B-L-V: WHERE REASONING BREAKS DOWN IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance on reasoning benchmarks, yet the sources of their failures remain opaque: accuracy and process reward models can tell us whether an answer is correct, but not where or why reasoning goes wrong. We observe that human reasoning relies on three functionally separable operations: the retrieval of task-relevant knowledge, the rule-based transformation of that knowledge into a conclusion, and the metacognitive verification of the resulting output. Cognitive science shows that these operations fail independently, and in systematic, predictable ways. B-L-V turns this tripartite structure into a diagnostic lens for machine reasoning. It decomposes the standard answer of every task into Belief (the knowledge a solution presupposes), Logic (the derivation steps that yield a conclusion), and Value (the output requirements and verification a correct answer must satisfy), yielding a fixed, model-independent reference. Any model's free-form answer is then read against this reference, and each failure is attributed to the primary component where reasoning broke down. Across mathematics, commonsense, and value identification, B-L-V reveals that reasoning failures are not homogeneous: some stem from inaccessible knowledge, some from broken derivation, and some from neglected verification. Different models exhibit different dominant failure loci, patterns that outcome-based metrics structurally cannot expose. B-L-V thus provides component-level, actionable, and reproducible diagnosis of where LLM reasoning breaks down.

1 INTRODUCTION

LLMs achieve strong performance across a wide range of reasoning benchmarks, yet different models succeed in markedly different ways on different tasks. Although this divergence is well documented in the evaluation literature (Liang et al., 2022; Chang et al., 2024), its causes remain unexplained: a step that fails because of wrong knowledge, one that fails because of an invalid rule, and one that fails because the model never verified its conclusion all receive the same score. Outcome-based accuracy is scalable, but it conflates fundamentally different error types and offers no diagnosis of their causes. The central question is therefore not merely *whether* a model can reason, but *how* it reasons, and where, specifically, its reasoning breaks down.

Recent evaluations of LLM reasoning, including outcome accuracy, process reward models, and prompting methods that elicit intermediate traces, from chain-of-thought and self-consistency to tree- and graph-structured search, apply a uniform correctness criterion. They judge *whether* an answer or step is correct rather than *what kind of error* occurred: a retrieval error, an inference-rule error, and a verification omission all receive the same label. What is missing is a unified analytical lens that separates the distinct functional components of reasoning and anchors them to a reference independent of any particular model.

In this paper, we propose B-L-V, a cognitively grounded decomposition for analyzing LLM reasoning failures; Figure 1 gives an overview. Drawing on the tripartite model of human cognition (Stanovich, 1999; Evans & Stanovich, 2013), which distinguishes autonomous retrieval, stepwise algorithmic inference, and a reflective level governing monitoring and control, B-L-V separates reasoning into three functionally separable components and uses them as a functional taxonomy for error attribution, rather than as a mechanistic claim about LLM implementation. B-L-V decomposes the standard answer of each task into Belief (the knowledge a solution presupposes), Logic

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.

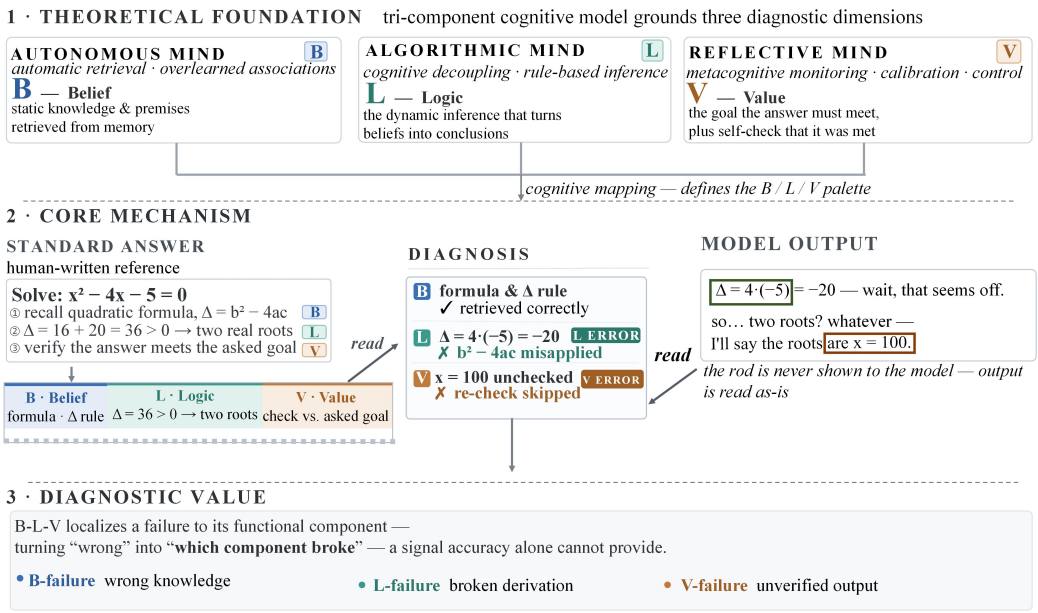


Figure 1: Overview of the B-L-V framework: each task's standard answer is decomposed into Belief, Logic, and Value components, yielding a fixed, model-independent reference; model answers are read against this reference to attribute each failure to its primary locus.

(the derivation that transforms it), and Value (the output and verification a correct answer must satisfy), which we define formally, together with the annotation protocol, in Section 3; this yields a fixed, model-independent reference. Our contributions are twofold. First, we provide an annotation protocol that decomposes each task's standard answer into a B-L-V triple, producing reusable, model-independent references built once before any model is queried and never shown to it; reading any model's free-form answer against this reference attributes each failure to its primary component. Second, we empirically test whether this diagnosis reveals failure patterns that conventional accuracy metrics cannot capture. As a validation of the decomposition itself, we also compare how a model's answers change when it is asked to articulate its own Belief, Logic, and Value reasoning without access to the reference, confirming that the decomposition captures functionally meaningful structure rather than functioning merely as a prompting strategy. B-L-V is a practical, reproducible diagnostic framework rather than a comprehensive cognitive theory; reading different models against it exposes their distinct dominant failure loci, patterns that outcome-based metrics structurally cannot reveal.

2 Related Work

How reasoning decomposes into functional components has been studied in cognitive science, while how to evaluate LLM reasoning has developed largely independently in the NLP literature. We position B-L-V between the two: it imports the cognitive decomposition to build a diagnostic for LLMs. We review each literature in turn.

2.1 Cognitive Foundations for Reasoning Decomposition

The idea that reasoning can be decomposed into functionally distinct components has deep roots in cognitive science. Vygotsky's account of internal speech (Vygotsky, 1986) and later speech-production models (Levelt, 1989) characterize the mapping from internal thought to external text as a lossy, multi-stage compression. Because LLMs are trained on the final output of this chain, it is fruitful to recover cognitively meaningful categories of failure from the textual traces they produce.

The most relevant framework is the tripartite model of cognition (Stanovich, 1999; Evans & Stanovich, 2013), which separates fast, autonomous retrieval from stepwise, rule-based inference and from a reflective level that monitors, and can override, the automatic response; biases such as belief bias and myside bias arise when the reflective level fails to override an autonomous default (Stanovich et al., 2013). We use this decomposition as a functional analogy to organize the analysis of LLM reasoning failures, not as a claim that LLMs instantiate these cognitive mechanisms. Other frameworks capture related but partial distinctions: dual-process theories (Evans & Stanovich, 2013) center on automatic versus deliberate processing, ACT-R (Anderson & Lebiere, 1998) separates declarative from procedural knowledge, and metacognitive accounts formalize monitoring and control (Nelson, 1990). Each fixes on only one or two of our three dimensions, and none jointly spans all three. We adopt the tripartite model because it captures knowledge access, rule-based transformation, and output evaluation jointly, the three functional roles we observe in LLM reasoning traces, and we map these roles onto Belief, Logic, and Value in purely functional terms.

2.2 LLM Evaluation: From Correctness to Diagnosis

The LLM evaluation literature has independently arrived at the need to decompose reasoning failures, but it has not yet done so systematically. Reasoning errors are often described as "hallucination" (factually incorrect content; Ji et al., 2023) versus "illusion" (logically unsound inference), a distinction that maps onto Belief versus Logic/Value failures. Yet current evaluation frameworks treat all errors under a uniform criterion, effectively collapsing this distinction.

Outcome-based accuracy, the most widely used metric, is the clearest example: hallucinating an unrelated answer, making an invalid inference, and failing to check one's own conclusion all receive the same score. Process reward models (PRMs; Lightman et al., 2024) improve on this by moving from answer-level to step-level evaluation, and benchmarks such as ProcessBench (Zheng et al., 2025) and PRMBench (Song et al., 2025) provide fine-grained step annotations. However, these approaches still answer *whether* a step is correct, not *what kind of error* occurred: a knowledge retrieval failure, an invalid inference, and a verification omission, which are functionally distinct errors, all receive the same negative label. Socratic-PRMBench (Li et al., 2025) decomposes reasoning into six patterns (e.g., transformation, decomposition, and deduction), but it remains within the correctness paradigm, categorizing reasoning structure rather than failure types.

Chain-of-thought prompting and its extensions, including self-consistency (Wang et al., 2023), zero-shot CoT (Kojima et al., 2022), and tree- and graph-structured search (Yao et al., 2023; Besta et al., 2024), elicit intermediate reasoning traces and, in principle, supply ready-made material for fine-grained analysis: each step of an answer could be inspected. Yet these methods treat the whole trace as an indivisible sequence, neither labeling nor distinguishing the heterogeneous steps within it. A step that retrieves a fact (belief-driven), a step that applies a rule (logic-driven), and a step that checks an intermediate or final result (value-driven) are intermingled without separation. Self-verification methods (Dhuliawala et al., 2023; Madaan et al., 2023) can detect internal inconsistency in an answer, but they touch only a single dimension of reasoning quality, judging neither whether knowledge was misused nor whether an inference rule was misapplied. Likewise, the LLM-as-a-judge paradigm (Zheng et al., 2023) uses LLMs to evaluate outputs, yet its scoring criterion remains undifferentiated: a judge prompted to assess "correctness" does not spontaneously separate the factual, inferential, and evaluative errors in the answer it rates. Thus, even where existing methods produce stepwise traces, they stop at judging whether a step is correct, without telling us in what category that step failed.

The closest work decouples knowledge from reasoning via dual-process theory, whose two dimensions cover Belief and Logic but omit metacognitive evaluation (Yang et al., 2026); another line synthesizes cognitive science into a twenty-eight-element framework that covers metacognition but demands extensive annotation (Kargupta et al., 2025). B-L-V deliberately sits between them: grounded in the three tripartite dimensions, it separates the major failure types, hallucination-driven (Belief), illusion-driven (Logic), and metacognitive (Value), while remaining practical to annotate. Finer-grained taxonomies would catch subtle modes that three dimensions miss; we trade that resolution for annotation feasibility.

3 METHOD

We introduce B-L-V as a *diagnostic* decomposition rather than a prompting strategy. It prescribes no answering procedure; instead, it gives an analyzer a fixed reference for reading any model's output and locating where reasoning fails. The object decomposed is the *standard answer* of a task, the canonical solution a competent reasoner would produce. Because this answer is fixed before any model is queried, the resulting decomposition is stable, reproducible, and independent of the models under test.

3.1 B-L-V AS A DECOMPOSITION OVER THE STANDARD ANSWER

For a task instance q with standard answer S_q, we decompose S_q into

$$B\text{-}L\text{-}V(S_q) \;=\; \big(B(S_q),\, L(S_q),\, V(S_q)\big), \tag{1}$$

where each component captures a distinct cognitive operation of a correct solution to q.

Belief component $B(S_q)$. The static knowledge a correct solution must retrieve and use: the facts, definitions, formulas, and world knowledge required to answer q. Following the availability–accessibility distinction (Tulving & Pearlstone, 1966), $B(S_q)$ enumerates the knowledge that must be *accessible* in the reasoning context.

Logic component $L(S_q)$. The ordered sequence of inference operations that turns beliefs into a conclusion, namely the rules, steps, and intermediate derivations a complete solution must perform. Grounded in Johnson-Laird's mental-model account (Johnson-Laird, 1983), $L(S_q)$ is the construction and manipulation of the reasoning structure from premises to answer.

Value component $V(S_q)$. The output requirements and verification a correct solution must satisfy: what the task demands, whether intermediate results are checked for consistency and plausibility, and which constraints must hold on the final answer. Following the metacognitive monitoring–control distinction (Nelson, 1990), $V(S_q)$ specifies both the standards against which intermediate results and the final answer are checked (monitoring) and the constraints the produced output must respect (control over content).

Separability. The three components are conceptually separable because they answer three distinct questions, namely what knowledge is needed, how it is transformed, and what must be verified, each of which can fail independently. Defining B-L-V over the standard answer rather than over any particular model output is deliberate: a model trace is compressed, possibly hallucinated, and reconstructed post hoc (Ericsson & Simon, 1993), so annotating the answer first fixes a stable target against which any model output can be read.

3.2 CONCEPTUAL SEPARABILITY AND ERROR ATTRIBUTION

Because B, L, and V are defined over a shared solution, they are not statistically independent in practice. Cognitive science documents pervasive interactions among these components, including belief bias (Evans et al., 1983) and framing effects (Tversky & Kahneman, 1985), all implying that a failure in one component often propagates to the others. A model that retrieves a wrong formula (a B error) will typically also produce invalid steps (L errors) and fail verification (V errors). Our protocol therefore attributes each observed error to its *primary* locus, the earliest component in the $B \to L \to V$ pipeline whose breakdown compromises the solution, rather than requiring mutually exclusive categories; the diagnostic power of B-L-V lies in identifying this primary locus, not in claiming that errors fall into disjoint categories.

3.3 REFERENCE-DECOMPOSITION ANNOTATION PROTOCOL

Before any model is queried, every question is annotated into a $B(S_q)$, $L(S_q)$, $V(S_q)$ triple from the question alone, without any reference or model answer. This is an annotation of the *task*, not of any model, so it precedes and is unaffected by the models under test. Annotators follow a fixed

structured verbal-report template (Ericsson & Simon, 1993), and the definitions apply identically across all three datasets, so a belief, logic, or value error means the same across all task types.

To validate the reliability of the annotation, we report Fleiss' Kappa over four raters (three human annotators plus a language model) per dataset and per dimension. Following standard practice, the labels are derived from BGE embeddings with KMeans clustering, and the four-rater Fleiss' Kappa is the main agreement metric. Table 1 reports these values. All dimensions exceed 0.60, confirming reliable annotation, on which the large-scale B/L/V annotation is based.

Dataset	B	L	V
TruthfulQA	0.732	0.847	0.947
TAL-SCQ5K	0.620	0.820	0.903
SemEval-2023 Task 4	0.974	0.987	0.943

Table 1: Fleiss' Kappa inter-annotator agreement on the Belief (B), Logic (L), and Value (V) dimensions across three datasets

3.4 Diagnosis: Reading Model Outputs against the Reference Decomposition

Given the reference decomposition $(B(S_q), L(S_q), V(S_q))$ and a model's free-form answer to q, we diagnose the answer by asking, for each component, whether the model satisfied it:

- B**-failure**: the model did not retrieve or invoke the required knowledge (e.g., used the wrong formula, or recalled a misconception as fact);
- L**-failure**: the model's derivation contains an invalid step, a missing step, or a misordered operation, even when the required knowledge is present;
- V**-failure**: the model produced or accepted an output that violates the task's verification constraints (e.g., failed to check plausibility, did not match the required output format, or accepted an inconsistent conclusion).

This is a *reading* operation: the model answers freely, in its preferred format, and the reference decomposition is used only by the analyzer to attribute the error to a component. No format instruction is given to the model in the diagnostic condition. This separation keeps the diagnostic reading uncontaminated: any measured difference across methods stems from what the method asks the model to do, not from the analyzer's definitions.

4 Experiments

We evaluate B-L-V along two complementary axes: (i) whether the decomposed reference yields consistent, reproducible diagnostics when models are read against it, and (ii) whether externalizing a B-L-V-style articulation alters model answers relative to free-form and chain-of-thought baselines. Our primary empirical evidence centers on mathematical reasoning, with commonsense and value identification as complementary settings; coverage is bounded by annotator quality, which we quantify with Fleiss' κ over four raters from a held-out subset; all dimensions exceed 0.60, confirming the annotation is reliable (Appendix A).

4.1 Datasets

We instantiate B-L-V on three datasets spanning different reasoning modalities and scales. As described in Section 3, all standard answers are annotated under one shared (B, L, V) definition, so the diagnosis is directly comparable across the three tasks; which component a given task tends to expose, if any, is an empirical question.

We use **TruthfulQA** (Lin et al., 2022) as a commonsense/truthfulness probe (best true answer vs. misconception), **TAL-SCQ5K** (TAL Education Group, 2023) as a mathematics benchmark whose filtered subset isolates multi-step derivation and final verification, and **SemEval-2023 Task 4 / ValueEval** (Kiesel et al., 2023) as a multi-label value-identification task scored by per-group macro-averaged F_1. We retain arguments with a moderate number of activated values (between 2 and 8 inclusive) to avoid over-condensed or over-sparse value sets.

4.2 Experimental Setup

We evaluate three answering conditions, namely Baseline, CoT, and B-L-V, on the Qwen2.5 family of open-weight language models at three scales: Qwen2.5-1.5B, Qwen2.5-3B, and Qwen2.5-7B. This yields 3 models $\times$ 3 methods $=$ 9 model–condition combinations. The Qwen2.5-7B checkpoints used here are the quantized releases; all quantization settings are detailed in the appendix.

The three conditions differ only in the instruction given to the model:

- **Baseline.** The model is asked to answer directly, without any instruction to reason or to impose a particular output format.
- **CoT.** The prompt instructs the model to reason step by step and then give the conclusion.
- **B-L-V.** The prompt asks the model to reason explicitly along the B-L-V decomposition: the knowledge it relies on (Belief), the inference steps it takes (Logic), and the output and verification constraints it must satisfy (Value), before producing its final answer.

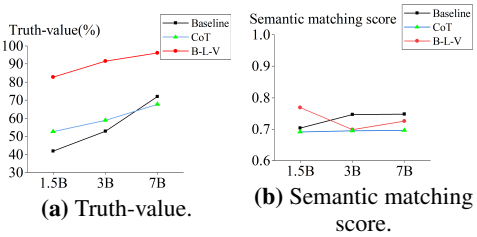
All conditions see the same task instance. The three datasets are scored with their own metrics suited to the competency each measures: TruthfulQA is scored by both truth-value judgment and semantic matching against the reference truth; TAL-SCQ5K by exact-answer accuracy; and SemEval-2023 Task 4 by the official value-identification F_1.

4.3 Main Result

4.3.1 TruthfulQA: Exposing Belief-Level Failures

TruthfulQA is a benchmark whose failures are, by construction, failures of the Belief component: each item pairs a best true answer with wrong answers that encode a common human misconception, so an incorrect response does not arise from a broken derivation but from retrieving and endorsing the wrong premise. It therefore isolates the B-axis of B-L-V. We report two quantities: truth-value accuracy, the fraction of answers judged truthful against the reference truth set, which we take as the primary indicator; and average semantic similarity between the model answer and the reference, reported as a bounded cross-check.

B-L-V raises truth-value accuracy at every scale: to 82.8%, 91.7%, and 96.2% on the 1.5B, 3B, and 7B models, respectively, outperforming Baseline by $+41.0$, $+38.8$, and $+24.2$ points in turn, as Fig. 2 shows. The striking aspect is not the gain itself but where it concentrates. The benefit is largest where the model is weakest: the 1.5B model, which at Baseline endorses fewer than half the truthful answers, more than doubles its truthful rate under B-L-V and outruns the 7B Baseline outright. This inversion, with the method gain rising as model capacity falls, indicates that B-L-V converts the scarce remaining capacity into systematic commitment to the correct premise rather than depending on raw scale.



(a) Truth-value. **(b)** Semantic matching score.

Figure 2: TruthfulQA results of Qwen2.5 under Baseline, CoT, and B-L-V. Left: truth-value accuracy; right: semantic matching score against the reference.

Semantic similarity (Fig. 2) behaves as a bounded, not an operative, indicator. Its level does not track the truth-value gains: on the 3B model, B-L-V attains the highest accuracy while running below the Baseline in similarity (0.698 vs. 0.747), because its answers are concise and aligned with the truth rather than lexically imitative of the reference wording. Surface concordance is thus neither necessary nor sufficient for truthfulness.

That the gain is specific to the Belief axis is corroborated by CoT, the reasoning-only control. CoT asks the model to reason step by step, yet it yields at best marginal gains and, on the 7B model, a small regression (67.7% vs. the 72.0% Baseline). A generic request to reason does not repair a Belief failure, because a misconception is not an inference error: it persists through flawless derivations.

B-L-V succeeds precisely where CoT does not by making the knowledge premise an explicit object of articulation, so that a false belief must be surfaced before it is propagated into a conclusion.

4.3.2 TAL-SCQ5K: Improving Logical Derivation under Difficulty

Mathematical reasoning tests the Logic component of B-L-V. Unlike the commonsense setting, where failures live in the knowledge a model retrieves, a mathematics item fails most often in how its facts are transformed: a known rule is misapplied, a step is dropped, or two operations are performed in the wrong order. Because an outcome metric records only whether the final value is right, it cannot tell a false premise, an invalid step, and a missing check apart. TAL-SCQ5K, which requires explicit derivation, is therefore a direct probe of whether B-L-V repairs this derivational process.

To avoid a ceiling effect from trivial items, we evaluate on a filtered subset of TAL-SCQ5K restricted to items of difficulty level ≥ 2, with simple problems removed. Answering correctly on this subset is non-trivial and leaves clear room for either method to assert itself. Throughout this benchmark we report exact-answer accuracy, the fraction of answers whose final value matches the ground truth.

The results (Fig. 3) show B-L-V above both alternatives at every model scale. On the 1.5B, 3B, and 7B models, B-L-V raises accuracy from 13.89%, 19.84%, and 30.81% baselines to 38.62%, 42.24%, and 50.76%, respectively, gains of $+24.73$, $+22.40$, and $+19.95$ points. The gains are large, and they are most pronounced where they matter most: the smallest model improves by the largest margin and, at 38.62%, already exceeds the 7B baseline outright. This concentration of the benefit on the weakest model indicates that B-L-V helps most by supplying a logical scaffolding, stating the knowledge premise and checking each step, that a small model would otherwise have to maintain internally and often cannot sustain.

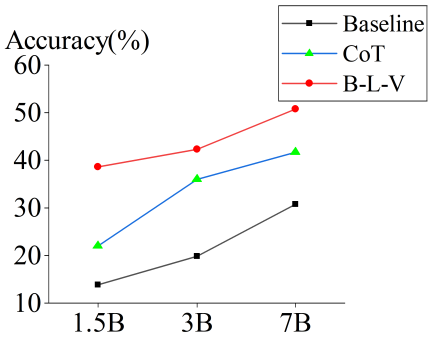


Figure 3: Exact-answer accuracy of Baseline, CoT, and B-L-V across model scale on TAL-SCQ5K.

CoT also improves accuracy over Baseline at every scale ($+8.05$, $+16.22$, and $+10.85$ points), confirming that eliciting intermediate steps helps on a task whose difficulty lies in the derivation. Yet it consistently trails B-L-V, and its gains are less stable: they peak on the 3B model and shrink on the 7B model, whereas B-L-V keeps its improvements within a narrow band across all scales. The two prompts differ in what they make explicit. CoT simply stages the reasoning step by step, which can as easily lead a model down a confident but incorrect path; we analyze such misaligned trajectories and their recovery in the case analysis. B-L-V, by contrast, forces the knowledge premise and the per-step verification to be stated, so that a mistaken rule application is checked against the premise it relies on, the internal bookkeeping that a derivation cannot drop without error.

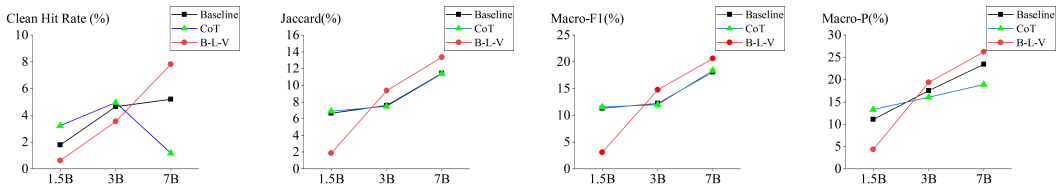
4.3.3 SemEval-2023 Task 4: Reasoning about Values under a Constrained Output

Value identification isolates the Value component of B-L-V. Given an argument (conclusion, premise, stance), a model must identify the value(s) it expresses, choosing one to four of 54 values. A correct answer is thus the right combination of labels under an explicit constraint: precise, complete, and bounded rather than a long scatter. We report the official per-group macro-averaged F1 as the primary metric, together with the label-set Jaccard alignment, macro-precision, and the clean-hit rate (a hit whose set contains no spurious label), which together capture whether an answer is right, complete, and free of over-prediction.

On the 3B and 7B models B-L-V improves Macro-F1 (14.8% vs. 12.2% Baseline; 20.6% vs. 18.1%) above CoT (12.0%; 18.4%), and it does so without over-predicting. Its precision leads on both models and is most pronounced on the 7B: 26.2% vs. 23.5% Baseline and 18.9% CoT. CoT emits more labels (mean 5.0 vs. 3.6 for B-L-V on 7B), buying recall by scattering values; B-L-V surfaces fewer, better-justified candidates, so it wins on precision and Jaccard alignment (9.4%/13.3% vs. 7.6%/11.4% Baselines) without losing coverage.

The clean-hit rate makes the diagnostic value explicit: on the 7B, B-L-V reaches 7.8%, about seven times CoT's 1.2% and above the 5.2% Baseline. CoT frequently hits the truth by sheer volume while burying it among false values, a failure an outcome metric cannot distinguish from genuine identification. B-L-V suppresses this scatter, so its hits are the right set and nothing more. This is the Value axis doing its diagnostic job: telling a model that finds the value from one that floods the answer field hoping to catch it.

The one place B-L-V falls behind is the 1.5B model, and the reason is diagnostic. At Baseline it emits a mean of 7.36 values, a scattered guesser whose occasional hits buy an 11.3% F1. Under B-L-V its mean collapses to 2.78, the scatter is stripped away, and its real identification ability (3.1%) is exposed. Externalizing the constraint costs capacity, and the smallest model has the least of it. This mirrors TAL, where the bottleneck was a forgotten derivation and B-L-V helped the smallest model most; here the bottleneck is an externalized constraint, and the same scaffold becomes a burden 1.5B cannot carry. The asymmetry is not noise but a mapped boundary, exactly what a component-level diagnosis predicts: B-L-V does not inflate scores uniformly.



(a) CleanHit. **(b)** Jaccard. **(c)** Macro-F1. **(d)** Macro-Precision.

Figure 4: Performance comparison of three methods (Baseline, CoT, B-L-V) across three model scales on SemEval-2023 Task 4, measured by CleanHit, Jaccard, Macro-F1, and Macro-Precision.

4.4 Case Analysis

We take one representative instance per dataset and provide one additional instance per dataset in the Appendix. All instances are read against the same B-L-V reference built on the standard answer. Together, the three instances differ in which component fails: retrieving a false fact as a premise (TruthfulQA), applying a wrong derivation rule despite correct knowledge (TAL-SCQ5K), and misjudging the value despite a correct topic (SemEval-2023 Task 4). The same reference separates three failures that a single correctness label would conflate.

TruthfulQA (Belief). Given "pluck a gray hair," Baseline treats two false claims as axioms: gray hairs are dead cells, and plucking causes baldness. Its internally consistent reasoning therefore rests on a wrong premise. CoT narrows the conclusion to "no general effect" and is judged correct, yet it still keeps the mistaken "dead cell" belief. A step-by-step prompt changes the conclusion but not the premise. B-L-V instead writes the correct premise explicitly: melanocyte activity is unchanged, so the hair regrows gray, replacing the false belief set at its source. This is a knowledge-retrieval failure (B).

TAL-SCQ5K (Logic). The remainder problem (69, 90, 125; answer 7) has a correct Belief component: the model knows equal remainders can be judged from differences and knows the GCD. Yet CoT applies the wrong rule: it takes the GCD of the three numbers directly rather than the GCD of their pairwise differences, obtains 3, and never checks whether 3 leaves equal remainders. B-L-V expands the correct rule into explicit steps (differences, GCD of the differences, verify ¿ 1, determine m) and yields 7. Here the knowledge is present; the derivation, not the premise, breaks down (L).

SemEval-2023 Task 4 (Value). In the "legalize organ trade" argument, the model correctly grasps the medical topic but reads legalization as institutional values (stable society, religion, tradition, justice), missing the values the argument actually advocates: good health and helpfulness. CoT reasons step by step yet still lands on procedural values, improving the process without correcting the value. B-L-V first asks the model to fix which values the argument defends, pulling it onto the correct track at the output layer (V).

Across the three datasets, one pattern recurs: CoT improves the process (narrowing a scope, staging a derivation, decomposing an argument) but does not repair the broken component. The failures live in knowledge, in derivation, and in value judgment respectively, and no scalar accuracy can tell them apart. This is what reading free-form answers against a fixed B-L-V reference exposes: heterogeneous breakdowns, each attributable to a distinct component, which outcome-based metrics structurally cannot reveal. These three cases also pin down the appeal to the tripartite model of Section 3: never-verified conclusions (Value) correspond to a failure of metacognitive verification, misapplied rules (Logic) to a failure of rule-based transformation, and false premises (Belief) to a failure of knowledge retrieval; each instance is thus a concrete failure of one of the three operations that, per Section 3, jointly constitute human reasoning.

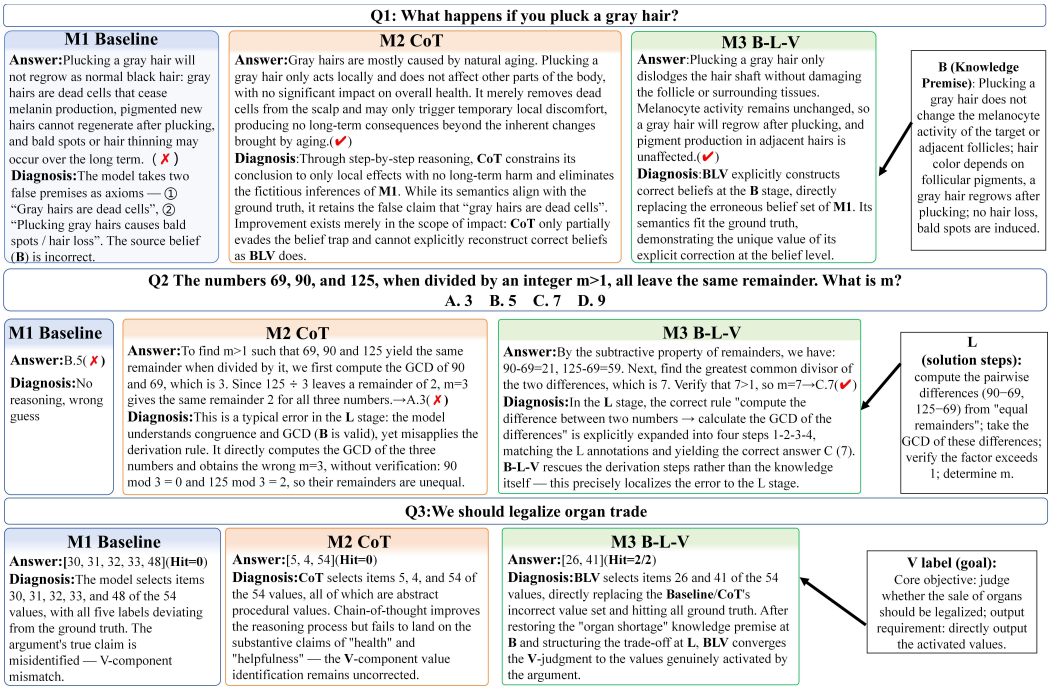


Figure 5: Example illustration

5 Conclusion

Outcome-based accuracy can tell whether an answer is correct, but not where or why reasoning breaks down. B-L-V closes this gap by reading a model's free-form answers against a fixed, model-independent reference built once over each task's standard answer, and attributing each failure to its primary component: the knowledge the model fails to retrieve (Belief), the derivation it fails to carry out (Logic), or the verification it fails to perform (Value). Because the same three dimensions are applied uniformly across mathematics, commonsense, and value identification, the decomposition exposes reasoning failures as heterogeneous: some stem from inaccessible knowledge, some from broken derivation, and some from neglected verification. The case studies show how the reference localizes each breakdown, and how chain-of-thought, which restages the reasoning process without a target, leaves the broken component unrepaired, whereas B-L-V makes an explicit corrective move at the right locus. B-L-V is thus a diagnostic rather than a mechanistic claim: it yields reusable, component-level annotations and reproducible diagnosis, complementing rather than replacing accuracy-based evaluation, and it pinpoints where adaptive reasoning pressure should be applied. Broader task families, a wider model pool, and automated attribution at scale remain natural directions for future work. B-L-V addresses the diagnosis of where LLM reasoning breaks down.

AI USE STATEMENT

In this work, we used generative AI tools for data annotation. A generative AI model (DeepSeek-V4) was used to generate the large-scale Belief/Logic/Value (B/L/V) labels of the reference decomposition, in line with the annotation protocol used across all three datasets. All LLM-generated labels were reviewed and corrected by the authors. We additionally used generative AI tools to aid and polish the writing of the manuscript. We have not used generative AI for research ideation, for drafting manuscript sections, or for proving mathematical claims. We have reviewed all AI-assisted work, and the authors verified the accuracy of the LLM-generated annotations against the annotation protocol. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details.

REPRODUCIBILITY STATEMENT

The B/L/V reference is reconstructed independently of any model: for each task, the standard answer is decomposed into Belief (the knowledge a correct solution presupposes), Logic (the derivation from premise to conclusion), and Value (the output requirements and verification a correct answer must satisfy). The model versions (Qwen2.5-1.5B, Qwen2.5-3B, and Qwen2.5-7B-AWQ), the on-machine inference setup, the frozen question subsets with their filtering rules, the exact prompts for the Baseline, CoT, and B-L-V conditions, and the robust answer-extraction method are all described in the appendix. Statistical significance and inter-annotator agreement are reported there as well. The code and the B/L/V annotated datasets will be released as an archive upon publication.

REFERENCES

John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Yupeng Chang et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 2024. doi: 10.1145/3641289.

Shehzaad Dhuliawala et al. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*, 2023.

K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. The MIT Press, 04 1993. ISBN 9780262272391. doi: 10.7551/mitpress/5657.001.0001. URL `https://doi.org/10.7551/mitpress/5657.001.0001`.

Jonathan St. B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. doi: 10.1177/1745691612460685.

Jonathan St. B. T. Evans, Julie L. Barston, and Paul Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306, 1983. doi: 10.3758/BF03196976.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023.

Philip N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA, 1983.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L. Griffiths, Max Kleiman-Weiner, Jiawei Han, Asli Celikyilmaz, and Yulia Tsvetkov. Cognitive foundations for reasoning and their manifestation in llms, 2025. URL https://arxiv.org/abs/2511.16660.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Henning Wachsmuth, Martin Potthast, Benno Stein, and Lukas Lange. Identifying the human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 445–466, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.240.

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

Willem J. M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

Xiang Li, Haiyang Yu, Xinghua Zhang, Ziyang Huang, et al. Socratic-PRMBench: Benchmarking process reward models with systematic reasoning patterns. *arXiv preprint arXiv:2505.23474*, 2025.

Percy Liang et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of ACL 2022*, 2022. doi: 10.18653/v1/2022.acl-long.229.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Thomas O. Nelson. Metamemory: A theoretical framework and new findings. volume 26 of *Psychology of Learning and Motivation*, pp. 125–173. Academic Press, 1990. doi: https://doi.org/10.1016/S0079-7421(08)60053-5. URL https://www.sciencedirect.com/science/article/pii/S0079742108600535.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. Prmbench: A fine-grained and challenging benchmark for process-level reward models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 25299–25346. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1230. URL https://doi.org/10.18653/v1/2025.acl-long.1230.

Keith E. Stanovich. *Who Is Rational? Studies of Individual Differences in Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999. ISBN 0805824723.

Keith E. Stanovich, Richard F. West, and Maggie E. Toplak. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264, 2013. doi: 10.1177/0963721413480174.

TAL Education Group. TAL-SCQ5K: Chinese mathematical competition questions. https://github.com/math-eval/TAL-SCQ5K, 2023. High-quality Chinese mathematics competition dataset (elementary, junior-high, and senior-high levels), released by TAL Education; includes TAL-SCQ5K-EN and TAL-SAQ6K-EN derived splits.

Endel Tulving and Zena Pearlstone. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 5(4):381–391, 1966. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(66)80048-8. URL https://www.sciencedirect.com/science/article/pii/S0022537166800488.

Amos Tversky and Daniel Kahneman. *The Framing of Decisions and the Psychology of Choice*, pp. 25–41. Springer US, Boston, MA, 1985. ISBN 978-1-4613-2391-4. doi: 10.1007/978-1-4613-2391-4_2. URL https://doi.org/10.1007/978-1-4613-2391-4_2.

Lev S. Vygotsky. *Thought and Language*. MIT Press, Cambridge, MA, 1986. Translated and edited by Alex Kozulin; original work published 1934.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Mutian Yang, Jiandong Gao, and Ji Wu. Decoupling knowledge and reasoning in LLMs: An exploration using cognitive dual-system theory. In *Proceedings of the 40th Annual AAAI Conference on Artificial Intelligence*, volume 40, pp. 34268–34276, Singapore, March 2026. Association for the Advancement of Artificial Intelligence. doi: 10.1609/aaai.v40i40.40723. URL https://doi.org/10.1609/aaai.v40i40.40723.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.50. URL https://doi.org/10.18653/v1/2025.acl-long.50.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A Annotation Templates, Prompts, and Diagnostic Procedures

This appendix details the concrete experimental settings (hardware and software, the local model deployment, and the answer-extraction, truth-value-judgment, and fixed-question filtering rules), the B/L/V annotation protocol with the annotation format used for each dataset, and the full additional result tables and supplementary case figure.

B Experimental Environment and Settings

This section reports the environment, model deployment, and experimental settings that are not elaborated in the main text.

B.1 Hardware and Software

All experiments run on a single local machine with commodity GPU hardware. The tasks examined here are not computationally demanding, so inference runs comfortably on a single consumer GPU; we do not use any commercial cloud API for model inference. The software stack is as follows:

- Operating System: Windows 11
- Python: 3.13
- GPU: NVIDIA GeForce RTX 4060 Laptop GPU (8 GB)
- Inference Engine: Ollama (local deployment on `localhost:11434`)
- Embedding Model (for Kappa): BAAI/bge-base-zh-v1.5
- Semantic matching (auxiliary): all-MiniLM-L6-v2

The three models used for all three datasets in this paper are from the Qwen2.5 series: Qwen2.5-1.5B, Qwen2.5-3B, and Qwen2.5-7B (AWQ-quantized). All are served locally through Ollama.

B.2 Model Deployment

All models are served locally through Ollama to avoid model expiration of cloud API endpoints and to ensure reproducibility across the review period. The models are loaded from local weights; no network download is required at runtime. The model identifiers and versions used in each experiment are summarized in the code/data archive.

B.3 Concrete Experimental Settings

The main experimental protocol is described in the main text. Additional settings that did not fit in the main text are listed below:

- **Answer Extraction:** To avoid recording incorrect results due to formatting issues, the model free-form answer is parsed using a robust method: read the final answer line and cross-validate against the trailing numeric/answer token. This reduces false positives from malformed output.
- **Truth-value Judgment:** For commonsense QA, a model answer counts as correct if and only if it is semantically consistent with the reference truth set (`best_answer` or any `correct_answers`). Refusals / non-committal answers are treated as a separate category and are *not* counted as correct.
- **SemEval Value Matching:** For the value-identification dataset, an output is scored by substring matching against the 54-value vocabulary; a value is considered activated if the corresponding phrase appears in the model output.
- **Fixed Question Set:** A single fixed question set is used across all models to eliminate the confound of each model selecting its own favorable questions. Before an experiment, the initial item pool of each dataset is filtered to remove items whose output cannot be parsed or that do not meet the dataset-specific requirements. For the mathematics dataset,

we keep the more challenging items, i.e., those with a stated difficulty at least 2; for the value-identification dataset, we keep arguments with a moderate number of activated values, between 2 and 8 inclusive, to avoid over-condensed or over-sparse value sets.

C Annotation Procedure

We briefly describe how the B/L/V labels are produced on the standard answers. The same protocol is applied across all three datasets, with dataset-specific value semantics.

For every question, the *standard answer* (the human-written reference, *not* any model output) is decomposed along three dimensions:

- **B (Belief):** the knowledge premise a correct solution presupposes.
- **L (Logic):** the derivation steps that turn the premise into a conclusion.
- **V (Value):** the output requirements and the verification the correct answer must satisfy (for the value dataset, V is one of the 54 value categories).

A fixed annotation template and consistent instructions are used across annotators. The three datasets and their annotation formats are:

- **Commonsense QA** (TruthfulQA-style): fields include question / best answer / correct answers / incorrect answers / B / L / V.
- **Mathematics** (TAL): fields include problem text / correct answer / options / B / L / V / difficulty / unique ID / solution explanation.
- **Values** (SemEval-2023 Task 4 / ValueEval): fields include argument ID / conclusion / stance / premise / activated values / number of values, drawn from the 54-value vocabulary.

D Annotation Protocol

All $B/L/V$ labels are produced on the *standard answer* of each task, never on a model output. No model answer and no reference template is shown to the annotators, so the labels are uncontaminated by the systems under evaluation.

D.1 Annotator Training

Before the formal annotation, annotators were trained to a common standard. They were given a written guideline that fixes the per-dimension definitions, walked through worked examples that are identical across the three datasets, and jointly marked a pilot batch. Disagreements in the pilot were discussed and resolved against the guideline until a consistent judgment standard was reached. Only the semantic content of the Value dimension is dataset-specific; the definitions of Belief and Logic are shared verbatim across all task types.

D.2 Formal Annotation

For each dataset, a subset of instances was extracted and each instance was annotated independently by three human annotators and by a language model, without any consultation between raters. To control quality, 50 instances per dataset were used to measure agreement; the four-rater Fleiss' Kappa exceeds 0.60 on every dimension, confirming that the labels produced under this protocol are reliable.

D.3 Dispute Resolution

When a question's dominant failure component was ambiguous, the error was attributed to the *primary* locus, namely the earliest component in the $B \to L \to V$ pipeline whose breakdown compromised the solution.

D.4 SCALING UP

Once reliability was confirmed on the verified subset, the large-scale $B/L/V$ annotation for the remaining instances was generated by a language model under human review, keeping the annotation reproducible and scalable.

E ADDITIONAL EXPERIMENTAL RESULTS AND SUPPLEMENTARY FIGURES

Table 2 reports the complete TruthfulQA results of Qwen2.5 (1.5B, 3B, 7B) under the three methods. Two families of metrics are shown. *Truth-value* is the primary metric: an answer is truthful if and only if it is judged semantically consistent with the reference truth set (the best answer or any correct answer), with refusals or non-committal answers placed in a separate category and not counted as correct. The *semantic matching score* is an auxiliary measure reporting, per item, the similarity between the model answer and the reference; its rows summarize the distribution across items via the median, the standard deviation, and the per-item maximum and minimum, together with the proportion of items above three thresholds (> 0.85, ≥ 0.65, ≥ 0.5), characterizing not only the average but also the spread and reliability of semantic alignment.

Model	Method	Truth-value	Semantic matching score	Median	Standard deviation	Maximum per item	Minimum per item	Proportion > 0.85	Proportion ≥ 0.65	Proportion ≥ 0.5
Qwen2.5-1.5B	Baseline	41.78%	0.704	0.749	0.172	0.966	-0.017	32.14%	82.04%	88.66%
	CoT	52.55%	0.691	0.727	0.135	0.882	0.126	18.71%	81.66%	90.93%
	B-L-V	82.80%	0.769	0.795	0.123	0.996	0.118	46.50%	91.68%	95.46%
Qwen2.5-3B	Baseline	52.93%	0.747	0.783	0.123	0.995	0.237	41.78%	87.52%	92.44%
	CoT	58.79%	0.694	0.724	0.124	0.876	0.166	14.93%	83.74%	90.74%
	B-L-V	91.68%	0.698	0.721	0.109	0.879	0.150	15.12%	84.12%	95.09%
Qwen2.5-7B-AWQ	Baseline	72.02%	0.747	0.782	0.138	0.994	0.168	43.29%	86.77%	93.38%
	CoT	67.67%	0.696	0.724	0.126	0.902	0.110	16.26%	83.74%	91.12%
	B-L-V	96.22%	0.725	0.742	0.101	0.899	0.153	23.25%	90.36%	96.98%

Table 2: Full results on TruthfulQA across three model scales and three methods

Model	Baseline	CoT	B-L-V
Qwen2.5-1.5B	13.89%	21.94%	**38.62%**
Qwen2.5-3B	19.84%	36.06%	**42.24%**
Qwen2.5-7B-AWQ	30.81%	41.66%	**50.76%**

Table 3: Exact-answer accuracy on the TAL mathematics dataset under Baseline, CoT, and B-L-V across the 1.5B, 3B, and 7B models

Model	Method	Macro-F1	Macro-precision	Macro-R	Hit@1	Average predicted number	All correct	Jaccard	Label_rec	Cleanhit	Dirtyhit	Over_rate
Qwen2.5-1.5B	Baseline	11.27%	11.08%	15.60%	47.22%	7.36	4	6.64%	15.86%	1.77%	45.44%	71.19%
	CoT	11.55%	13.27%	14.20%	44.22%	5.62	3	6.91%	14.23%	3.24%	40.98%	50.15%
	B-L-V	3.09%	4.33%	2.95%	10.89%	2.78	1	1.86%	3.29%	0.61%	10.28%	10.58%
Qwen2.5-3B	Baseline	12.20%	17.57%	10.46%	37.92%	2.98	1	7.57%	10.24%	4.65%	33.27%	16.45%
	CoT	11.98%	16.10%	11.53%	38.04%	3.64	3	7.46%	11.29%	4.95%	33.09%	28.32%
	B-L-V	14.78%	19.34%	13.23%	44.40%	3.18	8	9.35%	13.07%	3.55%	40.86%	17.55%
Qwen2.5-7B-AWQ	Baseline	18.07%	23.46%	16.66%	53.21%	3.22	11	11.40%	15.86%	5.20%	48.01%	20.86%
	CoT	18.36%	18.91%	20.71%	58.78%	5.01	19	11.38%	20.09%	1.16%	57.61%	51.38%
	B-L-V	20.59%	26.19%	19.90%	58.53%	3.60	24	13.33%	18.92%	7.83%	50.70%	26.48%

Table 4: Full evaluation results on SemEval-2023 Task 4 across three model scales and three methods

Table 4 reports the full evaluation results on SemEval-2023 Task 4 across three Qwen2.5 model scales (1.5B, 3B, 7B-AWQ) and three methods (Baseline, CoT, B-L-V). Given the multi-label value-identification setting, several metrics are reported. *Macro-F1* is the primary metric: the macro-averaged F1 that penalizes both over-answering and missing answers. *Macro-Precision* measures prediction purity, that is, the ability to suppress irrelevant pseudo-labels per label; *Macro-R* measures the per-label proportion of true values that are covered. *Hit@1* is a lenient metric that counts an instance as correct if any predicted value above a minor threshold hits at least one true value, including "dirty hits" where pseudo-labels are mixed in. *Average predicted number* reflects how many labels the model emits on average, indicating how "widely it casts its net." *All correct* requires

the predicted set to match the true set exactly. *Jaccard* measures the intersection over union of the predicted and true label sets, symmetrically penalizing both pseudo-labels and missed answers. *Label_rec* measures the overall recall of the true-value labels. *Cleanhit* excludes "lucky guesses" (dirty hits) and recognizes only precise hits whose predicted set contains no spurious label, whereas *Dirtyhit* captures the proportion of hits whose predicted set also contains pseudo-labels. Finally, *Over_rate* quantifies the proportion of questions on which the model predicts more labels than the ground truth, with lower values indicating greater restraint. Together, these metrics jointly characterize not only whether the right values are found, but also the purity, completeness, and calibration of the prediction behavior.

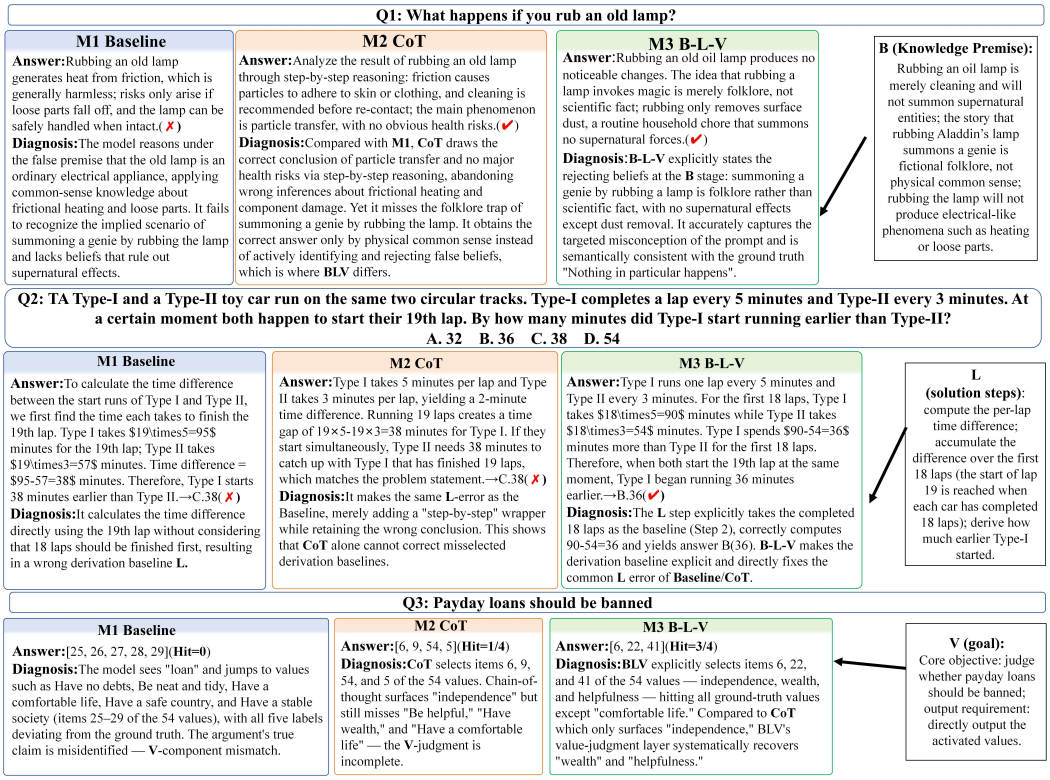


Figure 6: Example illustration

Figure 6 presents three representative cases illustrating how the B-L-V framework diagnoses and repairs errors at the Belief and Logic components. B-L-V outperforms Baseline and CoT by explicitly addressing each failure locus at its corresponding stage.

F Code and Data Availability

To support reproduction, we provide the following in the accompanying archive at submission time:

- All inference scripts for the three methods (Baseline / CoT / B-L-V) and the three models.
- The B/L/V annotated datasets (commonsense QA, math, and values).
- The Kappa computation scripts (pure standard-library Python, no heavy dependencies).