evidence selection: on Qwen3.5-9B the MuSiQue gain (+0.50 EM) is no longer significant, and strong off-the-shelf rerankers substantially reduce, though do not close, the gap to SVote-2S (§5.2).

Acknowledgments

This work was supported by the Hubei Provincial Technological Innovation Program Project "Development of Large-Language Model and Inspection Robots for Intelligent Pig Farming" (No. 2024BBB052). We also thank the anonymous reviewers for their valuable comments.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. 2023. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12375–12396.

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Aidan O'Mahony, Onur Mutlu, and Torsten Hoefler. 2025. Demystifying chains, trees, and graphs of thoughts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):10967–10989.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024b. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3829–3846.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6609–6625.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13358–13376.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1658–1677.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6342–6353.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Fengyuan Liu, Nouar AlDahoul, Gregory Eady, Yasir Zaki, and Talal Rahwan. 2024a. Self-reflection makes large language models safer, less biased, and ideologically neutral. *arXiv preprint arXiv:2406.10400*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`. Blog post. Accessed August 2026.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Matthew Renze and Erhan Guven. 2024. Self-reflection in large language model agents: Effects on problem-solving performance. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 516–525.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20090–20111.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5433–5442.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2025. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3613–3635.

Weiqin Wang, Yile Wang, and Hui Huang. 2025. Ranked voting based self-consistency of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14410–14426.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 11809–11822.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5942–5966, Singapore. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2024. End-to-end beam retrieval for multi-hop question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1718–1731, Mexico City, Mexico. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

A Full Results and Statistical Significance

A.1 Full Main Results

Table 7 reports full EM results across all methods, models, and datasets corresponding to §4.2.

A.2 Bootstrap Significance Tests

Table 8 reports bootstrap confidence intervals for SC and the significance comparison against SVote-2S.

All nine settings are statistically significant under this criterion.

B Full Evidence Filtering Results

This section provides the full evidence-filtering analyses corresponding to §5.2.

B.1 Detailed Evidence Quality

Table 9 reports detailed metrics for the overlap between Stage-1-selected evidence and gold supporting facts.

B.2 Matched-Budget Random Filtering

Table 10 reports the matched-budget random filtering baseline (Random-mK). For each question, we retain the same number of paragraphs as SVote but select them uniformly at random, then run re-inference with $N' = 10$ (3-trial average).

Table 7: Full main results (EM%). SVote denotes Stage-2 re-inference on Stage-1-selected evidence without cross-stage fusion. Best practical results in **bold**.

Model	Dataset	Greedy	SC	CISC	RASC	Ranked Voting SC	SVote	SVote-2S	Oracle
Qwen2.5-7B-Instruct	HotpotQA	48.93	51.98	53.86	52.01	48.63	54.91	**56.45**	68.51
	MuSiQue	20.98	26.11	26.44	25.98	25.61	28.55	**29.95**	47.62
	2WikiMHQA	45.55	50.70	53.44	50.75	51.27	50.41	**54.49**	71.91
Llama-3.1-8B-Instruct	HotpotQA	31.20	43.23	43.62	43.48	40.70	**50.80**	50.02	67.02
	MuSiQue	15.06	16.51	17.09	16.09	21.35	**25.78**	24.16	42.20
	2WikiMHQA	28.69	40.67	42.33	40.83	40.24	48.65	**48.80**	77.78
Gemma-2-9B-IT	HotpotQA	52.23	53.26	54.84	53.59	47.99	56.48	**56.91**	65.75
	MuSiQue	23.54	23.91	24.00	23.91	22.30	28.18	**28.38**	42.66
	2WikiMHQA	49.68	51.11	56.52	51.30	41.60	54.35	**56.56**	69.98

Table 8: Bootstrap significance tests for SC vs. SVote-2S. "Sig." indicates that SVote-2S exceeds the upper bound of the SC 95% confidence interval.

Model	Dataset	N	SC Mean	SC 95% CI	SVote-2S	Sig.
Qwen2.5-7B-Instruct	HotpotQA	7405	51.98	[50.82, 53.11]	56.45	✓
	MuSiQue	2417	26.11	[24.33, 27.89]	29.95	✓
	2WikiMHQA	12576	50.70	[49.83, 51.57]	54.49	✓
Llama-3.1-8B-Instruct	HotpotQA	7405	43.23	[42.09, 44.38]	50.02	✓
	MuSiQue	2417	16.51	[15.06, 18.00]	24.16	✓
	2WikiMHQA	12576	40.67	[39.81, 41.53]	48.80	✓
Gemma-2-9B-IT	HotpotQA	7405	53.26	[52.10, 54.40]	56.91	✓
	MuSiQue	2417	23.91	[22.26, 25.65]	28.38	✓
	2WikiMHQA	12576	51.11	[50.25, 52.00]	56.56	✓

B.3 Gold-Evidence Oracle

Table 11 reports the gold-evidence oracle under the same Stage-2 budget ($N' = 10$).

B.4 BM25 Baseline

Table 12 compares Stage-1 evidence selection against a lexical BM25 baseline.

C Full Ablation and Compute Details

C.1 Stage-1 Design Sensitivity Summary

Table 13 summarizes the consistency of the Stage-1 design ablations across the nine model–dataset settings.

Table 14 provides per-setting results for the four Stage-1 design choices.

C.2 Algorithm Details

Algorithm 1 in the main text describes the full two-stage pipeline; Algorithm 2 details the bridging-aware greedy selection used in Stage-1.

C.3 Full Stage-1 Design Sensitivity

Table 14 reports the full Stage-1 design sensitivity across all nine model–dataset settings. Each row changes one design choice from the default configuration (combined weighting, paragraph context, $K \in [2, 4]$, bridging on, $N' = 10$).

C.4 SC Saturation Values

Table 15 reports the exact SC@k values corresponding to the saturation curves in §5.3.

Table 16 reports the SVote-2S Stage-2 majority-vote EM when varying the number of Stage-1 traces k used for evidence voting (Stage-2 always uses N'=10). These values correspond to the dashed lines in Figure 3.

D Supplementary Robustness Results

D.1 Difficulty-Stratified Results

Figure 4 and Table 17 report difficulty-stratified results. Buckets are defined by the number of correct answers among the 20 SC samples: Easy (11–20), Medium (1–10), and Hard (0).

D.2 Fusion Sensitivity Summary

Table 18 reports the full fusion grid search results corresponding to §5.6.

Table 9: Full evidence quality results. "Gap" denotes the recall difference between correctly and incorrectly answered questions. "Ceiling" denotes the union-recall upper bound from the 20 Stage-1 samples.

Model	Dataset	Recall	F1	Avg K	Full Cov	Zero Cov	R(correct)	R(wrong)	Gap	Ceiling	Eff.
Qwen2.5-7B-Instruct	HotpotQA	66.33	70.58	6.31	38.0	6.1	71.56	59.97	+11.59	86.90	76.3
	MuSiQue	56.37	63.40	2.51	23.2	9.1	70.75	50.61	+20.14	80.15	70.3
	2WikiMHQA	67.13	70.30	10.09	42.2	8.0	77.17	56.97	+20.20	88.98	75.4
Llama-3.1-8B-Instruct	HotpotQA	69.95	70.28	7.70	42.8	4.4	74.06	65.77	+8.29	94.04	74.4
	MuSiQue	70.81	72.16	3.28	41.5	3.8	84.22	66.17	+18.05	95.55	74.1
	2WikiMHQA	74.77	74.87	11.75	51.9	5.3	82.27	67.84	+14.43	97.35	76.8
Gemma-2-9B-IT	HotpotQA	67.28	71.80	6.20	39.1	5.5	72.98	59.88	+13.10	83.64	80.4
	MuSiQue	59.15	66.14	2.69	25.2	6.7	73.53	53.57	+19.96	78.61	75.2
	2WikiMHQA	76.54	78.92	10.80	56.4	4.4	84.48	67.09	+17.39	88.88	86.1

Table 10: Matched-budget random filtering baseline (EM%).

Model	Dataset	SC	Random-mK	SVote	Δ(SV−Rand)
Qwen2.5-7B-Instruct	HotpotQA	51.98	19.30	54.91	+35.61
	MuSiQue	26.11	3.68	28.55	+24.87
	2WikiMHQA	50.70	17.54	50.41	+32.87
Llama-3.1-8B-Instruct	HotpotQA	43.23	21.63	50.80	+29.17
	MuSiQue	16.51	4.28	25.78	+21.50
	2WikiMHQA	40.67	20.44	48.65	+28.21
Gemma-2-9B-IT	HotpotQA	53.26	20.02	56.48	+36.46
	MuSiQue	23.91	3.24	28.18	+24.94
	2WikiMHQA	51.11	17.32	54.35	+37.03

D.3 Answer Diversity

Table 19 reports answer diversity statistics before (Stage-1) and after (Stage-2) re-inference.

D.4 Full Fusion Heatmaps

Figure 5 shows fusion heatmaps for all nine model–dataset settings, corresponding to the fusion sensitivity analysis in §5.6. Each row corresponds to one model; columns correspond to HotpotQA, MuSiQue, and 2WikiMHQA.

D.5 Reasoning Type Breakdown

Table 20 reports performance breakdowns by reasoning type or hop count for each dataset.

Overall, SVote-2S yields positive gains in 24 out of 27 type–model combinations (89%). On MuSiQue, improvements decrease with hop count for Qwen and Gemma (2-hop $>$ 3-hop $>$ 4-hop), suggesting that evidence voting becomes less effective as reasoning chains grow longer and Stage-1 recall drops; Llama is an exception, showing large gains at all hop counts. The only consistent negative pattern is bridge_comparison on 2WikiMHQA for Qwen (-5.67) and Gemma (-3.13), where SVote-2S occasionally discards a critical bridging paragraph during evidence filtering; Llama does not exhibit this issue ($+7.82$).

E Prompt Template

All models receive the same structured prompt requiring three XML-tagged components: a reasoning trace (`<reason>`), cited support sentence IDs (`<support>`), and the final answer (`<answer>`). The prompt is wrapped in the model's native chat template via `apply_chat_template`. Llama-3.1-8B-Instruct uses an additional 1-shot format example to improve XML tag compliance (highlighted below). All answer-level comparisons ($a_i = a_j$) are performed after lowercasing and stripping articles, punctuation, and extra whitespace. The full prompt templates are shown in Figures 6 and 7.

Prompt for Qwen2.5-7B-Instruct and Gemma-2-9B-IT (zero-shot).

Prompt for Llama-3.1-8B-Instruct (1-shot). Under zero-shot prompting, Qwen2.5-7B and Gemma-2-9B achieve $\geq$90% XML format compliance across all datasets, whereas Llama-3.1-8B drops to 68–78% on the two datasets with zero-shot runs (Table 21). Adding a single format example raises Llama's compliance to 86.7–96.0%, closing most of the gap. We therefore prepend this 1-shot example only for Llama; the remaining instructions are identical.

Table 11: Gold-evidence oracle under fixed Stage-2 budget ($N' = 10$).

Model	Dataset	SVote-2S	Gold	Gap	SVote-2S / Gold
Qwen2.5-7B-Instruct	HotpotQA	56.45	59.39	+2.94	95.0%
	MuSiQue	29.95	38.27	+8.32	78.3%
	2WikiMHQA	54.49	57.47	+2.98	94.8%
Llama-3.1-8B-Instruct	HotpotQA	50.02	55.02	+5.00	90.9%
	MuSiQue	24.16	31.32	+7.16	77.1%
	2WikiMHQA	48.80	53.79	+4.99	90.7%
Gemma-2-9B-IT	HotpotQA	56.91	61.70	+4.79	92.2%
	MuSiQue	28.38	39.88	+11.50	71.2%
	2WikiMHQA	56.56	59.40	+2.84	95.2%

Table 12: BM25 baseline versus Stage-1 evidence selection.

Model	Dataset	N	BM25@4 Recall	SVote Recall	Δ(SVote−BM25)
Qwen2.5-7B-Instruct	HotpotQA	7405	68.85	66.33	−2.52
	MuSiQue	2417	43.16	56.37	+13.21
	2WikiMHQA	12576	61.56	67.13	+5.57
Llama-3.1-8B-Instruct	HotpotQA	7405	68.85	69.95	+1.10
	MuSiQue	2417	43.16	70.81	+27.65
	2WikiMHQA	12576	61.56	74.77	+13.21
Gemma-2-9B-IT	HotpotQA	7405	68.85	67.28	−1.57
	MuSiQue	2417	43.16	59.15	+15.99
	2WikiMHQA	12576	61.56	76.54	+14.98

Table 13: Cross-setting summary of Stage-1 design ablations. Values are average EM drops relative to the default setting across all 9 model–dataset combinations. "Consistent" indicates how many settings show the expected direction. Summary statistics were computed from a separate sensitivity run that also covered the N'=1 setting; they may differ slightly from values derived from Table 14 due to sampling stochasticity.

Design choice	Avg. Δ EM	Consistent
$N' = 10$ vs. $N' = 1$	−3.15	9/9
Paragraph vs. sentence	−0.20	5/9
Combined vs. uniform	−0.18	6/9
$K = [2, 4]$ vs. fixed $K = 2$	−0.49	7/9
Bridging on vs. off	+0.04	5/9

Table 21: XML format compliance (%) by model and prompt style.

Model / Style	HotpotQA	MuSiQue	2WikiMHQA
Qwen2.5-7B (0-shot)	99.8	100.0	99.8
Gemma-2-9B (0-shot)	99.2	90.2	98.4
Llama-3.1-8B (0-shot)	78.3	68.1	–[†]
Llama-3.1-8B (1-shot)	96.0	86.7	93.6

[†]Evaluation run produced empty outputs; data unavailable.

E.1 Citation Faithfulness Classification

We classify each question by answer correctness (EM = 1 vs. 0) and citation correctness (Stage-1 evidence recall ≥ 0.5 vs. < 0.5), yielding four categories. Table 22 reports the proportion of each category across all nine settings.

Two patterns stand out. First, the AC+CW category (correct answer despite wrong evidence) is consistently small (average 3.96%), indicating that SVote-2S rarely "gets lucky"—correct answers almost always co-occur with correct evidence selection. Second, the AW+CC category (wrong answer despite correct evidence) is the largest failure mode (average 43.71%), suggesting that evidence selection is often adequate but the re-reasoning step still fails. This motivates future work on improving Stage-2 reasoning quality given good evidence.

E.2 Case Studies

We present three case studies from HotpotQA (Qwen2.5-7B-Instruct), covering three categories from Table 22: a success case where evidence voting corrects a wrong SC answer (AC+CC, Figure 8), a case with correct evidence but wrong reasoning (AW+CC, Figure 9), and a case where wrong evidence selection leads to an incorrect answer (AW+CW, Figure 10). Each case study is self-contained in its figure below.

Algorithm 2 BRIDGINGSELECT: Bridging-Aware Greedy Selection

Require: Scores $\{\text{score}(s)\}$, paragraph mapping $\text{para}(\cdot)$, $K_{\min}$, $K_{\max}$
Ensure: Selected support set E

Sort candidates by score descending; let $s^* \leftarrow \max_s \text{score}(s)$
$E \leftarrow \emptyset$; $P_{\text{covered}} \leftarrow \emptyset$
for each $(s, \text{score}(s))$ in sorted order **do**
  if $|E| \geq K_{\max}$ **then break**
  end if
  if $\text{score}(s) < 0.3 \cdot s^*$ **then break**
  end if ▷ Hard cutoff
  $\hat{s} \leftarrow \text{score}(s) \,/\, s^*$ ▷ Normalized score
  if $E = \emptyset$ **then** ▷ First item: accept
    $E \leftarrow E \cup \{s\}$; $P_{\text{covered}} \leftarrow P_{\text{covered}} \cup \{\text{para}(s)\}$
  else if $\text{para}(s) \notin P_{\text{covered}}$ **then** ▷ New paragraph
    if $\hat{s} \geq \tau_{\text{new}}$ **then** ▷ $\tau_{\text{new}} = 0.5$
      $E \leftarrow E \cup \{s\}$; $P_{\text{covered}} \leftarrow P_{\text{covered}} \cup \{\text{para}(s)\}$
    else if $|E| < K_{\min}$ **then**
      $E \leftarrow E \cup \{s\}$; $P_{\text{covered}} \leftarrow P_{\text{covered}} \cup \{\text{para}(s)\}$
    end if
  else ▷ Same paragraph: strict
    if $\hat{s} \geq \tau_{\text{same}}$ **then** ▷ $\tau_{\text{same}} = 0.7$
      $E \leftarrow E \cup \{s\}$
    end if
  end if
end for
if $|E| < K_{\min}$ **then** ▷ Fallback: relax bridging
  Fill E with top-scored remaining items until $|E| = K_{\min}$
end if
return E

Table 14: Full Stage-1 design sensitivity (EM%). Each non-default row changes one design choice from the default (combined weighting, paragraph-level context, adaptive $K \in [2, 4]$, bridging on, $N' = 10$). Best result per setting in **bold**. Numbers may differ slightly from Table 1 due to sampling stochasticity.

	Qwen2.5-7B-Instruct			Llama-3.1-8B-Instruct			Gemma-2-9B-IT		
Variant	**HotpotQA**	**MuSiQue**	**2WikiMHQA**	**HotpotQA**	**MuSiQue**	**2WikiMHQA**	**HotpotQA**	**MuSiQue**	**2WikiMHQA**
Default	**54.81**	27.97	**50.16**	49.87	24.49	47.71	**56.53**	28.01	**54.25**
Uniform weighting	54.36	27.68	49.48	49.52	**25.16**	47.52	56.08	27.39	53.65
Sentence context	52.82	**28.01**	49.05	**50.82**	25.44	**51.26**	54.71	27.97	54.56
Fixed $K = 2$	54.79	26.85	49.20	49.91	24.53	47.28	56.37	27.14	52.97
No bridging	54.69	27.97	50.02	49.80	24.95	47.63	56.34	**28.18**	54.13

Table 15: Full SC saturation values (EM%). SC@k denotes the mean EM of majority vote over repeated random subsamples of k samples from a pool of 30. SVote-2S uses $N = 20$, $N' = 10$.

Model	**Dataset**	**@1**	**@2**	**@3**	**@5**	**@7**	**@10**	**@13**	**@15**	**@17**	**@20**	**@25**	**@30**	**SVote-2S**
Qwen2.5-7B-Instruct	HotpotQA	48.40	48.32	49.98	50.97	51.29	51.62	51.71	51.78	51.84	51.89	51.95	52.06	56.45
	MuSiQue	21.17	21.10	22.53	24.23	24.86	25.32	25.61	25.79	25.87	25.98	26.03	26.19	29.95
	2WikiMHQA	45.72	45.72	47.70	49.01	49.57	50.01	50.21	50.26	50.39	50.43	50.54	50.62	54.49
Llama-3.1-8B-Instruct	HotpotQA	37.57	37.58	39.81	41.36	41.97	42.30	42.58	42.60	42.76	42.79	42.92	43.00	50.02
	MuSiQue	12.88	12.87	13.71	15.01	15.61	15.95	16.08	16.16	16.21	16.35	16.45	16.55	24.16
	2WikiMHQA	34.26	34.34	36.65	38.68	39.55	40.12	40.38	40.53	40.60	40.70	40.82	40.87	48.80
Gemma-2-9B-IT	HotpotQA	51.68	51.64	52.53	52.92	53.10	53.20	53.26	53.30	53.30	53.30	53.29	53.33	56.91
	MuSiQue	21.61	21.64	22.62	23.49	23.70	24.08	24.13	24.23	24.26	24.28	24.31	24.33	28.38
	2WikiMHQA	49.04	49.06	49.95	50.42	50.61	50.73	50.82	50.84	50.91	50.95	50.99	51.11	56.56

Table 16: SVote-2S saturation values (EM%). k = number of Stage-1 traces for evidence voting; Stage-2 uses N'=10. Total inference budget = k+10.

Model	Dataset	k=3	k=5	k=7	k=10	k=15	k=20
Qwen2.5-7B-Instruct	HotpotQA	53.84	54.22	54.19	54.71	54.58	54.91
	MuSiQue	27.89	27.89	27.47	27.51	27.60	28.55
	2WikiMHQA	48.66	49.09	49.59	49.90	49.99	50.41
Gemma-2-9B-IT	HotpotQA	55.87	56.03	55.99	56.58	56.46	56.48
	MuSiQue	27.06	27.14	27.72	28.30	27.97	28.18
	2WikiMHQA	53.59	53.90	53.91	54.25	54.45	54.35
Llama-3.1-8B-Instruct	HotpotQA	49.33	49.63	50.20	50.20	50.45	50.80
	MuSiQue	23.50	24.08	24.82	24.62	25.24	25.78
	2WikiMHQA	47.65	47.78	48.05	48.58	48.27	48.65

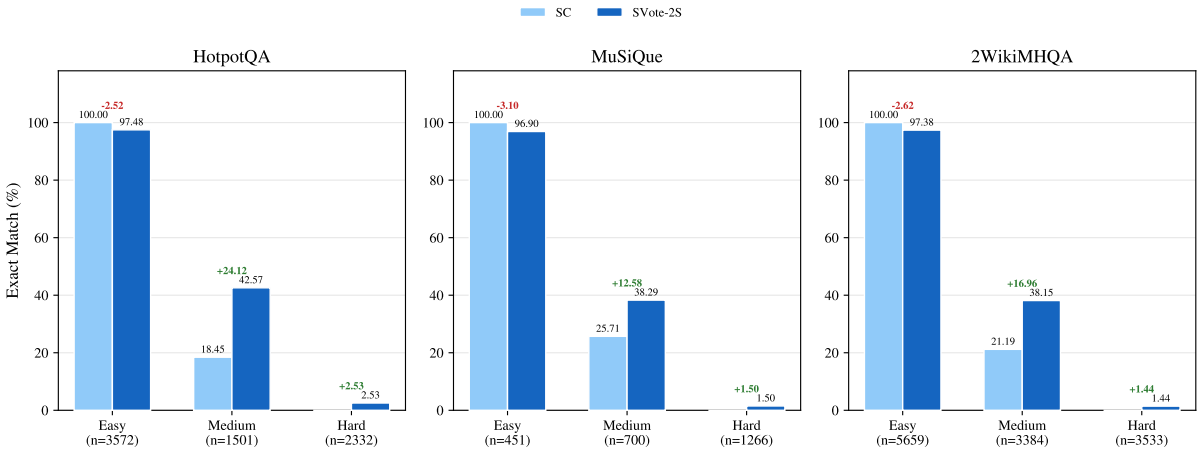


Figure 4: Difficulty-stratified EM across three datasets (Qwen2.5-7B-Instruct). Questions are grouped by SC correctness count.

Table 17: Difficulty-stratified EM for Qwen2.5-7B-Instruct. Questions are grouped by SC correctness count: Easy (11–20), Medium (1–10), Hard (0). Δ = SVote-2S − SC.

	Easy (11–20)			Medium (1–10)			Hard (0)		
Dataset	**SC**	**SVote-2S**	Δ	**SC**	**SVote-2S**	Δ	**SC**	**SVote-2S**	Δ
HotpotQA	100.00	97.48	−2.52	18.45	42.57	+24.12	0.00	2.53	+2.53
MuSiQue	100.00	96.90	−3.10	25.71	38.29	+12.58	0.00	1.50	+1.50
2WikiMHQA	100.00	97.38	−2.62	21.19	38.15	+16.96	0.00	1.44	+1.44

Table 18: Fusion sensitivity summary. "Spread" = Best EM − Worst EM across the grid. "All > SC" confirms every tested configuration exceeds the SC baseline.

Model	Dataset	#Configs	Best T	Best α	Best EM	Worst EM	Spread	SC EM	All > SC
Qwen2.5-7B-Instruct	HotpotQA	22	8.0	2.5	56.45	54.85	1.59	51.98	22/22
	MuSiQue	55	15.0	2.5	29.95	27.64	2.32	26.11	55/55
	2WikiMHQA	24	5.0	1.5	54.49	53.17	1.32	50.70	24/24
Llama-3.1-8B-Instruct	HotpotQA	24	2.0	3.0	50.02	48.28	1.74	43.23	24/24
	MuSiQue	24	10.0	3.0	24.16	22.13	2.03	16.51	24/24
	2WikiMHQA	24	2.0	3.0	48.80	46.75	2.05	40.67	24/24
Gemma-2-9B-IT	HotpotQA	24	5.0	3.0	56.91	56.16	0.74	53.26	24/24
	MuSiQue	24	8.0	3.0	28.38	26.44	1.94	23.91	24/24
	2WikiMHQA	24	2.0	2.0	56.56	55.22	1.34	51.11	24/24

Table 19: Answer diversity statistics before (Stage-1, S1) and after (Stage-2, S2) re-inference.

Model	Dataset	Unique Ratio (S1 / S2)	Majority Share (S1 / S2)	Entropy (S1 / S2)
Qwen2.5-7B-Instruct	HotpotQA	0.17 / 0.21	0.76 / 0.83	0.92 / 0.61
	MuSiQue	0.35 / 0.40	0.54 / 0.61	1.96 / 1.42
	2WikiMHQA	0.21 / 0.27	0.69 / 0.75	1.22 / 0.88
Llama-3.1-8B-Instruct	HotpotQA	0.21 / 0.24	0.68 / 0.78	1.24 / 0.77
	MuSiQue	0.34 / 0.35	0.53 / 0.64	1.97 / 1.29
	2WikiMHQA	0.26 / 0.30	0.57 / 0.67	1.65 / 1.11
Gemma-2-9B-IT	HotpotQA	0.11 / 0.15	0.84 / 0.91	0.58 / 0.31
	MuSiQue	0.26 / 0.28	0.63 / 0.74	1.49 / 0.93
	2WikiMHQA	0.13 / 0.19	0.79 / 0.85	0.75 / 0.50

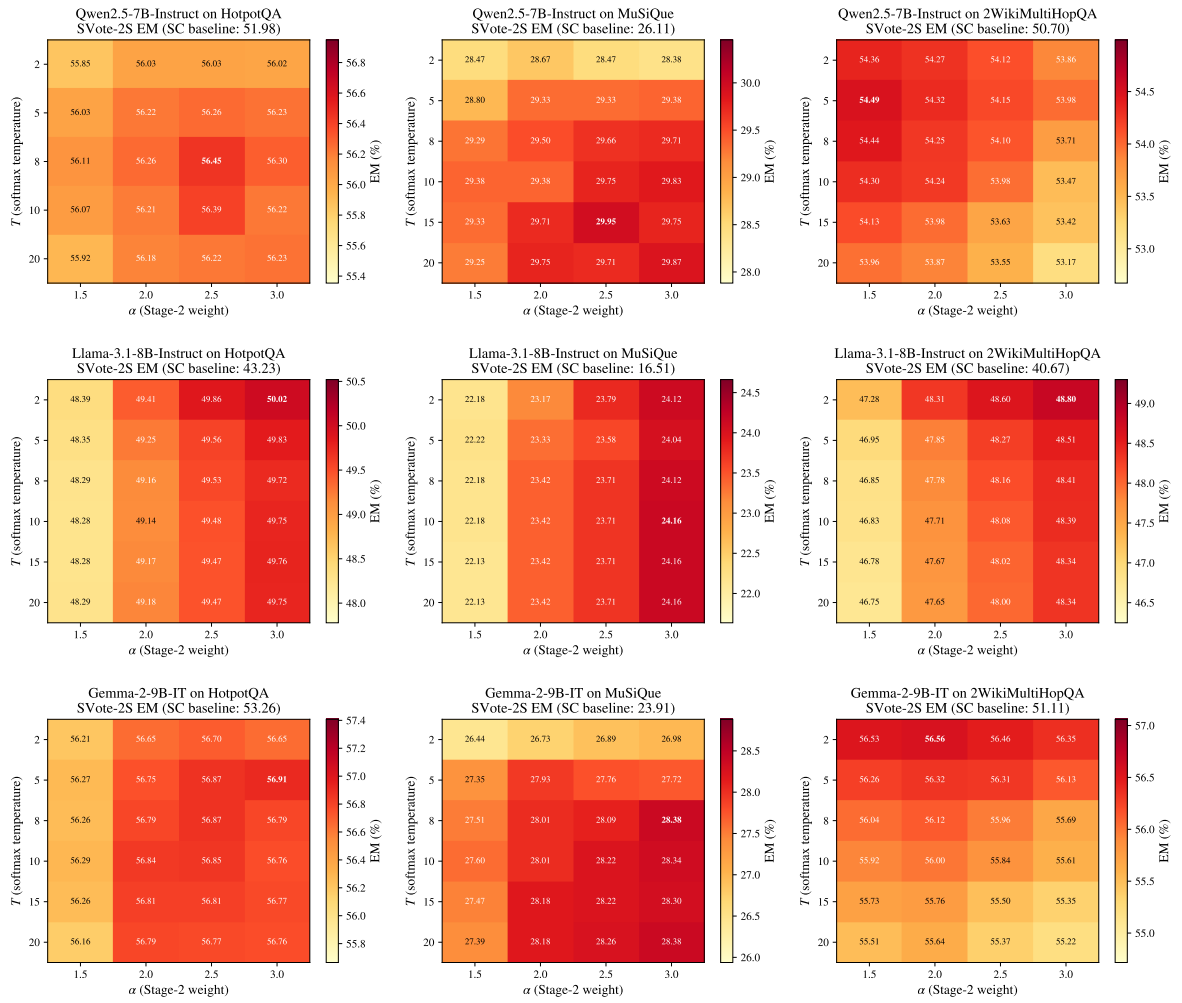


Figure 5: Fusion heatmaps (T vs. α) for all nine model–dataset settings. Rows: Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Gemma-2-9B-IT. Columns: HotpotQA, MuSiQue, 2WikiMHQA. All configurations exceed the corresponding SC baseline.

Table 20: Reasoning-type / hop-count breakdown (EM%). N denotes the number of questions per category. Rows are grouped by dataset: HotpotQA (bridge vs. comparison), MuSiQue (2/3/4-hop), and 2WikiMHQA (four reasoning types). $\Delta = \text{SVote-2S} - \text{SC}$.

			Qwen2.5-7B-Instruct			Llama-3.1-8B-Instruct			Gemma-2-9B-IT		
Dataset	**Type / Hops**	N	**SC**	**2S**	Δ	**SC**	**2S**	Δ	**SC**	**2S**	Δ
HotpotQA	bridge	5,918	48.36	51.84	+3.48	41.31	49.34	+8.03	50.74	54.02	+3.28
	comparison	1,487	66.24	67.11	+0.87	47.88	56.62	+8.74	63.08	66.24	+3.16
MuSiQue	2-hop	1,252	26.76	30.03	+3.27	18.53	29.95	+11.42	24.52	31.95	+7.43
	3-hop	760	23.68	26.45	+2.76	16.58	23.68	+7.11	22.63	23.82	+1.18
	4-hop	405	28.89	27.90	−0.99	9.38	16.79	+7.41	24.44	24.69	+0.25
2WikiMHQA	comparison	3,040	84.84	85.95	+1.12	61.64	72.40	+10.76	74.57	83.03	+8.45
	inference	1,549	29.63	33.57	+3.94	25.31	33.05	+7.75	41.32	48.22	+6.91
	compositional	5,236	32.47	33.38	+0.92	28.65	37.64	+9.00	33.00	35.83	+2.83
	bridge_comparison	2,751	58.71	53.04	−5.67	44.31	52.13	+7.82	64.49	61.36	−3.13

User Message: *You are given a question and a set of context sentences. Answer the question using only the provided context. You must cite the sentence IDs that directly support your reasoning.*

Question: *{question}*

Context:

[S1] Title_A: sentence text . . .
[S2] Title_A: sentence text . . .
. . .

Please output exactly three XML fields in this order:

`<reason>`*Use at most 5 concise reasoning sentences. Cite supporting sentence IDs when needed.*`</reason>`
`<support>`*List only the supporting sentence IDs, such as [S3], [S7].*`</support>`
`<answer>`*Give the shortest possible final answer.*`</answer>`

Rules: 1. Always close every XML tag. 2. The final output must end with `</answer>`*. 3. Do not output any text before* `<reason>` *or after* `</answer>`*. 4. Keep the reasoning concise. 5. Output ONLY the three XML fields. No other text.*

Figure 6: Prompt template for Qwen2.5-7B-Instruct and Gemma-2-9B-IT (zero-shot). The prompt is wrapped in each model's native chat template.

Table 22: Citation faithfulness classification (%). AC/AW = answer correct/wrong; CC/CW = citation correct/wrong (evidence recall $\geq / < 0.5$).

Model	**Dataset**	**AC+CC**	**AC+CW**	**AW+CC**	**AW+CW**
Qwen2.5-7B-Instruct	HotpotQA	47.68	7.13	34.69	10.49
	MuSiQue	24.45	3.52	47.79	24.24
	2WikiMHQA	47.00	3.16	41.02	8.82
Llama-3.1-8B-Instruct	HotpotQA	43.28	6.59	39.53	10.60
	MuSiQue	23.42	1.08	59.08	16.43
	2WikiMHQA	45.37	2.34	45.57	6.72
Gemma-2-9B-IT	HotpotQA	49.67	6.86	33.65	9.82
	MuSiQue	25.16	2.85	50.85	21.14
	2WikiMHQA	52.18	2.08	41.24	4.51
Average		**39.80**	**3.96**	**43.71**	**12.53**

User Message: *You are given a question and a set of context sentences. Answer the question using only the provided context. You must cite the sentence IDs that directly support your reasoning.*

Here is an example of the expected output format:

`<reason>`*[S3] states that Alice was born in 1910. [S7] states that Bob was born in 1923. Therefore Alice was born first.*`</reason>`
`<support>`*[S3], [S7]*`</support>`
`<answer>`*Alice*`</answer>`

—

Now answer the following question using the same XML format.

Question: *{question}*

Context:
[S1] Title_A: sentence text . . .
. . .

Please output exactly three XML fields in this order: . . .

Rules: 1–5 (same as above).

Figure 7: Prompt template for Llama-3.1-8B-Instruct (1-shot). A format example is prepended to improve XML tag compliance; the remaining instructions are identical to Figure 6.

Success case: evidence voting corrects a wrong SC answer (AC+CC).

Question: When was the composer of "Persian Surgery Dervishes" born? **Gold answer:** June 24, 1935 **Context:** 36 sentences from 10 paragraphs (including *Persian Surgery Dervishes*, *Terry Riley*, and 8 distractors).

SC (Stage-1 samples): The 20 sampled answers split into four groups: *terry riley* (8), *june 24 1935* (7), *1935* (4), and one verbose variant (1). Majority vote selects "*terry riley*" — **wrong** (EM = 0).

SVote Stage-1: Evidence voting selects paragraphs *Persian Surgery Dervishes* and *Terry Riley*, retaining only 5/36 sentences (13.9% of the original context). Gold support recall = 1.00.

SVote Stage-2: With the filtered context, the model reasons: "*From [S24], we learn that Terry Riley, the composer of 'Persian Surgery Dervishes,' was born on June 24, 1935.*" Final answer: "*June 24, 1935*" — **correct** (EM = 1).

Analysis: Under full context with 8 distractor paragraphs, 8/20 samples answered with the composer's *name* instead of the *birth date*. After evidence filtering removed all distractors, the model focused on the relevant sentences and correctly extracted the date.

Figure 8: AC+CC case study (HotpotQA, Qwen2.5-7B). SVote-2S corrects a wrong SC answer by filtering distractors via evidence voting.

AW+CC case: correct evidence but wrong reasoning.

Question: Jens Risom introduced what type of design, characterized by minimalism and functionality? **Gold answer:** Scandinavian design **Context:** 57 sentences from 10 paragraphs (including *Jens Risom*, *Scandinavian design*, and 8 distractors).

SC (Stage-1 samples): Only 4/20 samples correctly answer "*Scandinavian design*"; others answer "*minimalism and functionality*" (5 votes) or other variants. SC majority vote selects "*Scandinavian design*" — **correct** (EM = 1).

SVote Stage-1: Evidence voting correctly selects paragraphs *Jens Risom* and *Scandinavian design*, retaining only 4/57 sentences (7.0%). Gold support recall = 1.00.

SVote Stage-2: The model reasons: "*Jens Risom introduced a type of design characterized by minimalism and functionality*" and answers "*minimalism and functionality*" — **wrong** (EM = 0).

Analysis: Stage-1 evidence selection is perfect, yet Stage-2 reasoning fails. The question asks for the *name* of the design movement, but the model extracts its *characteristics* instead—a semantic confusion between "what type" (requesting a name) and "what kind" (requesting properties). This illustrates the AW+CC failure mode (43.71% of all cases on average): correct evidence is necessary but not sufficient for correct reasoning.

Figure 9: AW+CC case study (HotpotQA, Qwen2.5-7B). Evidence selection is perfect (recall = 1.0) but Stage-2 reasoning extracts characteristics instead of the design name.

AW+CW case: wrong evidence selection leads to an incorrect answer.

Question: Pierre Benjamin Monteux conducted the world premiere of which play by Maurice Ravel? **Gold answer:** Daphnis et Chloé **Context:** 45 sentences from 10 paragraphs (including *Pierre Monteux*, *Daphnis et Chloé*, and 8 distractors).

SC (Stage-1 samples): 10/20 samples correctly answer "*Daphnis et Chloé*" while the remaining 10 scatter across wrong answers. Majority vote selects the correct answer — **correct** (EM = 1).

SVote Stage-1: Evidence voting selects paragraphs *La reine Fiammette* and *Mârouf, savetier du Caire* — both about Monteux but unrelated to Ravel. Gold support recall = 0.00. Context retains 14/45 sentences (31.1%).

SVote Stage-2: The model reasons: "*There is no mention of Maurice Ravel's play in the context.*" and returns an empty answer — **wrong** (EM = 0).

Analysis: The gold-support sentences (*S7, S11, S13*) linking Monteux to Ravel's *Daphnis et Chloé* were cited by only a minority of Stage-1 samples. Stage-1 voting instead favored other Monteux-related paragraphs that lacked the Ravel connection. This illustrates SVote-2S's main failure mode: when the correct evidence is cited by fewer than half of the Stage-1 responses, voting may discard it.

Figure 10: AW+CW case study (HotpotQA, Qwen2.5-7B). Stage-1 voting discards gold evidence cited by a minority of samples, causing Stage-2 to fail.