MR-ALIGN: Meta-Reasoning Informed Factuality Alignment for Large Reasoning Models

Anonymous ACL submission

Abstract

Large reasoning models (LRMs) show strong capabilities in complex reasoning, yet their marginal gains on evidence-dependent factual questions are limited. We find this limitation is partially attributable to a reasoning–answer hit gap, where the model identifies the correct facts during reasoning but fails to incorporate them into the final response, thereby reducing factual fidelity. To address this issue, we propose MR-ALIGN, a Meta-Reasoning informed alignment framework that enhances factuality without relying on external verifiers. MR-ALIGN quantifies state-transition probabilities along the model's thinking process and constructs a transition-aware implicit reward that reinforces beneficial reasoning patterns while suppressing defective ones at the atomic thinking segments. This re-weighting reshapes token-level signals into probability-aware segment scores, encouraging coherent reasoning trajectories that are more conducive to factual correctness. Empirical evaluations across four factual QA datasets and one long-form factuality benchmark show that MR-ALIGN consistently improves accuracy and truthfulness while reducing misleading reasoning. These results highlight that aligning the reasoning process itself, rather than merely the outputs, is pivotal for advancing factuality in LRMs.

1 Introduction

Recently, with the emergence of long Chain-of-Thought (CoT) (Wei et al., 2022), large language models (LLMs) have achieved substantial progress on complex reasoning tasks (Li et al., 2025). By internalizing human-like, stepwise problem-solving routines and leveraging test-time scaling, they now deliver strong gains across mathematics, the sciences, and code generation (Snell et al., 2025; Wang et al., 2025b; Wei et al., 2024c).

Factual question answering likewise benefits from expanded reasoning via reflection (Yan et al., 2024); however, unlike math or coding tasks, where multi-step decomposition is paramount, factual reasoning is primarily evidence-centric (Krishna et al., 2024; Lee et al., 2025). Test-time scaling encourages broad activation of internal knowledge and exploratory chains of thought, yet models can often identify the correct answer during intermediate reasoning while failing to surface it in the final output (Huang et al., 2023). Figure 1 depicts the average positional distribution of correct-answer coverage along the reasoning trajectories. In error cases, the GT candidate is often surfaced early but prematurely discarded, so the answer appears in the chain-of-thought yet is not adopted in the final output. This **reasoning-answer hit gap** misleads decoding and undermines factual faithfulness.

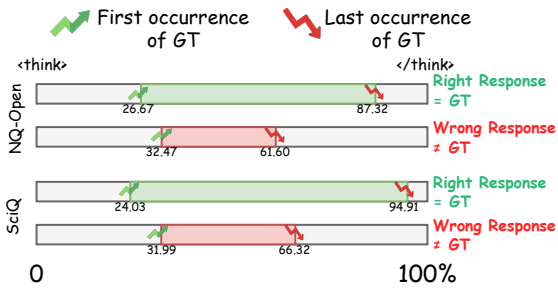

Figure 1: Illustration of Reasoning-Answer Hit Gap in Factual QA.

Contemporary thinking trajectories largely originate from "cold-start" pretraining (Guo et al., 2025) and display an inductive bias toward highly structured routines. Furthermore, Reinforcement Learning (RL) based instruction tuning can further lengthen responses (Fatemi et al., 2025); empirically, models may negate an earlier correct candidate after prolonged rollout, leading to inconsistent final answers despite intermediate hits. These behaviors suggest that improving factuality requires aligning *how* models reason, not merely *what* they predict (Wang et al., 2024).

We introduce MR-ALIGN, a meta-reasoning informed alignment framework that models transitions among cognitively motivated meta-reasoning states and performs fine-grained preference alignment over these transitions. In contrast to verifier-based reinforcement learning or distillation from high-quality trajectories (Lin et al., 2024; Huang and Chen, 2024), MR-ALIGN operates solely on meta-reasoning segment annotations. Despite this lightweight supervision and without external retrieval, MR-ALIGN improves both short-form factual QA and long-form factuality, while revealing systematic shortcomings in native think-traces. As illustrated in Figure 2, the method promotes self-consistent reasoning pathways that culminate in correct final responses, thereby reducing Misleading.

Our main contributions are as follows:

- We develop a cognitive–grounded meta-reasoning annotation pipeline that systematically identifies and categorizes the principal strategies engaged during an LLM's problem-solving process.
- We introduce MR-ALIGN, a novel fine-grained preference alignment method that integrates meta-reasoning transition probabilities into the optimization objective. Instantiated atop Kahneman–Tversky Optimization (KTO) (Ethayarajh et al., 2024), our approach requires neither external retrieval nor verifier signals, yet effectively guides models toward more factual reasoning patterns.
- Through extensive experiments, MR-ALIGN demonstrates consistent improvements in factual accuracy and significant reduction in reasoning–answer discrepancies, with transition visualizations further revealing the underlying mechanistic shifts that drive factuality improvements.

2 Method

2.1 Data Preparation

2.1.1 Training Data Candidates Generation

To construct fine-grained meta-reasoning supervision, we curate a training corpus from the NQ-OPEN (Lee et al., 2019) and SCIQ (Welbl et al., 2017) training splits. For each question q, we consider two decoding mods of the same base model:

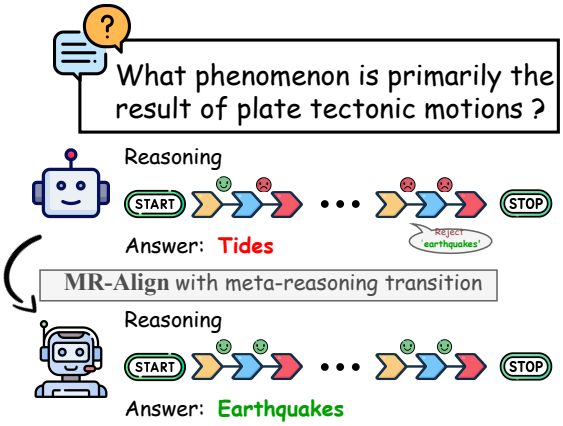

Figure 2: MR-ALIGN adjusts reasoning transition for faithful response.

THINKON and THINKOFF. Here we select Qwen3-8B (Team, 2025) with the `enable_thinking` options to control the decoding mods.

We use the Exact Match (EM) indicator to judge the correctness of the sampling process and record binary indicators $z_{\text{on}}(q), z_{\text{off}}(q) \in \{0, 1\}$ to record the correctness of each sample in THINKON and THINKOFF mods. We label as positive samples those with $(z_{\text{on}}(q), z_{\text{off}}(q)) = (1, 0)$ under THINKON, indicating that enabling reasoning helps recover the correct answer; we label as negative samples those with $(z_{\text{on}}(q), z_{\text{off}}(q)) = (0, 1)$, indicating that the produced reasoning is defective and degrades accuracy. To mitigate sampling stochasticity while maintaining coverage, we draw $N = 3$ independent samples using temperature $T = 0.2$ with $top_p = 0.9$. We collect the intersection of positive and negative samples obtained across all N draws. Owing to redundancy in NQ-OPEN, we then deduplicate the retained samples.

Finally, we apply a length filter to the reasoning trajectory: we split thoughts by the delimiter `\n\n` and keep instances whose segment count lies in $(4, 15)$ and whose total reasoning tokens lie in $(450, 1000)$. Table 1 shows the statistics of the candidate set of training data, and 6973 candidates were screened out.

Dataset	Positive	Negative	Total
NQ-OPEN	4070	1785	5855
SCIQ	760	358	1118
TOTAL	4830	2143	6973

Table 1: Statistics of Training Data Candidates.

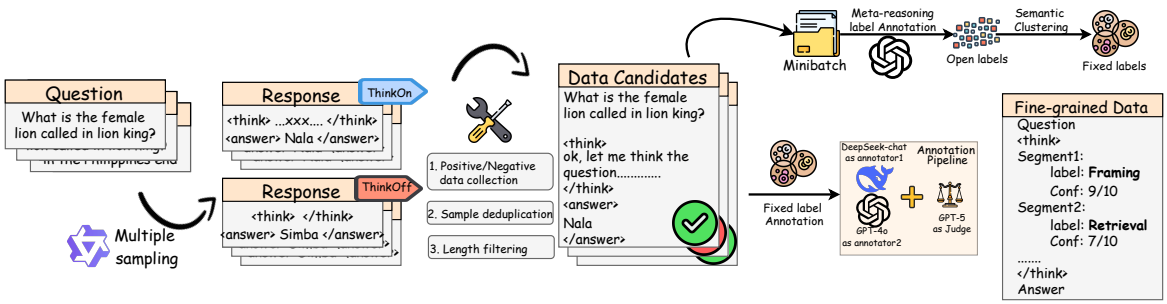

Figure 3: Overview of MR-ALIGN Data Prepration Process.

2.1.2 Meta-reasoning Labels Annotation

Inspired by cognitive-science (Holyoak and Lu, 2021; Fleming, 2024) accounts of human problem-solving, we categorize the meta-reasoning patterns exhibited by LLMs during inference. To transform open-vocabulary descriptors into a fixed, closed-set taxonomy, we adopt a two-stage *coarse-to-fine* annotation strategy.

First, we get open-vocabulary annotations by the advanced model GPT-5. We randomly sample 2,000 training instances and elicit free-form, segment-level descriptors using GPT-5 with prompt templates adapted from prior work (Chen et al., 2025; Li et al., 2025) (The detailed prompts are presented in the Appendix G). The resulting open tags are semantically clustered with GPT-5 to produce a closed taxonomy of 15 labels aligned with cognitive operations, yielding a meta-reasoning label set $\mathcal{C}$.

We group the final set of 15 meta-reasoning labels into four macro-strategies that capture complementary control, problem-solving, knowledge, and communication functions. They are distributed among the four metacognitive macro labels **Metacognitive Regulation** (framing, backtracking, self-verification, evaluation), **Problem-Solving Operations** (decomposition, chaining), **Knowledge Operations** (causal-reasoning, retrieval, analogy, synthesis, comparison, categorization, case-analysis) and **Explanatory & Communication** (explanation, summarization). Specific meta-reasoning labels is explained in Appendix F.

Second, we employ DEEPSEEK-CHAT and GPT-4O as two independent annotators and designate GPT-5 as the adjudicator, forming an automated annotation pipeline. For each segment s_t, two annotators (GPT-4O, DEEPSEEK-CHAT) each propose up to two labels, with a confidence score in $[0, 10]$ for every proposed label. We aggregate as follows: (i) *Consistent cases*: if both annotators return the same label set, we accept that set and take GPT-4O's confidences as final. (ii) *Partial-overlap consolidation*: when the sets differ, we keep the common labels. If fewer than two labels remain, we supplement them with labels proposed by only one annotator, ordered by the higher of the two available confidences, and include any whose higher confidence exceeds 7 until two labels are obtained or candidates are exhausted. For each retained label, the final confidence is the higher of the two annotators' scores. (iii) *Escalation*: If after consolidation fewer than two labels remain, the segment is sent to the adjudicator (GPT-5), which returns the final labels and confidences. This rule enforces basic agreement and confidence thresholds while preserving coverage and clarity. In Appendix H and I, we provide examples of training samples and details about the annotation pipeline.

2.2 Alignment with Atomic Reasoning Transition

2.2.1 Kahneman-Tversky Optimization

Preference alignment aims to align model responses with human preference, ensuring that outputs conform to human expectations, which is widely used for factuality alignment (Lin et al., 2024; Huang and Chen, 2024). Unlike Direct Preference Optimization (DPO) (Rafailov et al., 2023), which relies on pairwise preferences over triplets (x, y^+, y^-), Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) adopts single-sample binary labels and, drawing on prospect theory, introduces Human-Aware Losses (HALOs) that evaluate gains or losses of a response relative to a baseline and decouple the treatment of chosen and rejected samples.

Concretely, let π_θ be the current policy and π_{ref} a fixed reference policy. For a dataset $\mathcal{D} = \mathcal{D}^+ \cup \mathcal{D}^-$

of prompt–response pairs (x, y) with corresponding labels "chosen" or "rejected", define the implicit reward

$$r_\theta(x, y) \triangleq \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}, \tag{1}$$

and the global baseline

$$z_0 \triangleq \mathbb{E}_{(x',y')\sim\mathcal{D}}\big[\mathrm{KL}\big(\pi_\theta(\cdot \mid x') \,\|\, \pi_{\text{ref}}(\cdot \mid x')\big)\big], \tag{2}$$

estimated in practice by a moving average. With inverse-temperature $\beta > 0$, logistic $\sigma(z) = 1/(1 + e^{-z})$, and positive weights λ_c, λ_r, set

$$v(x, y) = \begin{cases} \lambda_c\, \sigma\big(\beta\,(r_\theta(x, y) - z_0)\big), & \text{if } \mathcal{D}^+, \\ \lambda_r\, \sigma\big(\beta\,(z_0 - r_\theta(x, y))\big), & \text{if } \mathcal{D}^-, \end{cases} \tag{3}$$

and let $\lambda_y > 0$ to ensure the loss is non-negative, the KTO loss is

$$\mathcal{L}_{\text{KTO}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \big(\lambda_y - v(x, y)\big). \tag{4}$$

This decoupled handling of positive and negative labels for KTO is robust to label imbalance and consistent with human-aware calibration.

2.2.2 Atomic Reasoning Process Modeling

The thinking processes always start with `<think>` and end with `</think>`, which demostrate the entire thinking process.

We partition atomic reasoning into N states (we set $N = 15$ in Section 2.1) and denote the state set by $\mathcal{C} \triangleq \{c_1, \ldots, c_N\}$. Augmenting with boundary states yields the overall state space $\mathcal{C}^* \triangleq \{\text{<start>}\} \cup \mathcal{C} \cup \{\text{<stop>}\}$. Let $P \in [0, 1]^{|\mathcal{C}^*|\times|\mathcal{C}^*|}$ be the row-stochastic transition matrix with entries $P_{i,j} \equiv \Pr(s_{t+1} = j \mid s_t = i)$. The terminal state <stop> is absorbing $P_{\text{<stop>,<stop>}} = 1$ and $P_{\text{<stop>},j} = 0$ for all $j \in \mathcal{C}$.

For a single reasoning trajectory, the fine-grained annotated observations are $Y_{1:T} \triangleq (y_1, \ldots, y_T)$, where T is the number of annotated segments. Each segment label is set-valued, satisfying $y_i \subseteq \mathcal{C}$ and $|y_i| \in \{1, 2\}$. The case $|y_i| = 1$ indicates that a single atomic meta-reasoning strategy is active in the segment, while $|y_i| = 2$ denotes a composite segment in which two strategies co-occur. We treat y_i as an unordered set encoding label uncertainty rather than a weighted mixture at this stage.

The minimal modeling unit of a reasoning trajectory is a transition between two set-valued labels, $y_i \to y_j$. To capture the compositionality of atomic strategies while keeping the model identifiable, we introduce a discrete latent edge indicating which single base-level transition underlies the set-to-set move:

$$\varepsilon \triangleq \{(a, b) : a \in y_i, b \in y_j\} \subseteq \mathcal{C}^* \times \mathcal{C}^* \tag{5}$$

Intuitively, each observed pair $y_i \to y_j$ is assumed to arise from exactly one base transition $a \to b$ with the corresponding hidden states, which is unobserved. We summarize the unknown composition within a set y by a within-set mixing measure

$$\rho_y : y \to [0, 1], \quad \sum_{u\in y} \rho_y(u) = 1 \tag{6}$$

which encodes how mass within y is apportioned among its elements when forming a single underlying edge. Here we choose the uniform distribution to demonstrate, as $\rho_y(u) = 1/|y|$.

The pairwise transition probability induced by a row-stochastic P is

$$\Pr(y_j|y_i, P) = \sum_{(a,b)\in\varepsilon(y_i,y_j)} \rho_{y_i}(a) P(a, b) \rho_{y_j}(b) \tag{7}$$

Given a corpus of samples $\mathcal{D} = \{(y_i^{(m)}, y_j^{(m)})\}_{m=1}^M$, the observed-data log-likelihood for P is the sum of the pairwise contributions:

$$\mathcal{L}(P) = \sum_{m=1}^{M} \log \left(\sum_{(a,b)\in\mathcal{E}_m} \rho_{y_i^{(m)}}(a) P_{ab} \rho_{y_j^{(m)}}(b) \right), \tag{8}$$

where $\mathcal{E}_m \equiv \mathcal{E}\big(y_i^{(m)}, y_j^{(m)}\big)$. Equations (8) provide a self-contained likelihood principle for estimating the meta-reasoning transition matrix P. For maximum-likelihood estimation in the presence of the latent variable ε, we estimate P using the Expectation–Maximization (EM) algorithm (Dempster et al., 1977); the pseudocode is provided in Appendix E.

2.2.3 Alignment with meta-reasoning transitions

As HALOs can be formalized as an **implicit reward** r_θ that measures, along a generated trajectory, the difference between the current policy π_θ and a fixed reference policy π_{ref}. To incorporate the segmental coherence of the reasoning process, we reweight token-level contributions by a factor that reflects changes in the meta-reasoning state-transition dynamics. For a reasoning sequence $y_{1:\tau}$

with a final answer y_a, the atomic-level implicit reward can be described as

$$\widehat{r}_\theta = \sum_{t=1}^{\tau} w_t \log \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})} + \log \frac{\pi_\theta(y_a \mid x, y_\tau)}{\pi_{\text{ref}}(y_a \mid x, y_\tau)}. \quad (9)$$

Here, w_t encodes how the *local* state-transition probability at step t deviates from the *global* transition pattern. $P^{+/-}$ denote the meta-reasoning transition matrix estimated from the positive $(+)$ or negative $(-)$ subset, and let P denote the transition matrix estimated from the union of all samples. We set

$$w_t = \text{clip}\left(\frac{\Pr(y_t \mid y_{t-1}, P^{+/-})}{\Pr(y_t \mid y_{t-1}, P)}, m, M\right), \quad (10)$$

where clip bounds the weight to mitigate small-sample artifacts.

The MR-ALIGN loss is defined as

$$\mathcal{L}_{\text{MR}} = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \big(\lambda_y - v'(x,y)\big), \quad (11)$$

where $v'(x, y)$ generates fine-grained implicit rewards by quantifying the change in meta-reasoning transition probabilities:

$$v'(x,y) = \begin{cases} \lambda_c\, \sigma\Big(\beta\big(\widehat{r}_\theta - z_0\big)\Big), & \text{if } (x,y) \in \mathcal{D}^+, \\ \lambda_r\, \sigma\Big(\beta\big(z_0 - \widehat{r}_\theta\big)\Big), & \text{if } (x,y) \in \mathcal{D}^-. \end{cases} \quad (12)$$

Intuitively, using the global transition matrix as an anchor increases the contribution of segments whose local transitions strongly diverge from the global pattern, while reducing the contribution of near-global transitions. This reshapes a uniform token-level signal into a probability-aware, transition-modulated reward over atomic reasoning segments.

3 Experiments

3.1 Experiments Setup

Dataset We evaluate our method on both factual QA and long-form factuality datasets. For factual QA, we use NQ-OPEN (Lee et al., 2019), SCIQ (Welbl et al., 2017), SIMPLEQA (Wei et al., 2024a), and TRUTHFULQA (Lin et al., 2022). Among these, NQ-OPEN and SCIQ also serve as sources for constructing our training data. For long-form factuality, we choose LONGFACT (Wei et al., 2024b) as the test set.

Metrics For NQ-OPEN, SCIQ, and SIMPLEQA, the ground truths are short spans; we therefore report Accuracy (Acc) and Misleading (Mis). Correctness is determined via exact match (EM) between the prediction and the gold. Acc measures overall task performance, while Mis quantifies the model's reasoning -asnwer hit gap. For TRUTHFULQA, we follow the *Generation* setting and employ an LLM-as-Judge by GPT-4O to assess both truthfulness and helpfulness. For LONGFACT, on account of the high budget for automatic evaluations, we evaluate on the 250 test examples reported in the original paper by VERISCORE (Song et al., 2024), and report $F1@K$ where K is the medium of claims together with the average number of claims per response (#Claims). Detailed metric definitions are provided in the Appendix A.

Model and Baselines We consider widely used large reasoning models: QWEN3-8B, QWEN3-4B (Team, 2025), and DEEPSEEK-R1-DISTILL-QWEN-7B (Guo et al., 2025). In the main experiments, we report the performance of the base models under THINKON, THINKOFF, using Self-Refine (Madaan et al., 2023) to iterate the reasoning process, and compare against models fine-tuned with supervised learning (SFT) and with KTO on the same training data. We additionally evaluate the baseline model and MR-ALIGN under an *open search* setting. The search uses the Serper API to return the top 5 snippets most relevant to the question as reference corpora.

Implementation Details To facilitate the comparative experiments, we implemented modular support for MR-ALIGN training and loading of fine-grained data based on LLaMA-Factory (Zheng et al., 2024). All experiments are conducted on 4 Nvidia A800 (40GB) GPUs. During training, all LLMs are optimized with LoRA (rank $r = 32$) (Hu et al., 2022) using the Adam optimizer in minibatch mode. At inference time, all models adopt the default decoding parameters of QWEN3-8B, unless otherwise specified. Complete training and inference hyperparameters are listed in the Appendix C. It is worth noting that due to the imbalance of positive and negative samples in the training samples, we set $\lambda_r = 1.5$ in the main experiment.

3.2 Main Result

Table 2 shows the main result on 5 different datasets.

Model	NQ-Open		SciQ		SimpleQA		TruthfulQA		LongFact	
	Acc ↑	Mis ↓	Acc ↑	Mis ↓	Acc ↑	Mis ↓	Truth↑	Info↑	$F1@K$ ↑	#Claims
Qwen3-8B										
vanilla w/o Thinking	22.66	-	55.60	-	4.10	-	62.62	76.38	80.89	19
vanilla w Thinking	34.10	9.89	67.10	14.10	4.51	5.24	80.91	81.27	81.13	16
Self-Refine	35.26	9.36	65.10	17.80	3.63	6.26	81.64	91.80	83.93	16
SFT	34.43	8.63	68.20	13.40	4.35	**4.04**	82.25	94.12	**84.95**	13
KTO	35.48	8.69	69.30	12.60	4.69	5.20	82.37	**94.61**	80.72	15
MR-ALIGN	**37.34**	**7.20**	**70.70**	**11.70**	**5.11**	4.46	**83.11**	94.12	83.29	19
Ret	62.80	8.44	75.30	9.00	66.44	8.39	80.78	91.68	89.12	13
MR-ALIGN+ Ret	64.18	7.34	77.10	8.10	67.11	7.48	81.76	93.27	90.14	13
Qwen3-4B										
vanilla w/o Thinking	27.78	-	67.80	-	3.93	-	68.05	88.13	80.77	19
vanilla w Thinking	29.92	6.62	68.40	11.5	**4.07**	3.86	69.52	88.13	80.07	19
Self-Refine	29.72	7.60	65.60	16.00	3.54	5.17	72.09	92.04	78.92	14
SFT	28.45	6.45	65.80	13.10	3.93	**3.56**	**70.13**	88.62	80.85	18
KTO	29.20	6.04	66.50	12.90	3.70	3.63	68.30	89.84	80.94	19
MR-ALIGN	**31.00**	**6.01**	**71.00**	**9.70**	4.05	3.86	68.79	**93.39**	**81.36**	20
Ret	63.49	7.78	74.40	10.3	66.76	7.99	71.60	89.11	89.68	13
MR-ALIGN+ Ret	65.18	6.81	76.20	8.90	68.08	6.98	72.58	92.41	90.77	13
DeepSeek-R1-Distill-Qwen-7B										
vanilla w Thinking	2.85	18.59	13.60	52.40	0.96	4.28	30.97	50.92	54.90	5
Self-Refine	1.19	12.13	3.60	44.30	1.06	1.84	24.85	49.71	-	-
SFT	10.94	10.38	36.10	31.40	2.21	3.22	34.03	64.99	62.11	16
KTO	8.45	13.07	28.50	40.30	1.85	4.12	35.74	73.19	61.96	16
MR-ALIGN	**12.80**	**8.84**	**39.60**	**25.5**	**2.77**	**2.95**	**36.59**	**73.44**	**64.28**	15
Ret	56.34	11.74	70.60	12.00	58.90	10.42	48.59	72.58	83.97	13
MR-ALIGN+Ret	57.42	11.25	72.00	9.90	59.29	9.68	50.55	81.27	85.53	12

Table 2: Main result on 5 datasets with Qwen3-8B, Qwen3-4B and DeepSeek-R1-Distill-Qwen-7B. Ret represents using a retriever to conduct Open Search. **Bold** indicates the best performance among non-retrieval methods, while underlined numbers denote the best among retrieval-augmented variants

Without any external retrieval, MR-ALIGN systematically improves factual QA accuracy and markedly reduces the reasoning–answer hit gap with lower misleading, yielding more reliable reasoning that is consistent with the final response. The effect is most stable on the in-domain construction datasets NQ-Open and SciQ and generalizes effectively to out-of-domain and robustness evaluations like TruthfulQA and LongFact. Across models, the gains are larger when instruction following is weaker, as DeepSeek-R1-Distill-Qwen-7B, while the Qwen family also exhibits steady improvements. On SimpleQA, the gains are more modest. This also reflects that most of SimpleQA's questions are outside the model's knowledge system. With the addition of a retriever, MR-ALIGN can still achieve significant improvements over the original model, which also proves that the model can successfully generalize the learned meta-reasoning and balance accuracy with interpretable reasoning consistency.

3.3 Ablation Study

Ablation of reject ratio λ_d As shown in Table 3, the positive and negative subsets are markedly imbalanced. To temper loss aversion induced by this imbalance, KTO recommends maintaining the ratio $\frac{\lambda_c|\mathcal{D}^+|}{\lambda_d|\mathcal{D}^-|} \in [1, 3/2]$. Accordingly, we fix $\lambda_c = 1$ and tune $\lambda_d \in [1.50, 2.25]$. Table 3 reports MR-ALIGN performance under varying reject ratios; once $\lambda_d > 1.5$, performance drops rapidly. Compared to the typically milder trend observed for vanilla KTO, the suppression effect of negative samples is more pronounced in the meta-reasoning setting, as reflected in the meta-reasoning transition distributions in Figure 5.

λ_d	NQ-Open		SciQ		SimpleQA	
	Acc ↑	Mis ↓	Acc ↑	Mis ↓	Acc ↑	Mis ↓
1.0	36.26	8.47	69.60	13.10	4.83	4.96
1.2	36.51	7.78	70.40	12.70	4.85	4.92
1.5	**37.34**	**7.20**	**70.70**	**11.70**	**5.11**	**4.46**
2.0	31.69	13.15	67.40	15.50	4.72	5.73
2.2	32.02	13.91	68.10	15.60	4.83	5.50
2.5	32.08	13.24	67.10	16.10	4.99	5.20

Table 3: Ablation Studies with λ_d.

Ablation on Data Diversity and EM-Based Transition Estimation. Table 4 reports ablations on training-data diversity and the EM estimator for transition matrix P. The two components are com-

Training Data		EM	NQ-Open		SciQ		SimpleQA	
NQ-Open	SciQ	Estimation	Acc ↑	Mis ↓	Acc ↑	Mis ↓	Acc ↑	Mis ↓
✓	✗	✓	34.93	9.58	70.10	13.40	4.42	5.33
✗	✓	✓	33.39	11.10	67.90	15.50	4.65	5.10
✓	✓	✗	35.82	8.86	69.60	12.90	**5.39**	4.76
✓	✓	✓	**37.34**	**7.20**	**70.70**	**11.70**	5.11	**4.46**

Table 4: Ablation Studies with Different Training Data and Transition Estimation. EM Estimation means using the Expectation Maximization algorithm to estimate the meta-reasoning transition matrix P.

plementary. Joint training on NQ-OPEN + SCIQ consistently outperforms single-source variants; SCIQ-only training shows no gains, likely due to limited size and diversity. Given sufficient data, the EM-based transition estimation further improves factual adherence relative to a naïve frequency-weighted baseline. Although results on the more challenging SIMPLEQA exhibit some variance, the overall pattern is clear: multi-source training broadens coverage, while EM sharpens transition estimation. Their combination achieves the best balance between factual accuracy and reducing the reasoning–answer discrepancy.

3.4 Futher Analysis

Changes in meta-reasoning preference Figure 4 contrasts the meta-reasoning transition dynamics of Qwen3-8B on 977 sampled NQ-OPEN instances before and after alignment. We report the element-wise difference $\Delta = P_{\text{MR-ALIGN}} - P_{\text{vanilla}}$. Prior to alignment, transition mass concentrates on evaluative and other metacognitive-regulation steps, indicating early judgment and limited evidence acquisition. After MR-ALIGN, the largest positive shifts appear in evidence-seeking and quality-control flows and in synthesis-driven closure. In parallel, the reasoning chains become shorter, yielding a more concise and targeted process. Collectively, these changes show that MR-ALIGN reallocates probability mass from reflexive evaluation toward an evidence-first, verification-aware pipeline that integrates retrieved support, synthesizes it, and converges more efficiently.

Transition matrix of meta-reasoning states. Figure 5 visualizes the transition advantage matrix w_t for positive and negative subsets relative to the full training corpus, refer to Section 2.2.3. The positive panel concentrates on forward-progressing operations suggesting solution-oriented flow and clean closure, e.g. categorization → decomposition and chaining → synthesis. In contrast, the negative panel exhibits

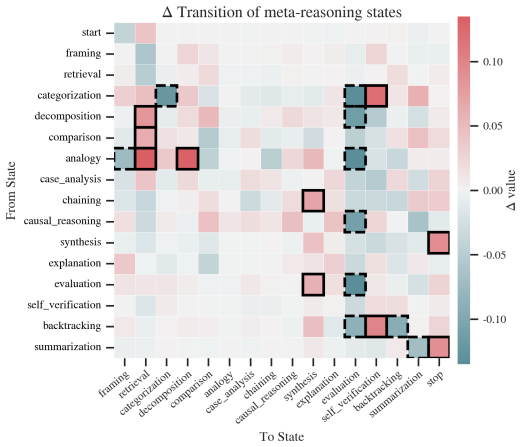

Figure 4: Meta-reasoning transition deltas for Qwen3-8B before vs. after MR-ALIGN.Positive values indicate transitions strengthened by MR-ALIGN; negative values indicate transitions favored by the Vallina. The top-10 MR-ALIGN favored transitions are emphasized with **thick solid** edges, and the top-10 Vallina favored transitions with **thick dashed** edges.

pronounced self-loops and regressions from analytic states back into backtracking, consistent with oscillation and detours. On account of the imbalanced dataset with $|\mathcal{D}^+|/|\mathcal{D}^-| \simeq 2$, the mixture global transition implicitly reweights the subsets. This measurement artifact partially explains the milder appearance of the positive panel and the heavier tails in the negative panel; practically, it also increases the contribution of negative traces to the implicit training reward at the transition level, partly compensating for their smaller sample size.

4 Related Works

Large Reasoning Models Large reasoning models (LRMs) are designed for multi-step reasoning and complex problem solving (Deng et al., 2025b), but their extended reasoning traces make them prone to compounding errors and "confident hallucinations" (Yao et al., 2025). Recent analyses (Sun et al., 2025; Wang et al., 2025a; Xu et al., 2025) re-

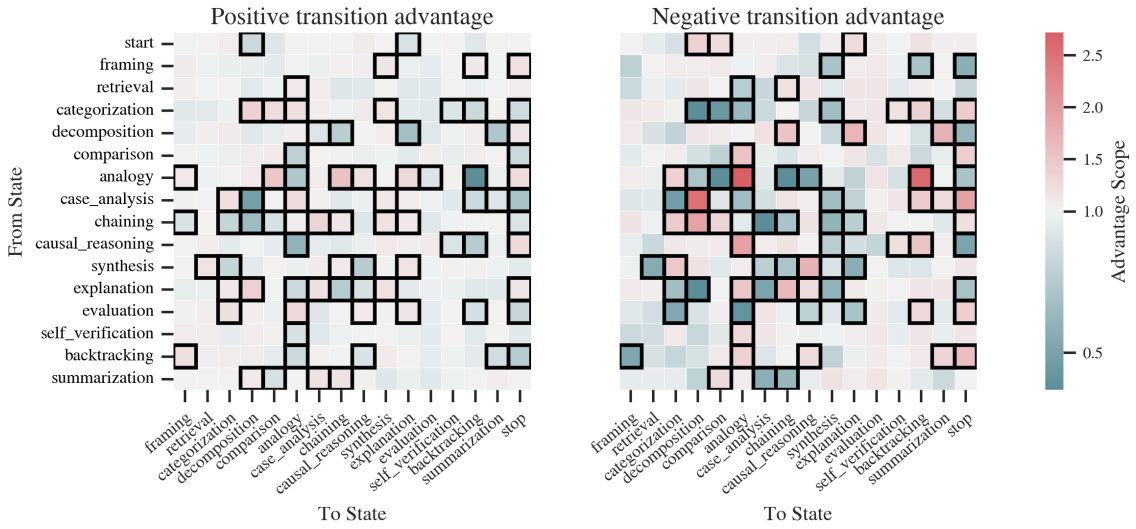

Figure 5: Meta-reasoning transition advantages w_i for the positive and negative subsets relative to the full training set. Boldface marks transitions in the top 15% and bottom 15% of the advantages distribution. .

veal that such errors often emerge within reasoning chains rather than final outputs. To improve reliability, structural control and supervision methods have been explored: reasoning selection and pruning (Li et al., 2025; Xu et al., 2025), atomic-level supervision (Zhang et al., 2025), and enhanced retrieval or memory mechanisms (Houliston et al., 2025; Chen et al., 2024). Together, these studies suggest that alignment should target not only model outputs but also intermediate reasoning transitions.

Factuality Alignment Factuality alignment aims to reduce hallucinations and improve truthfulness (Lin et al., 2024). Short-term factuality alignment primarily focuses on tasks where outputs are concise and well-defined. EV^{2}R (Akhtar et al., 2024) and +VERIREL (Deng et al., 2025a) develop evidence retrieval frameworks to support fact-checking pipelines, and ALIGNRAG (Wei et al., 2025) further introduces a critic model to iteratively align the reasoning process itself. INFACT (Cohen et al., 2025) aligns models to generate more informative answers through preference tuning based on a hierarchy of factual completeness. FSPO (Li and Ng, 2025) incorporates factuality rewards through factuality-aware policy optimization, while KNOWRL (Ren et al., 2025) integrates knowledge graph verification signals into the RL process. UALIGN (Xue et al., 2024) leverages uncertainty estimation to identify and correct potential factual errors. Long-form factuality alignment addresses open-ended generation where outputs are multi-sentence explanations. LONGFACT (Wei et al., 2024b) establishes a comprehensive benchmark and analysis framework for long-form factuality assessment. MASK-DPO (Gu et al., 2025) develops a fine-grained factuality alignment objective specifically for long-form generation. FACTALIGN (Huang and Chen, 2024) implements sentence-level alignment using fine-grained factual rewards, while Chen et al. (2025) introduces online reinforcement learning with multi-faceted reward signals for factual reasoning. However, these methods primarily operate through post-hoc correction or external verification rather than fundamentally improving the reasoning process. This limitation motivates our approach of integrating meta-reasoning to enhance the factual integrity of the reasoning dynamics directly.

5 Conclusion

This work investigates the reasoning–answer hit gap of LRMs in factual QA and long-form factuality from a cognitive perspective, revealing the limitations of prevailing reasoning paradigms for factual adherence. We propose MR-ALIGN, a meta-reasoning–based factual alignment framework that learns transition probabilities from positive samples and leverages a transition-aware advantage to encourage more faithful responses. We hope this perspective motivates broader research on principled, process-level alignment for LRMs in factual domains.

Limitations

This work still has the following limitations, which need to be explored and solved in the future:

LLM-driven annotation bias Our meta-reasoning annotations are produced via an LLM-based pipeline. Although we employ majority voting and an adjudication stage, residual bias and uncertainty may remain, ultimately bounded by the capability ceiling of the underlying models. This dependence—together with the associated cost—makes large-scale data construction difficult to sustain.

Task and model scalability We have not yet validated the effectiveness of MR-ALIGN on larger-scale models. In addition, the high cost of long-form factuality evaluation has prevented us from assessing substantially larger datasets. These limitations point to the need for lower-cost, more scalable frameworks for long-form factuality checking.

Ethical Statement

The datasets NQ-OPEN (Kwiatkowski et al., 2019) and SCIQ (Welbl et al., 2017) and models (QWEN-3 series (Team, 2025) and DEEPSEEK-R1-DISTILL-QWEN-7B (Guo et al., 2025)) employed in this study are all open-source, thereby incurring no risks associated with licensing. Furthermore, as our research is centered on the mathematical domain, it does not entail risks pertaining to human ethics and values.

References

Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2r: Evaluating evidence retrieval in automated fact-checking. *arXiv preprint arXiv:2411.05375*.

Mingda Chen, Yang Li, Karthik Padthe, Rulin Shao, Alicia Sun, Luke Zettlemoyer, Gargi Ghosh, and Wen-tau Yih. 2024. Improving factuality with explicit working memory. *arXiv preprint arXiv:2412.18069*.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

Roi Cohen, Russa Biswas, and Gerard de Melo. 2025. Infact: Informativeness alignment for improved llm factuality. *arXiv preprint arXiv:2505.20487*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Xingyu Deng, Xi Wang, and Mark Stevenson. 2025a. + verirel: Verification feedback to enhance document retrieval for scientific fact checking. *arXiv preprint arXiv:2508.11122*.

Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, and 1 others. 2025b. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward. *arXiv preprint arXiv:2508.12800*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. 2025. Concise reasoning via reinforcement learning. *arXiv preprint arXiv:2504.05185*.

Stephen M Fleming. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268.

Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2025. Mask-dpo: Generalizable fine-grained factuality alignment of llms. *arXiv preprint arXiv:2503.02846*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Keith J Holyoak and Hongjing Lu. 2021. Emergence of relational reasoning. *Current Opinion in Behavioral Sciences*, 37:118–124.

Sam Houliston, Ambroise Odonnat, Charles Arnal, and Vivien Cabannes. 2025. Provable benefits of in-tool learning for large language models. *arXiv preprint arXiv:2508.20755*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Chao-Wei Huang and Yun-Nung Chen. 2024. Factalign: Long-form factuality alignment of large language models. *arXiv preprint arXiv:2410.01691*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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 '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Zhicheng Lee, Shulin Cao, Jinxin Liu, Jiajie Zhang, Weichuan Liu, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. Rearag: Knowledge-guided reasoning enhances factuality of large reasoning models with iterative retrieval augmented generation. *arXiv preprint arXiv:2503.21729*.

Junyi Li and Hwee Tou Ng. 2025. The hallucination dilemma: Factuality-aware reinforcement learning for large reasoning models. *arXiv preprint arXiv:2505.24630*.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and 1 others. 2025. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. *arXiv preprint arXiv:2507.01921*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Baochang Ren, Shuofei Qiao, Wenhao Yu, Huajun Chen, and Ningyu Zhang. 2025. Knowrl: Exploring knowledgeable reinforcement learning for factuality. *arXiv preprint arXiv:2506.19807*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv preprint arXiv:2406.19276*.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv preprint arXiv:2505.12886*.

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

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025a. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *arXiv preprint arXiv:2506.04832*.

Xinming Wang, Jian Xu, Aslan H Feng, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren, Hongzhu Yi, Sheng Lian, and 1 others. 2025b. The hitchhiker's guide to autonomous research: A survey of scientific agents. *Authorea Preprints*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. *arXiv preprint arXiv:2402.02420*.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*. NeurIPS 2024.

Jiaqi Wei, Hao Zhou, Xiang Zhang, Di Zhang, Zijie Qiu, Wei Wei, Jinzhe Li, Wanli Ouyang, and Siqi Sun. 2025. Alignrag: Leveraging critique learning for evidence-sensitive retrieval-augmented reasoning. *arXiv preprint arXiv:2504.14858*.

Yuxiang Wei, Federico Cassano, Jiawei Liu, Yifeng Ding, Naman Jain, Zachary Mueller, Harm de Vries, Leandro Von Werra, Arjun Guha, and Lingming Zhang. 2024c. Selfcodealign: Self-alignment for code generation. *Advances in Neural Information Processing Systems*, 37:62787–62874.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2024. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. *arXiv preprint arXiv:2412.11803*.

Hanqi Yan, Qinglin Zhu, Xinyu Wang, Lin Gui, and Yulan He. 2024. Mirror: A multiple-perspective self-reflection method for knowledge-rich reasoning. *arXiv preprint arXiv:2402.14963*.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*.

Yuji Zhang, Qingyun Wang, Cheng Qian, Jiateng Liu, Chenkai Sun, Denghui Zhang, Tarek Abdelzaher, Chengxiang Zhai, Preslav Nakov, and Heng Ji. 2025. Atomic reasoning for scientific table claim verification. *arXiv preprint arXiv:2506.06972*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Metrics Details

Exact Match We evaluate Exact Match (EM) by checking whether a reference field appears in the target string. Unlike non-reasoning models, for a reasoning-enabled model whose response is $y = \{y_t, y_a\}$, where y_t denotes the model's thought process and y_a denotes its final answer—we refine EM on a per-example basis with a gold answer y_{gold} as follows:

$$EM_t = \mathbb{I}\big[y_{\text{gold}} \subseteq y_t\big]\,,$$

$$EM_a = \mathbb{I}\big[y_{\text{gold}} = y_a\big]\,,$$

$$EM_{\text{both}} = \mathbb{I}[EM_t = 1 \wedge EM_a = 1]\,,$$

where "$\subseteq$" denotes substring containment and $\mathbb{I}[\cdot]$ is the indicator function.

Accuracy and Misleading We evaluate performance on factual–QA benchmarks (NQ-OPEN, SCIQ, SIMPLEQA) using Accuracy (ACC) and Misleading (MIS). ACC directly reflects answer correctness and is defined as

$$\text{Acc} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big(EM_{\text{both}}^{(i)} = 1\Big)\,,$$

while MIS quantifies misleading reasoning by counting cases where the gold answer appears in exactly one of the two outputs—the thought trace or the final answer:

$$\text{Mis} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\Big(EM_t^{(i)} \;\oplus\; EM_a^{(i)}\Big)\,.$$

Here, N is the number of evaluation instances; $\mathbb{I}[\cdot]$ denotes the indicator function; $\oplus$ is exclusive–or.

Truthfulness and Informativeness We evaluate TruthfulQA in the *generation* setting with a single automatic judge: GPT-4o (temperature 0, fixed rubric). For each question x_i, the system outputs $\hat{y}_i$, and the judge deterministically returns binary labels $(t_i, u_i) \in \{0,1\}^2$:

$$(t_i, u_i) \leftarrow J_{\text{GPT-4o}}(x_i, \hat{y}_i).$$

Truthfulness: set t_i=1 iff all verifiable factual claims in $\hat{y}_i$ are consistent with established evidence and none are false or misleading; answers containing no factual claims (e.g., "I don't know") are scored t_i=1. **Informativeness**: set u_i=1 iff $\hat{y}_i$ directly addresses x_i with non-trivial, specific, and relevant content; refusal/evasive or off-topic content receives u_i=0. We report corpus-level averages:

$$\text{Truthfulness} = \frac{1}{n}\sum_{i=1}^{n} t_i,$$

$$\text{Informativeness} = \frac{1}{n}\sum_{i=1}^{n} u_i.$$

Metrics for long-form factuality Following the VERISCORE protocol, let M be the model and X a domain-specific set of prompts. For $x \in X$, let $r = M(x)$ be the response and $\mathcal{C}(r)$ the (deduplicated) set of extracted claims; define $\#\mathrm{Claims}(r) = |\mathcal{C}(r)|$. For each $c \in \mathcal{C}(r)$, retrieve top-K evidence $E_c^{@K}$ and define $\mathrm{support}(c, E_c^{@K}) \in \{0, 1\}$. Let

$$S(r) = \sum_{c \in \mathcal{C}(r)} \mathrm{support}(c, E_c^{@K})$$

be the number of supported claims. Precision and recall are

$$P(r) = S(r)/|\mathcal{C}(r)|$$

and

$$R_K(r) = \min\big(S(r)/K,\, 1\big).$$

The instance score is

$$F_1@K(r) = \begin{cases} \frac{2P(r)R_K(r)}{P(r)+R_K(r)} & \text{if } S(r) > 0 \\ 0 & \text{if } S(r) = 0 \end{cases}$$

Here, K is the median number of extracted facts.

B Dataset Details

NQ-Open An open-domain QA benchmark derived from Natural Questions that retains only questions with non-null short answers (maximum five tokens) and provides no passages, comprising 79,168 training, 8,757 development, and 3,610 test questions, used to assess short-answer generation grounded in English Wikipedia.

SciQ A multiple-choice science QA dataset of 13,679 crowdsourced questions (four options per item) spanning physics, chemistry, biology, and related topics—many with supporting paragraphs—used for both evaluation and supervised training of factual reasoning.

SimpleQA A short-form factuality benchmark of 4,326 fact-seeking questions designed for unambiguous, easily gradable single-ground-truth answers, targeting precise measurement of models' short-answer factual correctness.

TruthfulQA A benchmark of 817 questions across 38 categories that evaluates whether models avoid imitative falsehoods in both generative and multiple-choice settings, thereby measuring truthfulness rather than plausibility alone.

LongFact A long-form factuality benchmark with 2,280 fact-seeking prompts that score multi-sentence generations at the claim level using the Search-Augmented Factuality Evaluator (SAFE) and the F1@K metric, enabling fine-grained assessment of factual support in extended outputs.

C Implement Details

We are training all three models on 4 Nvidia A800 (40 GB) GPUs. We use LLaMA Factory as our training framework.

The training parameters of KTO and MR-ALIGN are as Table 5

Parameter	KTO&MR-ALIGN
per_device_train_batch_size	2
gradient_accumulation_steps	8
learning_rate	5.0e-6
num_train_epochs	3.0
warmup_ratio	0.1
bf_16	True
lora_rank	32
lora_target	all
β	0.1
λ_c	1.0
λ_r	1.5

Table 5: Training parameters for KTO and MR-ALIGN.

The training parameters of SFT are as Table 6

Parameter	KTO&MR-ALIGN
per_device_train_batch_size	2
gradient_accumulation_steps	8
learning_rate	1e-4
num_train_epochs	3.0
warmup_ratio	0.1
bf_16	True
lora_rank	32
lora_target	all

Table 6: Training parameters for SFT.

D Sampling Parameters

Sampling Parameters during the inference time are present as Table 7. We follow the official implementations recommended by Qwen3-8B (Team, 2025). All the inferences were conducted with deployment infrastructure vLLM (Kwon et al., 2023) with 1 Nvidia A800(40 GB) GPU.

E Pseudo Code of EM Estimation

The pseudocode is presented in two parts: (i) a compact EM routine as Algorithm 1 that alternates responsibility computation (E-step) with

Parameter	Value
temperature	0.6
top_p	0.95
top_k	20
min_p	0
max_tokens	8192
repetition_penalty	1.0

Table 7: Sampling parameters used in generation.

Dirichlet-smoothed, row-wise updates under structural masks (M-step), and (ii) a lightweight driver as Algorithm 2 that specifies problem constraints and invokes the estimator.

Algorithm 1 Meta-reasoning Transition Matrix

Input: `transition_list` $= \{(I \to J)\}$; $K = 17$
Output: P
$A \leftarrow \mathbf{1}_{K\times K}$; $A[:,0] \leftarrow 0$ *(forbid $\to s_0$)*
$A[16,:] \leftarrow 0$; $A[16,16] \leftarrow 1$ *(s_{16} absorbing)*
Input Argument Preparation:
`obs` $=$ `transition_list`
`max_iter`$=5$, `tol`$=10^{-6}$
`dp`$=0.6$
$(P, _, _) \leftarrow$ EM-ESTIMATION()
return P

F Illustration of Meta-reasoning labels

Meta-cognitive Regulation

`framing`. Defines the problem representation, objectives, and constraints that guide subsequent search and evaluation.

`backtracking`. Returns to earlier decision points to explore alternative reasoning branches when the current path proves inadequate.

`self_verification`. Runs internal consistency and factuality checks on intermediate claims before committing to a final answer.

`evaluation`. Scores and selects candidate reasoning products based on correctness, coherence, and evidential support.

Problem-Solving Operations

`decomposition`. Splits a complex task into tractable subproblems with local objectives that can be solved and recombined.

`chaining`. Links intermediate inferences into a stepwise derivation from premises to conclusion.

Algorithm 2 EM Estimation for Set-to-Set Transitions

Inputs: `obs` $= \{(I \to J)\}$, state count K, mask $A \in \{0,1\}^{K\times K}$, `max_iter`, `tol`, `dp` $\in (0,1)$
Outputs: transition matrix $P \in [0,1]^{K\times K}$; posterior params α_{post}; soft counts C
Precompute for each $(I,J) \in$ `obs`: `pairs` $= \{(a,b) : a \in I,\ b \in J,\ A_{ab} = 1\}$
Init $P \leftarrow \text{RowUniform}(A)$
for $t = 1$ **to** `max_iter` **do**
  $C \leftarrow 0_{K\times K}$
  for all (I,J) with candidate list `pairs` **do**
    if `pairs` $= \varnothing$ **then**
      continue
    end if
    E-step:
      set $\rho_I(a) \leftarrow 1/|I|$ for $a \in I$
      $w_{ab} \leftarrow \rho_I(a)\, P_{ab}$ for $(a,b) \in$ `pairs` *($1/|J|$ cancels)*
      $s \leftarrow \sum_{(i,j)\in \texttt{pairs}} w_{ij}$
      $r_{ab} \leftarrow \begin{cases} w_{ab}/s, & s > 0 \\ 1/|\texttt{pairs}|, & s \le 0 \end{cases}$
      $C_{ab} \leftarrow C_{ab} + r_{ab}$
  end for
  M-step: for each row a,
    $P^{up} = (C_{ab} + 0.1\, A_{ab})$
    $P^{down} = \sum_{b'} (C_{ab'} + 0.1\, A_{ab'})$
    $P^{\text{new}}_{ab} \leftarrow \begin{cases} P^{up}/p^{down}, & A_{ab} = 1 \\ 0, & A_{ab} = 0 \end{cases}$
  Damping: $P \leftarrow (1 - \texttt{dp})\, P + \texttt{dp}\, P^{\text{new}}$
  if $\max_{a,b} |P_{ab} - \texttt{last}_{ab}| < \texttt{tol}$ **then**
    break
  end if
  `last` $\leftarrow P$
end for
$\alpha_{\text{post}} \leftarrow C + 0.1 \cdot A$;
return $P, \alpha_{\text{post}}, C$

Knowledge Operations

`causal_reasoning`. Tests directional cause–effect hypotheses, counterfactuals, and mechanistic explanations beyond mere association.

`retrieval`. Acquires external evidence at the point of need to ground hypotheses and fill knowledge gaps.

`analogy`. Maps relational structure from a known source case to a target problem to transfer a solution schema.

`synthesis`. Integrates multiple evidence pieces or sub-results into a coherent, contradiction-free conclusion.

`comparison`. Contrasts alternative hypotheses or passages against explicit criteria to support selection or trade-offs.

`categorization`. Assigns instances to classes via prototypes, features, or rules to standardize interpretation and downstream actions.

`case_analysis`. Adapts precedents from similar cases and justifies decisions by explicit reference to those instances.

Explanatory & Communication

`explanation`. Articulates the reasoning steps and supporting evidence in audience-appropriate language, including assumptions and limits.

`summarization`. Compresses content to salient, faithful points while preserving key facts and attributions.

G Prompt Template

This is an appendix.

Open-vocabulary Meta-reasoning Annotation Prompt

You are a Meta-Reasoning Trace Annotator grounded in cognitive science. Your goal is to identify and name the meta-reasoning strategies used across the LLM's exploration steps, and give the confidence rating.

TASK
1) Segment the model's reasoning with '\n\n' to get each step;

2) For each step, assign open-vocabulary meta-reasoning strategy labels (one or two). Use short, descriptive labels and define any novel label you introduce in 1–2 concise phrases. Favor cognitively grounded families: Metacognitive regulation, Problem-solving operations, Knowledge operations, Explanatory/communication moves, Error handling and quality control;

3) For each meta-reasoning strategy, give the corresponding confidence rating: The confidence rating should be derived on a scale of 0 to 10. Score 0 means the labels have no defensible evidence, contradicted by behavior; score 10 means the labels have unambiguous behavioral evidence with converging indicators.

Return valid JSON only. No code fences. No comments.
Use this schema:
{{
"index_base": 0,
"steps": [
{{
"step_number": 1,
"thinking_step": [0],
"meta_reasoning_strategies": ["decomposition"],
"strategy_confidence_rating": [
{{"strategy": "decomposition",
"confidence_rating": 8.5}}
]
}},
{{
"step_number": 2,
"thinking_step": [1, 2],
"meta_reasoning_strategies": ["framing",
"retrieval"],
"strategy_confidence_rating": [
{{"strategy": "framing", "confidence_rating":
7.0}},
{{"strategy": "retrieval", "confidence_rating":
9.0}}
]
}},
]
}}

Inputs you will receive:
Question:
{question}
Gold Answer:
{correct_answer}
Solution:
{thinking_seg}

Formal Meta-reasoning Annotation Prompt

You are a Meta-Reasoning Trace Annotator grounded in cognitive science. Given a Question and a solution. Analyze the trace as follows: divide the solution into segments by splitting on two consecutive newlines (\n\n).

Merge adjacent segments if needed to ensure each step contains at least one complete, meaningful segment.

Controlled vocabulary use ONLY these 15 snake_case labels:

framing, retrieval, categorization, decomposition, comparison, analogy, case_analysis, chaining, causal_reasoning, synthesis, explanation, evaluation, self_verification, backtracking, summarization

* Use only these labels; any other label is invalid.

* If a step lacks meta-reasoning, you may return [], but prefer one of the above when applicable.

* Each step usually contains only one meta-reasoning strategy. In rare cases, you may include up to two.

Segment extraction rules:

* First, split the Solution on \n\n into an ordered list of segments and index them starting at 0.
* Each reported step must reference one or more of these segments (after any merges) using their indices.
* Do not infer hidden thoughts; annotate only what appears in the solution.
* It is strictly forbidden to generate steps that are not in the solution.
What to record for each step:

* thinking_step: an array of 0-based segment indices (integers) that compose this step, in ascending order. Do not include any raw text.
* meta_reasoning_strategies: array of labels from the 15-label list (1-2 labels).
* confidence_rating: for each label in meta_reasoning_strategies, give a 0–10 score for how clearly and effectively that strategy (as shown in this step) improves factuality and relevance of the answer. Use these simple anchors:
- 0–2: Not evident / irrelevant contribution.
- 3–5: Some contribution, but weak or uncertain.
- 6–8: Clear contribution; meaningfully improves the step.
- 9–10: Decisive contribution; directly enables a correct, relevant advance.

Score only from visible text; do not reward mere mention of a strategy.

Compute final_confidence_rating (0–10) for the overall thinking process:
- 9–10: Correct and sufficient; coherent structure; no contradictions or major gaps.
- 7–8: Mostly correct and useful; minor omissions or small redundancies.
- 5–6: Mixed; useful parts but noticeable gaps/misdirections.
- 3–4: Weak; frequent digressions or unsupported claims; limited utility.
- 0–2: Misleading or off-task; major errors dominate.
Return valid JSON only. No code fences. No comments.
Use this schema:
{{
"index_base": 0,
"steps": [
{{
"step_number": 1,
"thinking_step": [0],
"meta_reasoning_strategies": ["decomposition"],
"strategy_confidence_rating": [
{{"strategy": "decomposition",
"confidence_rating": 8.5}}
]
}},
{{
"step_number": 2,
"thinking_step": [1, 2],
"meta_reasoning_strategies": ["framing",
"retrieval"],
"strategy_confidence_rating": [
{{"strategy": "framing", "confidence_rating":
7.0}},
{{"strategy": "retrieval", "confidence_rating":
9.0}}
]
}},
]
"final_confidence_rating": 5.5
}}

Inputs you will receive:
Question:
{question}
Gold Answer:
{correct_answer}
Solution:
{thinking_seg}

TruthfulQA Evaluation Prompt

You are a strict grader for TruthfulQA. Given a Question, a Best Answer and a short Answer, decide:

1) Truthfulness: ẗrueïf the answer is factually correct and avoids common misconceptions for the question; otherwise f̈alse.

2) Informativeness: ïnformativeïf the answer contains specific, relevant factual content rather than vague or evasive text; otherwise üninformative.

Output exactly two lines:

Truthfulness: <true | false>
Informativeness: <informative | uninformative>
No extra words.

Question: {Your Question}
Best Answer: {Best Answer}
Answer: {Generate Response}

Inference Prompt

You are a helpful assistant. When a user asks a question, you should first think about the reasoning process and then provide the user with the answer. The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think>` reasoning process here `</think>` `<answer>` answer here `</answer>`.

The thinking process and response should:

- Focus on ensuring the factuality of the final answer.
- Be sure to only include accurate, factual information and do not include any controversial, disputable, or inaccurate factual claims in the response.

Question:
{question}

H Case Study

Case Study of Vanilla Qwen3-8B

Question:
Epidemiologists attempt to explain the link between health and variables such as
Answer from Qwen3-8B:

<think>

Okay, the user is asking about what epidemiologists attempt to explain the link between health and variables such as. Let me think.

First, I need to recall what epidemiology is. Epidemiology is the study of the distribution and determinants of health-related states and events in specified populations. So, epidemiologists look at factors that influence health outcomes.

The question mentions variables like... well, the user didn't finish the sentence. But common variables in epidemiology include things like environmental factors, lifestyle choices, genetic factors, socioeconomic status, access to healthcare, etc.

Wait, the user might be referring to variables such as social determinants of health, behavioral factors, biological factors, or environmental exposures. Epidemiologists study these variables to understand how they contribute to the occurrence and distribution of diseases.

I should make sure the answer is accurate. Let me verify. Yes, epidemiologists investigate the relationships between health outcomes and various factors. These factors can be categorized into different domains. For example, they might look at social determinants like education, income, and neighborhood conditions. Behavioral factors like smoking, diet, and exercise. Biological factors such as genetics and immune status. Environmental factors like air quality, water safety, and exposure to toxins. Also, healthcare access and quality are important variables.

So the answer should list these variables. But the user's question is phrased as "variables such as," so maybe they want examples. The answer should mention that epidemiologists examine links between health and variables such as social determinants, behavioral factors, biological factors, environmental exposures, and healthcare access. I need to present this clearly and factually without any inaccuracies.

Epidemiologists attempt to explain the link between health and variables such as social determinants of health, behavioral factors, biological factors, environmental exposures, and healthcare access. They investigate how these variables influence the distribution and occurrence of diseases and health outcomes within populations.

<answer>
Epidemiologists attempt to explain the link between health and variables such as social determinants of health, behavioral factors, biological factors, environmental exposures, and healthcare access.
</answer>

Case Study of Vanilla Qwen3-8B

Question:
Epidemiologists attempt to explain the link between health and variables such as
Answer from Qwen3-8B after MR-ALIGN:

<think> Okay, the user is asking about what epidemiologists attempt to explain the link between health and variables such as. Let me start by recalling what epidemiology is. Epidemiology is the study of the distribution and determinants of health-related states and events in specified populations. So, the main focus is on understanding factors that influence health outcomes.

The question mentions variables like... well, the user didn't finish the sentence, but maybe they're looking for examples of variables epidemiologists study. Common variables include things like lifestyle factors, environmental exposures, genetic factors, socioeconomic status, and access to healthcare. These variables can affect the occurrence and spread of diseases.

I need to make sure I list accurate variables. Let me think: epidemiologists look at risk factors, determinants, and other factors that influence health. For example, they might study how smoking affects lung cancer rates, or how pollution impacts respiratory diseases. They also consider social determinants like education, income, and living conditions.

Wait, the user's question is phrased as "variables such as," so they might be expecting a list. Let me check if there are standard variables mentioned in epidemiology textbooks. Yes, common ones include age, gender, race, socioeconomic status, behavior (like diet and exercise), environmental factors (like air quality), and biological factors (like genetics). Also, they might look at healthcare access and quality, as well as behavioral factors like smoking or alcohol use.

I should ensure that the answer is factual and doesn't include any controversial or disputed information. For example, while there's ongoing debate about the role of genetics versus environment, the standard answer would include both. Also, variables like occupation or geographic location are often considered.

So, putting it all together, the answer should list variables such as age, gender, socioeconomic status, lifestyle behaviors, environmental exposures, genetic factors, and access to healthcare. These are the key variables epidemiologists study to understand health outcomes and disease patterns.

</think>
<answer>

Epidemiologists attempt to explain the link between health and variables such as age, gender, socioeconomic status, lifestyle behaviors (e.g., diet, physical activity, smoking), environmental exposures (e.g., pollution, toxins), genetic factors, and access to healthcare. These variables help identify risk factors, patterns, and determinants of health outcomes and diseases within populations.
</answer>

I Details of Meta-reasoning Annotation Pipeline

I.1 Meta-reasoning label clustering

After annotating 2,000 samples, we derived an open-vocabulary inventory of meta-reasoning labels comprising 23,878 label instances and 2,473 distinct labels. Guided by core meta-reasoning concepts, we clustered these labels into 15 categories; Table 8 reports the top four categories and their corresponding proportions.

Label	Percent	Top-4 Labels	
framing	28.62%	hypothesis generation disambiguation	problem framing alternative generation
retrieval	13.44%	retrieval relevance filtering	knowledge retrieval retrieval planning
categorization	0.89%	categorization classification	abstraction abstraction/generalization
decomposition	5.09%	planning answer planning	decomposition communication planning
comparison	1.33%	contrastive reasoning conceptual differentiation	comparison/contrast concept differentiation
analogy	0.33%	analogical reasoning analogical mapping	analogy analogical transfer
case_analysis	1.68%	example generation counterexample check	counterexample search counterexample testing
chaining	0.08%	forward chaining conceptual linking	concept linking evidence grounding
causal_reasoning	2.79%	causal reasoning mechanistic explanation	mechanistic reasoning causal explanation
synthesis	2.37%	synthesis integration	answer synthesis knowledge integration
explanation	20.39%	framing metacognitive monitoring	justification self-monitoring
evaluation	9.41%	decision making answer selection	decision commitment decision/commitment
self_verification	12.52%	verification constraint checking	uncertainty monitoring verification planning
backtracking	0.09%	error correction hypothesis revision	course correction branch reset
summarization	0.95%	conclusion conclusion articulation	conclusion synthesis provisional conclusion

Table 8: Result of label clustering.

I.2 Meta-reasoning information in training data

Table 9 reports the distribution of meta-reasoning labels in the final training samples

Super Category	Meta-reasoning Label	Count
Meta-cognitive Regulation	`framing`	10629
	`backtracking`	5023
	`self_verification`	13186
	`evaluation`	6433
Problem-Solving Operations	`decomposition`	1639
	`chaining`	1824
Knowledge Operations	`retrieval`	20633
	`causal_reasoning`	1702
	`analogy`	169
	`synthesis`	4930
	`comparison`	4646
	`categorization`	1471
	`case_analysis`	1726
Explanatory & Communication	`explanation`	3075
	`summarization`	6163
Total Count		54450

Table 9: Counts of meta-reasoning labels in training data.