READABLE CONSTRAINTS, UNRELIABLE DECISIONS: A CONSTRAINT-INSTANTIATION LENS ON LANGUAGE MODEL ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural answers expose a model's final judgment, but may not express all task-relevant information recoverable from its internal states. We study whether such recoverable information tracks the truth required by the current input and whether disagreement between recoverable truth and the model's decision can diagnose errors. In a verifiable Fact–Rule–Action (FRA) setting, we distinguish task-defined local truth, hidden-state readouts, and natural responses. Correct local truth remains recoverable under answer errors, and frozen readouts systematically track truth-changing edits to facts, rules, and action arguments across new worlds. On a held-out controlled cohort, readout–answer disagreement strongly ranks natural-answer errors. We further extend the same truth-readout principle to natural QA, where the model's realized answer instantiates the proposition to be evaluated. The resulting error-oriented post-answer readout, FRA-ER, improves error ranking over generation confidence on several tasks across two model families.

1 INTRODUCTION

A wrong natural answer does not reveal whether task-relevant truth remains recoverable from a model's internal states. Prior work has shown that correct content can be decoded from hidden states even when the generated answer is wrong (Azaria and Mitchell, 2023; Orgad et al., 2025); other studies document failures to use evidence already present in the prompt, including sensitivity to evidence position (Liu et al., 2024). These observations motivate a focused question: *when the model's answer is wrong, does an internal measurement still track the truth required by the current input, and can that comparison help diagnose the error?*

Aggregate probe accuracy alone does not show whether a fixed internal score follows the specific information that determines the required answer. We therefore need a setting in which the target is independently verifiable and its determining factors can be changed under explicit controls. This lets us test both semantic responsiveness and the diagnostic value of disagreement with the natural answer.

We instantiate this design using Fact–Rule–Action (FRA) tasks, where facts specify the current state, a rule defines the condition to be checked, and action arguments bind that condition to concrete entities. The resulting local proposition has an externally verifiable truth, and each of its determining factors can be manipulated while keeping the others fixed; the query q identifies which precondition is evaluated. Figure 1 illustrates this instantiation and how its verified truth can differ from the model's natural answer. Section 3 formalizes the controlled setting, after which we extend the same truth-readout principle to propositions instantiated by model-generated answers in natural QA.

Frozen readouts recover local truth when natural answers fail and systematically track truth-changing edits to facts, rules, and action arguments across new worlds. On a held-out two-precondition cohort, a post-hoc analysis shows that disagreement between the truth readout and the model's realized decision strongly ranks answer errors. We then extend the same truth-readout perspective to natural QA: the generated answer itself instantiates the proposition whose verifier-defined truth is estimated from post-answer states. The resulting error-oriented readout, FRA-ER, is evaluated across five QA datasets and two model families, where it provides gains over generation confidence on several tasks.

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

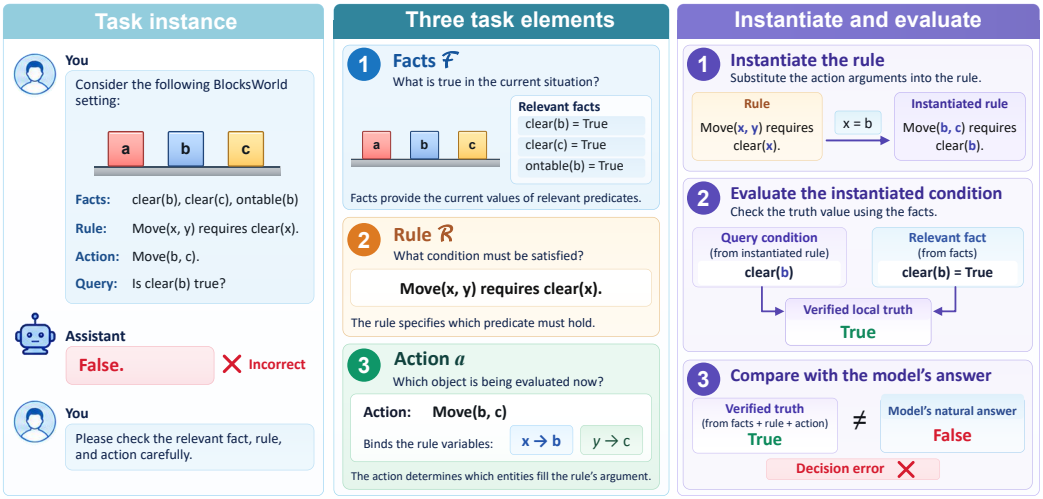

Figure 1: An illustrative Fact–Rule–Action task.

Our contributions are threefold. First, we introduce a controlled evaluation that separates externally verified local truth, frozen readout predictions, and natural decisions. Second, we provide empirical evidence that task-defined truth remains recoverable under answer errors, that frozen readouts respond to truth-changing semantic edits, and that readout–answer disagreement supports error diagnosis. Third, we extend the same truth-readout principle to model-generated answers in natural QA through FRA-ER, and evaluate its behavior across tasks, model families, feature sets, model scales, and computational settings.

2 Related Work

Internal truth and probing. Truthfulness benchmarks evaluate model answers, while probing studies ask which properties are recoverable from hidden states (Lin et al., 2022b; Alain and Bengio, 2016; Belinkov, 2022). Prior work has shown that internal truth signals can diverge from natural answers (Azaria and Mitchell, 2023; Burns et al., 2023; Marks and Tegmark, 2024; Liu et al., 2023; Orgad et al., 2025). Probe controls and causal interventions further distinguish recoverability from claims about model use (Hewitt and Liang, 2019; Pimentel et al., 2020; Voita and Titov, 2020; Richardson and Sabharwal, 2020; Ravichander et al., 2021; Elazar et al., 2021; Geiger et al., 2021; Wu et al., 2023; Li et al., 2023). We build on this distinction by testing a fixed truth readout under edits with independently known effects on the required answer.

Constraints, binding, and planning. A complementary line of work studies contextual entity representations and binding (Kim and Schuster, 2023; Li et al., 2021; Feng and Steinhardt, 2024; Feng et al., 2025), as well as rule-governed reasoning and planning (Clark et al., 2020; Tafjord et al., 2021; Valmeekam et al., 2023a;b; Kokel et al., 2025; Xie et al., 2025; Dziri et al., 2023; McCoy et al., 2019). The closest constraint-based precedent, SAT, relates attention to constraint tokens to output-level satisfaction (Yuksekgonul et al., 2024). We define the target as the externally verified truth of an instantiated requirement and test whether a fixed internal readout tracks edits that change this target. Attention allocation remains an association measure rather than an explanation of behavior, consistent with broader cautions about interpreting attention weights (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Abnar and Zuidema, 2020; Kobayashi et al., 2020; Geva et al., 2021).

Error detection from internal states. Reliability methods use calibration, verbal confidence, self-consistency, semantic uncertainty, or evidence checking (Jiang et al., 2021; Kadavath et al., 2022; Lin et al., 2022a; Tian et al., 2023; Xiong et al., 2024; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Agrawal et al., 2024). Other detectors use hidden activations or attention,

including ReDeEP and CHARM (Azaria and Mitchell, 2023; Orgad et al., 2025; Yuksekgonul et al., 2024; Sun et al., 2025; Frasca et al., 2026). Our controlled analysis reads out the verified truth of an explicitly instantiated requirement and compares it with the model's natural decision. FRA-ER extends this truth-readout perspective to natural QA: the realized answer instantiates the proposition to be evaluated, task-specific verifiers define its truth, and post-answer states are used to estimate that answer-conditioned target. Section 5 specifies the corresponding task-adapted comparisons.

3 Task-Conditioned Constraint Instantiation

3.1 Task-Defined Local Constraint

We consider an input $X = (E, R, a, q)$, where E is the current set of facts, R contains action requirements, a is a candidate action, and q selects the requirement being queried. Write the action as $a = \rho(o_1, \ldots, o_m)$, with action schema ρ and entity arguments o_i. Let $P(R, a)$ denote the index set of preconditions applicable to this action, and let $q = k \in P(R, a)$ select the queried precondition. Its task-defined local truth is distinct from whether the full action is executable: overall applicability is $\mathrm{APP}(E, R, a) = \bigwedge_{j \in P(R,a)} b_j^*(X)$, whereas the query asks for only $b_k^*(X)$. Consequently, neither overall applicability nor the total number of failed preconditions generally determines the queried local truth; the controlled cohorts use this distinction to construct matched cases (Appendix B.1).

3.2 Fact–Rule–Action Instantiation

For the selected requirement template R_k, the action arguments determine a binding $\theta_a = \{x_i \mapsto o_i\}$, mapping each requirement variable in argument slot i to the corresponding action argument. The instantiated constraint and its external truth are

$$\phi_k(X) = R_k \theta_a, \qquad b_k^*(X) = \mathrm{Eval}(\phi_k(X), E). \tag{1}$$

The rule and action binding specify which concrete proposition is checked; the facts determine its truth. The controlled symbolic task uses the verifier's closed-world semantics: an atomic fact is true when it is present in E, and an absent atomic fact is false. For example, if the rule requires $\mathrm{clear}(u)$ and the action binds $u \mapsto b$, the verifier evaluates $\mathrm{clear}(b)$ against E.

3.3 Local Truth, Readout, and Natural Decision

We distinguish three objects for the later evaluations. The external verifier supplies task-defined local truth $b = b_k^*(X)$; each valid natural response is semantically mapped to a Boolean answer $y \in \{0, 1\}$, with 1 denoting True; and a readout applied to a hidden state yields a continuous score s_ℓ and a thresholded prediction $r_\ell = \mathbf{1}[s_\ell > 0]$. Invalid or unparsable responses are handled under the cohort-specific protocol rather than assigned an arbitrary Boolean value. The verifier defines the target, the natural answer records the model's decision, and the readout is an independently fitted measurement. Section 4 defines how the latter is obtained and used.

Figure 2 summarizes how these quantities are obtained and compared: the verifier computes the reference truth directly from the task input, while the natural answer and the frozen truth readout are distinct measurements associated with the model. **Summary.** The FRA formulation gives us a task-defined local truth that is distinct from both overall action success and the model's natural decision. This provides a controlled target for studying what information is recoverable from model states. We next define internal readouts of this target and show how their relation to realized decisions can support error diagnosis.

4 Internal Readouts for Decision Error Diagnosis

4.1 Reading Task-Defined Local Truth

To estimate local truth from model states, we fit an affine readout at a specified block and observation position. Let $H_\ell \in \mathbb{R}^{n \times d}$ contain the post-block residual-stream states, with $h_{\ell,t}^{\mathrm{out}}$ the state at position

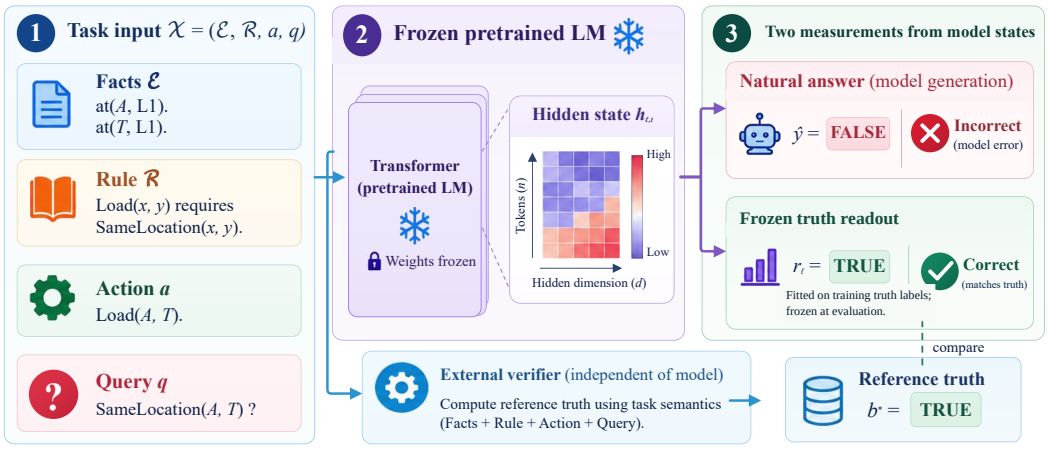

Figure 2: The FRA measurement framework.

t. At the designated observation event, the truth-readout score and Boolean prediction are

$$s_{\ell,t}(X) = w_{\ell,t}^\top \mathrm{Std}_{\mathrm{train}}[h_{\ell,t}^{\mathrm{out}}(X)] + \beta_{\ell,t}, \qquad r_{\ell,t}(X) = \mathbf{1}[s_{\ell,t}(X) > 0]. \tag{2}$$

The score ranks cases by the readout's estimated truth; thresholding gives a Boolean prediction.

Readout weights and standardization statistics are fitted on training groups and held fixed for evaluation; the base language model is also frozen. Each study specifies its own readout source and observation event. Prompt-final states contain the task input before generation, whereas later selection-position states can include a generated prefix. This distinction determines the information available to the external readout (Alain and Bengio, 2016; Hewitt and Liang, 2019; Pimentel et al., 2020; Voita and Titov, 2020; Richardson and Sabharwal, 2020; Ravichander et al., 2021). Hooks, layers, and fitting protocols are given in Appendices D.1 and D.2.

4.2 Diagnosing Errors by Readout–Answer Disagreement

Given the frozen truth-readout score s_ℓ and observed Boolean answer y, we orient the score toward answer error:

$$D_{\mathrm{TD},\ell} = (1 - 2y)s_\ell. \tag{3}$$

For the reported five-layer variant, let $\mathcal{L}$ be the five measured blocks. We scale each layer by the standard deviation of its truth score on the training group and then average:

$$D_{\mathrm{multi}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{(1 - 2y)s_\ell}{\sigma_{\ell,\mathrm{train}}}. \tag{4}$$

At test time these scores use only frozen readout scores and y; verified truth is used to evaluate ranking, not to construct the scores.

For binary labels, let $b = b_k^*(X)$, $e = \mathbf{1}[y \neq b]$ be answer error, and $d_\ell = \mathbf{1}[y \neq r_\ell]$ be hard disagreement. Since $e = y \oplus b$ and $d_\ell = y \oplus r_\ell$, the natural answer cancels:

$$\mathbf{1}[d_\ell \neq e] = \mathbf{1}[r_\ell \neq b]. \tag{5}$$

Thus hard disagreement has the same classification errors as the truth readout on those examples. This pointwise identity does not imply equal AUROC for continuous truth scores and disagreement scores; their ranking performance is measured empirically in Section 5.3. A complete statement of the binary conditions is in Appendix A.2.

4.3 Post-Answer Truth Readouts for Natural-QA Error Diagnosis

Natural QA admits the same truth-readout view once the model's realized answer is treated as the candidate whose validity is being evaluated. For a dataset-specific task input X_d and generated answer $\hat{Y}$, let

$$b_d^*(X_d, \hat{Y}) = V_d(X_d, \hat{Y}) \tag{6}$$

denote the task-defined truth of the realized answer proposition, where V_d is the dataset-specific external verifier. Thus $b_d^* = 1$ when the generated answer satisfies the task-defined answer condition and $b_d^* = 0$ otherwise. Unlike the controlled Boolean setting, the realized answer is part of the proposition whose truth is being read out.

We estimate this answer-conditioned truth from hidden states associated with the realized answer. Let $t_1, \ldots, t_m$ be the parsed answer-content positions and $\mathcal{L}$ the five selected blocks. The Post representation concatenates the mean and final answer-content states at each block:

$$\bar{h}_\ell^{\text{ans}} = \frac{1}{m} \sum_{j=1}^{m} h_\ell(t_j), \qquad v_{\text{post}} = \text{Concat}_{\ell \in \mathcal{L}}[\bar{h}_\ell^{\text{ans}}; h_\ell(t_m)]. \tag{7}$$

Token masks, layer mappings, and extraction rules are in Appendix D.1.

For error diagnosis, we use the equivalent error-oriented target $e_d = 1 - b_d^*(X_d, \hat{Y})$. Training-only normalization of the Post representation is followed by an affine readout:

$$z_{\text{ER}} = B_{\text{tr}}^{(\text{post})}(v_{\text{post}}), \qquad s_{\text{ER}} = w_{\text{ER}}^\top z_{\text{ER}} + \beta_{\text{ER}}. \tag{8}$$

We report s_{ER} in error-positive orientation for error-ranking evaluation; equivalently, $-s_{\text{ER}}$ is the corresponding truth-oriented score for the realized answer proposition. Each model–task readout is fitted on training answer-error labels and frozen before test labels are joined. The fitting protocol is specified in Appendices D.2 and C.8.

Summary. Across both controlled FRA tasks and natural QA, the diagnostic centers on the truth of the proposition instantiated by the current task. In the controlled setting, the truth readout is compared with a separate Boolean decision. In natural QA, the realized answer itself instantiates the proposition, and FRA-ER reads out whether that answer satisfies the task-defined condition in error-positive orientation. We next evaluate truth recoverability, semantic responsiveness, and error-ranking performance under these two forms.

5 Experiments

Research questions. We evaluate four questions: (RQ1) whether a frozen readout recovers task-defined local truth when natural answers err; (RQ2) whether its score tracks controlled edits that flip that truth; (RQ3) whether readout–answer disagreement ranks errors on the held-out controlled cohort; and (RQ4) whether the same truth-readout principle extends to model-generated answers in natural QA.

Tasks and Metrics. We evaluate controlled FRA tasks with verified local truth using Qwen2.5-7B-Instruct, and natural QA using Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct on NQ, Movies, AnswerableMath, PopQA, and HotpotQA. NQ supplies answer-bearing passages; HotpotQA uses distractor contexts. Additional evaluations examine Qwen14B/32B. We report accuracy and AUROC for truth recovery, matched-pair ranking for semantic edits, and error-positive AUROC for detection, supplemented by AP and Recall@20%. Task construction, data splits, and metric definitions are described in Appendices B.1–B.2 and C.1–C.4.

Baselines and Protocol. Controlled error diagnosis is compared with generation confidence and a direct error probe. Natural-QA comparisons include Confidence, SAT, and SAPLMA, with task-adapted ReDeEP on NQ and CHARM on Movies and AnswerableMath reported separately. Methods score the same valid natural answers within each comparison. Paired detector differences use 2,000 parent- or dependency-group bootstrap resamples to obtain 95% confidence intervals. Comparator definitions, adaptations, and statistical protocols are provided in Appendices C.3–C.4 and D.3.

Training Details. All base language models remain frozen. Readouts and normalization statistics are fitted on training data, with development-based configuration selection where specified and no test-set tuning. Natural-QA features are collected by replaying each model's generated answers. Parent or dependency groups determine data partitions. Decoding settings, study-specific fitting procedures, hyperparameter choices, and seed protocols are documented in Appendices B.1–B.2, C.8, and D.1–D.3.

Method Settings. The controlled diagnostic combines a frozen local-truth readout with the observed Boolean answer, without using reference truth in test-time scoring. For natural QA, task-specific verifiers define the truth of each realized answer proposition, and FRA-ER uses the equivalent error-positive target with mean and final answer-content states from five layers. Query and answer-prediction states are included only in feature ablations. Study-specific readout sources, observation events, normalization, and feature variants are described in Appendices B.1, C.5, and D.1–D.2.

5.1 Is Constraint Truth Readable Under Natural-Answer Errors?

The same-output evaluation tests whether the readout distinguishes verified targets that the natural answer category does not. Across four held-out parent topologies, all 16 grounded `on` answers are False, while the verifier labels eight targets True and eight False. Frozen readouts at blocks 20, 25, and 27 classify all 16 cases correctly (Figure 3). Identical natural answers therefore accompany distinct task-defined truths that remain recoverable from the measured states.

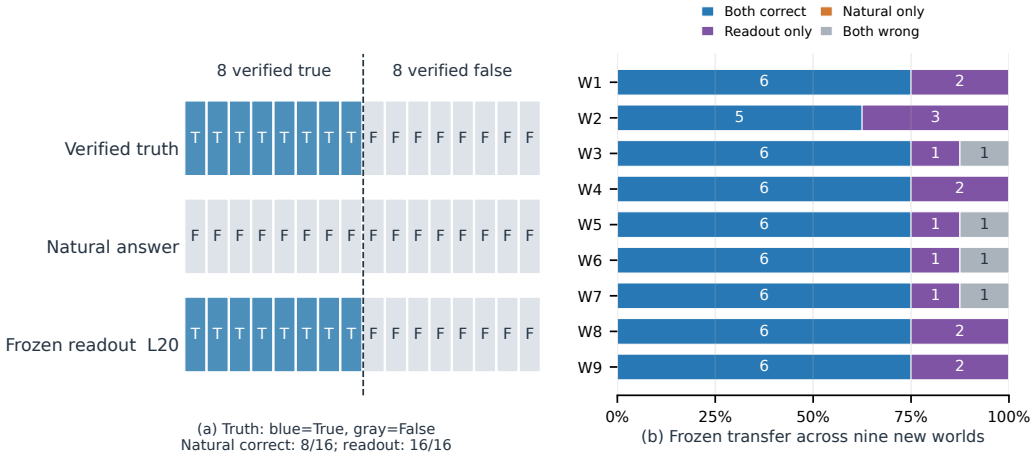

Figure 3: Recoverable distinctions beyond natural answers. (a) Sixteen grounded `on` cases with identical natural answers, balanced verified targets, and block-20 readout predictions. (b) Joint natural-answer/readout outcomes by world in the separate nine-world frozen-transfer evaluation (eight prompts per world).

We next apply readouts fitted on the original prompts, without refitting, to 72 grounded prompts from nine new complete worlds. Of the 19 natural-answer errors, the readout recovers the verified target in 15 cases. This extends the observation to new worlds, while retaining some shared local structural motifs. A further, separate two-precondition evaluation uses 1,152 examples from 48 held-out parents. At block 20, truth-readout AUROC/accuracy is 97.56%/93.23% under answer encoding 0 and 97.59%/92.53% under encoding 1; natural-answer accuracy is 60.76%. These results establish recoverability for the evaluated targets and readouts; they do not imply that the model uses the fitted direction to generate its answer. Complete layer-by-encoding results are in Appendix C.1.

5.2 Do Readouts Track Truth-Changing Task Factors?

Recoverability across examples does not by itself show that the score responds to semantic changes that determine the answer. We evaluate matched inputs that differ in one task factor and for which the external verifier confirms a truth flip. A pair is correctly ordered when the true endpoint receives the higher frozen score; ties count as one half. Facts, rule clauses, and action arguments define the edit types, while the prediction target remains the instantiated local truth in every pair.

The transfer evaluation applies the frozen factorial readouts to 320 prompts from 40 new parent worlds. At block 20, ranking rates are 88.8% for fact edits, 96.3% for rule edits, and 94.4% for action-argument edits. Across the prespecified blocks 20, 25, and 27, rates range from 86.9% to 98.1% across factors; blocks 6 and 13 are weaker (Figure 4). These are paired score-ordering rates, not classification accuracy. Parent-level summaries show world-to-world variation, with parents as

the resampling unit. The frozen readouts thus track truth-changing edits across all three factors in this new-world evaluation.

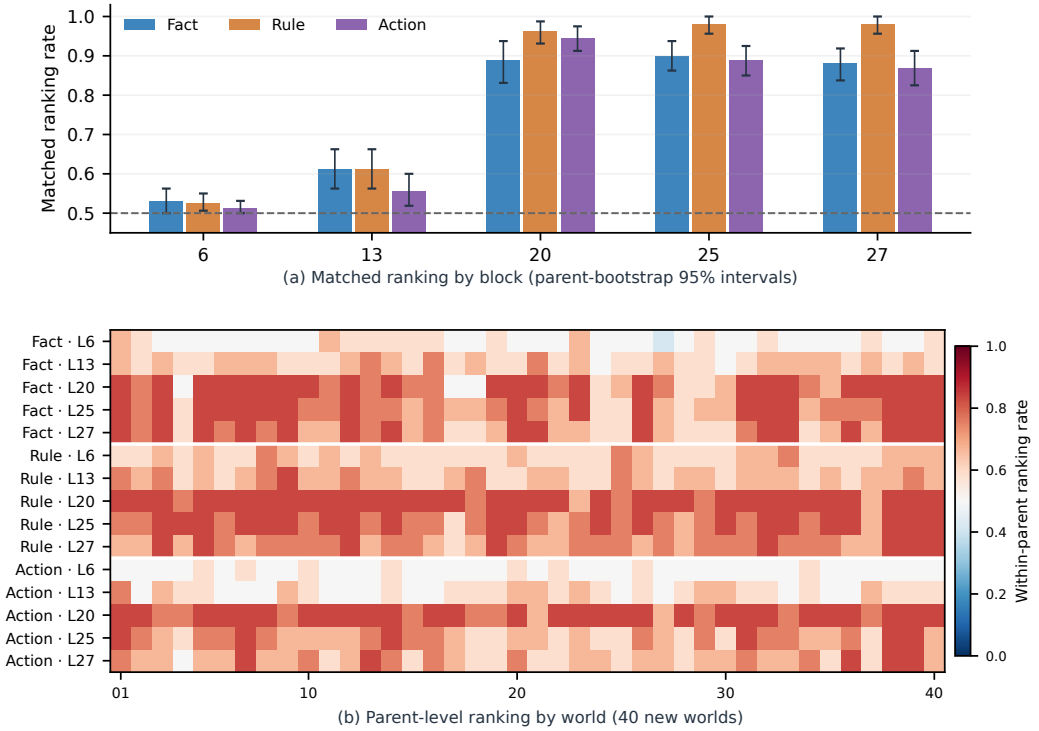

Figure 4: Factor responses across layers and worlds. (a) Matched ranking rates with nominal 95% parent-bootstrap intervals. (b) Parent-level ranking rates for 40 new worlds, three factors, and five blocks; each cell averages available matched edges, and parents are the resampling units.

Pairwise factor geometry, sublayer projections, and calibration analyses characterize the score within the separate factorial development split (Appendix C.2).

5.3 Does Readout–Answer Disagreement Diagnose Errors?

We evaluate readout–answer disagreement on the held-out two-precondition cohort using the separately trained, frozen truth readouts from the recoverability evaluation in Section 5.1. These readouts differ from those reused for the 40-world edit-transfer analysis. Section 4.2 defines the error-positive disagreement score and its hard-label identity: on the same examples, hard disagreement has the same classification error as the truth readout. The identity motivates the comparison but does not determine continuous-score ranking quality, which we evaluate on held-out predictions.

This post-hoc analysis uses 1,152 stored Qwen answers from 48 test parents, with no new generation or refitting. The block-20 disagreement score reaches 97.54% error AUROC; an equal-weight five-layer score reaches 96.46%, compared with 47.11% for generation confidence. A Direct error probe trained on controlled-task error labels is an auxiliary comparator (86.19% AUROC); its full specification, paired comparisons, and score-construction diagram are in Appendix C.3.

These results show that, when a verified local-truth readout is available, comparing it with the model's realized decision provides an effective error-ranking signal on the held-out controlled cohort. Section 5.4 extends the same diagnostic principle to natural QA. There, the realized answer itself instantiates the proposition to be evaluated, so its verifier-defined truth can be read from post-answer states and expressed in error-positive orientation. A supplementary source-allocation analysis is reported in Appendix C.7; it does not define or explain the disagreement score.

5.4 POST-ANSWER TRUTH DIAGNOSIS ON NATURAL QA

Natural QA instantiates the same truth-readout principle at the level of the realized answer. For each generated answer $\hat{Y}$, the task verifier determines whether $\hat{Y}$ satisfies the answer condition defined by the current input, providing the truth target $b_d^*(X_d, \hat{Y})$. FRA-ER estimates this answer-conditioned truth from post-answer hidden states and reports the resulting score in error-positive orientation.

Main Results. We fit FRA-ER with three training seeds and report mean test AUROC with standard deviation. FRA-ER has the highest mean AUROC among the four shared methods in nine of ten 7B/8B model–task settings (Table 1). The largest and most consistent gains over generation confidence occur on PopQA and HotpotQA across both model families, with positive gains also observed on NQ. Movies and AnswerableMath show smaller or task-dependent differences. Paired comparisons with generation confidence and secondary metrics are reported in Appendix C.4.

Table 1: Error detection across natural-QA tasks. Acc. is natural-answer accuracy, and detector scores are error-positive AUROC percentages. N is the shared valid-answer count. FRA-ER reports mean $\pm$ standard deviation over three training seeds; baseline aggregation follows each method's protocol in Appendix D.3. Confidence is deterministic for the fixed generations. Bold marks the highest AUROC mean in each row. Paired bootstrap comparisons and secondary metrics are reported in Appendix C.4.

Model	Task	N	Acc.	Conf.	SAT	SAPLMA	FRA-ER
Qwen7B	NQ	633	58.77	77.00	76.97	80.85	82.61 ± 0.31
	Movies	1478	3.86	89.49	83.42	86.97	91.64 ± 0.27
	Math	591	33.33	**87.87**	85.51	85.98	86.49 ± 0.38
	PopQA	984	15.55	83.04	90.53	92.98	93.90 ± 0.22
	HotpotQA	845	44.85	65.61	66.20	72.50	76.07 ± 0.41
Llama8B	NQ	697	61.41	70.47	77.60	77.19	78.25 ± 0.34
	Movies	1467	21.47	80.94	81.39	81.25	82.04 ± 0.29
	Math	585	27.35	90.52	89.00	90.01	91.72 ± 0.36
	PopQA	995	22.11	81.78	88.93	91.07	92.78 ± 0.25
	HotpotQA	911	54.56	62.86	69.18	72.38	75.01 ± 0.43

Table 2: Post-answer feature and depth ablations.

Variant	NQ	Math	HotpotQA
Confidence	77.00	**87.87**	65.61
Output-only	73.66	86.00	67.16
Event-only	78.53	85.57	70.56
Last-layer Post	79.28	86.96	73.61
Late-3 Post	81.30	87.19	75.31
FRA-ER	**82.63**	86.51	**76.05**
Full (+Event)	81.94	86.36	75.29

Qwen7B fixed-fit error-positive AUROC (%). Paired comparisons are in Appendix C.5.

Table 3: FRA-ER across Qwen model sizes.

Task	Size	Acc.	Conf.	FRA-ER
NQ	7B	58.77	77.00	**82.63**
	14B	65.36	74.23	**80.89**
	32B	65.35	74.46	**83.46**
Math	7B	33.33	**87.87**	86.51
	14B	48.50	88.99	**89.55**
	32B	65.04	**89.23**	85.16
PopQA	7B	15.55	83.04	**93.92**
	14B	20.53	83.75	**91.91**
	32B	20.84	88.52	**93.72**

Natural-answer accuracy and fixed-fit error-positive AUROC (%). Valid-answer counts are in Appendix C.6.

Feature and depth. The Post representation benefits from aggregating information across multiple layers on NQ and HotpotQA. NQ AUROC rises from 79.28% with the final layer to 81.30% with the late three layers and 82.63% with the fixed five-layer representation; HotpotQA shows the same ordering. AnswerableMath does not follow this pattern, and adding query and answer-prediction states does not consistently improve performance. We therefore retain the prespecified five-layer Post representation across tasks rather than selecting feature depth separately for each dataset.

Scaling behavior. We fit the same FRA-ER design separately at each Qwen model size. FRA-ER remains above generation confidence on NQ and PopQA at 7B, 14B, and 32B, whereas AnswerableMath shows a mixed pattern. Detector AUROC itself is not monotonic with model size. These results show that post-answer diagnostic information is not confined to the 7B model, while its value relative to generation confidence remains task-dependent.

Computational cost. Acquiring Post features adds measurable cost, whereas normalization and scoring are inexpensive. On an A800 80GB, generation takes .1685 s/example for Qwen7B and .2097 s/example for Llama8B; generation plus Post acquisition takes .2514 and .3063 s/example.

The corresponding acquisition overheads are 49.21% and 46.06%, while normalization and scoring add .2253 and .2493 ms/example. The summed totals are accounting estimates from separately measured components, not independent end-to-end measurements (Table 4; Appendix D.5).

Table 4: Generation and post-answer acquisition cost.

Model	Generation (s/ex)	Gen.+Post (s/ex)	Norm.+score (ms/ex)	Sum est. (s/ex)	Acq. overhead (%)	Throughput retention (%)
Qwen7B	.1685	.2514	.2253	.2516	49.21	67.02
Llama8B	.2097	.3063	.2493	.3066	46.06	68.46

Means over three measured repeats of 200 fixed inputs per model after 10 warm-up inputs. CPU normalization and scoring were measured separately; Sum is an accounting estimate. Repeat standard deviations and resource details are in Appendix D.5.

6 Discussion and Conclusion

The controlled FRA experiments show that language-model errors need not coincide with an absence of task-defined truth from the measured hidden states. Frozen readouts remain accurate when natural answers fail, systematically track truth-changing edits to facts, rules, and action arguments, and provide strong error-ranking signals when compared with the model's realized decision. Natural QA extends the same truth-readout perspective to generated answers: the realized answer instantiates the proposition being evaluated, the task verifier defines its truth, and FRA-ER estimates this answer-conditioned target from post-answer states in error-positive orientation.

The controlled results establish recoverability and semantic responsiveness, but not a general causal mechanism for target selection or decision formation. In the supplementary query-switch analysis, no tested block consistently favors the newly queried truth across both query conditions, and the tested activation replacements do not reliably repair model outputs (Appendix C.7). We therefore interpret the readouts as diagnostic measurements rather than demonstrated causal control variables. More broadly, the results suggest that part of language-model unreliability can arise even when task-relevant truth remains internally recoverable, making the relation between recoverable truth and realized decisions a useful target for error diagnosis.

Reproducibility Statement

The appendices specify the controlled-task construction and verifier, natural-QA datasets and answer checks, grouped data partitions, observation positions, readout and detector fitting, baseline adaptations, and paired resampling. They report the extended truth-readout, semantic-edit, diagnostic, and natural-QA results together with feature and scale comparisons and resource accounting. Base language models remain frozen; fitted measurements use the stated training groups and held-out evaluation groups.

AI Use Statement

We employed large language models (LLMs) to support the writing process, specifically for improving clarity in technical explanations, refining grammar and style, and enhancing overall readability. LLMs were also used to a limited extent to aid the process of finding related works. All research contributions, including the design of experiments, data analysis, and conclusions, are entirely our own. The LLMs were used strictly as writing aids to improve presentation quality, not for generating research content or shaping the substance of our work.

References

Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. https://aclanthology.org/2020.acl-main.385/.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. Do Language Models Know When They're Hallucinating References? In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 912–928, 2024. https://aclanthology.org/2024.findings-eacl.62/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. https://arxiv.org/abs/1610.01644v1.

Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023. https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1), pp. 207–219, 2022. https://aclanthology.org/2022.cl-1.7/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2212.03827.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3882–3890, 2020. https://www.ijcai.org/Proceedings/2020/537.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al.. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. https://arxiv.org/abs/2407.21783v1.

Nouha Dziri, Ximing Lu, Melanie Sclar, et al.. Faith and Fate: Limits of Transformers on Compositionality. In *Advances in Neural Information Processing Systems*, pp. 70293–70332, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/deb3c28192f979302c157cb653c15e90-Abstract.html.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, pp. 160–175, 2021. https://aclanthology.org/2021.tacl-1.10/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017), pp. 625–630, 2024. https://www.nature.com/articles/s41586-024-07421-0.

Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *International Conference on Learning Representations*, pp. 36391–36413, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring Latent World States in Language Models with Propositional Probes. In *International Conference on Learning Representations*, pp. 19337–19359, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/3132d0405fabe24b2a7b6cd7ba9de6b5-Abstract-Conference.html.

Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Haggai Maron. Neural Message-Passing on Attention Graphs for Hallucination Detection. In *International Conference on Learning Representations*, pp. 449–475, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/014898b821e3ce966ae1db68cb8d9966-Abstract-Conference.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 9574–9586, 2021. https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021. https://aclanthology.org/2021.emnlp-main.446/.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. https://aclanthology.org/D19-1275/.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019. https://aclanthology.org/N19-1357/.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9, pp. 962–977, 2021. https://aclanthology.org/2021.tacl-1.57/.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al.. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. https://arxiv.org/abs/2207.05221.

Najoung Kim and Sebastian Schuster. Entity Tracking in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3835–3855, 2023. https://aclanthology.org/2023.acl-long.213/.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020. https://aclanthology.org/2020.emnlp-main.574/.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. ACPBench: Reasoning About Action, Change, and Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25), pp. 26559–26568, 2025. https://doi.org/10.1609/aaai.v39i25.34857.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2302.09664.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al.. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7, pp. 452–466, 2019. https://aclanthology.org/Q19-1026/.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit Representations of Meaning in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1813–1827, 2021. https://aclanthology.org/2021.acl-long.143/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, pp. 41451–41530, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *arXiv preprint arXiv:2205.14334*, 2022. https://arxiv.org/abs/2205.14334.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 3214–3252, 2022. https://aclanthology.org/2022.acl-long.229/.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, 2023. https://aclanthology.org/2023.emnlp-main.291/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, pp. 157–173, 2024. https://aclanthology.org/2024.tacl-1.9/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. https://aclanthology.org/2023.acl-long.546/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. https://aclanthology.org/2023.emnlp-main.557/.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling*, 2024. https://arxiv.org/abs/2310.06824.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019. https://aclanthology.org/P19-1334/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *International Conference on Learning Representations*, pp. 66880–66913, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020. https://aclanthology.org/2020.acl-main.420/.

Qwen, An Yang, Baosong Yang, et al.. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2025. https://arxiv.org/abs/2412.15115v2.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. https://aclanthology.org/2021.eacl-main.295/.

Kyle Richardson and Ashish Sabharwal. What Does My QA Model Know? Devising Controlled Probes Using Expert Knowledge. *Transactions of the Association for Computational Linguistics*, 8, pp. 572–588, 2020. https://aclanthology.org/2020.tacl-1.37/.

Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019. https://aclanthology.org/P19-1282/.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability. In *International Conference on Learning Representations*, pp. 50250–50279, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/7daf60e805e596c3bd1e843e72ea5560-Abstract-Conference.html.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021. https://aclanthology.org/2021.findings-acl.317/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 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*, pp. 5433–5442, 2023. https://aclanthology.org/2023.emnlp-main.330/.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the Planning Abilities of Large Language Models - A Critical Investigation. In *Advances in Neural Information Processing Systems*, pp. 75993–76005, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/efb2072a358cefb75886a315a6fcf880-Abstract-Conference.html.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Advances in Neural Information Processing Systems*, pp. 38975–38987, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. https://aclanthology.org/2020.emnlp-main.14/.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019. https://aclanthology.org/D19-1002/.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at Scale: Identifying Causal Mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, pp. 78205–78226, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract.html.

Jian Xie, Kexun Zhang, Jiangjie Chen, Siyu Yuan, Kai Zhang, Yikai Zhang, Lei Li, and Yanghua Xiao. Revealing the Barriers of Language Agents in Planning. 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)*, pp. 1872–1888, 2025. https://aclanthology.org/2025.naacl-long.93/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2306.13063.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. https://aclanthology.org/D18-1259/.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models. In *International Conference on Learning Representations*, pp. 32097–32121, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/880ee6e23b6ee548ebcccafe0ef0f513-Abstract-Conference.html.