

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 TRUST: A DECENTRALIZED FRAMEWORK FOR AUDIT- ING LARGE LANGUAGE MODEL REASONING

Anonymous authors

Paper under double-blind review

Large Language Models (LLMs) can produce complex reasoning chains, offering a window into their decision-making processes. However, verifying the quality (e.g., faithfulness and harmlessness) of these intermediate steps is a critical, unsolved challenge. Current auditing methods are often centralized, opaque, and struggle to scale, creating significant risks for the deployment of proprietary models in high-stakes domains. This paper addresses four key challenges in reasoning verification: (1) *Robustness*: Centralized systems are single points of failure, vulnerable to attacks and systemic bias. (2) *Scalability*: The length and complexity of reasoning traces create a severe bottleneck for human auditors. (3) *Opacity*: Internal auditing processes are typically hidden from end-users, undermining public trust. (4) *Privacy*: Model providers risk intellectual property theft or unauthorized model distillation when exposing complete reasoning traces. To overcome these barriers, we introduce TRUST, a decentralized framework for auditing LLM reasoning. TRUST makes the following contributions: (1) It establishes a decentralized consensus mechanism among a diverse set of auditors, provably guaranteeing audit correctness with up to 30% malicious participants and mitigating single-source bias. (2) It introduces a scalable decomposition method that transforms reasoning traces into hierarchical directed acyclic graphs, enabling atomic reasoning steps to be audited in parallel by a distributed network. (3) All verification decisions are recorded on a transparent blockchain ledger, creating a permanent and publicly auditable record. (4) The framework is privacy-preserving by distributing only partial segments of the reasoning trace to auditors, thus protecting the full proprietary logic from distillation. We provide theoretical guarantees for the security and economic incentives of the TRUST framework. Experiments across multiple LLMs (e.g., GPT-OSS, DeepSeek-r1, Qwen) and reasoning tasks (e.g., mathematical, medical, science, and humanities) demonstrate that TRUST is highly effective at identifying reasoning flaws and is significantly more resilient to corrupted auditors than centralized baselines. Our work pioneers the field of decentralized AI auditing, offering a practical pathway for the safe and secure deployment of AI systems.

## 1 INTRODUCTION

The capabilities of large language models (LLMs) have expanded from text generation to complex, multi-step reasoning, leading to the development of Large Reasoning Models (LRMs) that produce explicit reasoning traces (Wei et al., 2022). While offering a view of a model’s logical flow, this explicit reasoning also exposes potential flaws, including logical errors, a lack of faithfulness to the model’s true internal state (Turpin et al., 2023), and safety vulnerabilities. The verification of these intermediate steps is a critical prerequisite for the safe and reliable deployment of LRMs in high-stakes domains such as medicine (Singhal et al., 2023), law (Chalkidis et al., 2021), and finance (Wang et al., 2023). The urgency of this task is underscored by emerging regulatory frameworks like the EU AI Act (COM, 2021) and the NIST AI RMF (AI, 2023), which mandate rigorous documentation and monitoring (OECD, 2019). However, even as recent works advance auditing for semi-structured reasoning (Leng et al., 2025) or propose new faithfulness metrics (Lanham et al., 2023), prevailing auditing methods remain misaligned with this paradigm. Their centralized, opaque, and unscalable nature creates unacceptable risks, as they either rely on a single trusted entity, cannot process the volume and complexity of reasoning traces, or force a dangerous trade-off between public transparency and the protection of proprietary models.

The inadequacy of current auditing systems stems from four interconnected challenges. A primary issue is a lack of **Robustness** since systems relying on a single auditor, whether a human expert or another LLM, constitute a “single point of failure” and are vulnerable to targeted attacks like prompt

054 injection (Zou et al., 2023; Perez and Ribeiro, 2022) and susceptible to systemic biases (Bender et al.,  
 055 2021; Liang et al., 2022). Compounding this issue is a severe **Scalability** bottleneck. The volume  
 056 and combinatorial complexity of reasoning traces from modern LRM, especially those employing  
 057 branching search (Lightman et al., 2023; Yao et al., 2023a), make comprehensive manual verification  
 058 practically and economically infeasible, a fact evidenced by the massive human effort required for  
 059 existing process supervision datasets (Bai et al., 2022; Lightman et al., 2023). Furthermore, the  
 060 **Opacity** of internal auditing processes at proprietary model providers erodes public trust and prevents  
 061 independent verification of safety claims, conflicting with established principles of transparent  
 062 reporting (Bommasani et al., 2023; Mitchell et al., 2019). In parallel, addressing opacity creates a  
 063 critical tension with **Privacy**, since exposing complete reasoning traces for public audit risks the theft  
 064 of valuable intellectual property through model distillation (Carlini et al., 2021) and increases the  
 065 surface area for extracting sensitive training data (Nasr et al., 2023).

066 Addressing these simultaneous challenges of robustness, scalability, opacity, and privacy demands a  
 067 new approach to the auditing paradigm. Our work is guided by the following research questions:

068 **RQ 1.** *How can we design an auditing system that is robust to malicious participants and systemic  
 069 bias without relying on a central trusted authority*

070 **RQ 2.** *How can this system scale to audit complex reasoning traces while preserving the intellectual  
 071 property of the model provider and ensuring public transparency*

072 Answering these questions naturally leads to a framework that integrates decentralized consensus,  
 073 privacy-preserving protocols, and a novel representation for reasoning itself.

074 We introduce TRUST, a decentralized framework for  
 075 auditing LLM reasoning (see Figure 1). To achieve  
 076 **robustness**, TRUST establishes a consensus mech-  
 077 anism among a diverse, multi-tier set of auditors,  
 078 drawing on principles from Byzantine Fault Tol-  
 079 erant systems (Castro et al., 1999; Lamport et al.,  
 080 2019) to provably guarantee audit correctness even  
 081 with a significant fraction of malicious participants.  
 082 For **scalability**, the framework introduces a novel  
 083 decomposition method that transforms reasoning  
 084 traces into *Hierarchical Directed Acyclic Graphs*  
 085 (*HDAGs*), a structured representation that permits  
 086 parallel verification of atomic reasoning steps by a  
 087 distributed network. To jointly address **opacity** and  
 088 **privacy**, all verification decisions are recorded on a  
 089 transparent blockchain ledger for public auditability,  
 090 while the protocol preserves confidentiality by  
 091 distributing only partial, disconnected trace segments to  
 092 logic from reconstruction.

093 The design of TRUST is supported by rigorous theoretical guarantees for security and economic  
 094 viability. We prove a *Safety-Profitability Guarantee* theorem, which formally ensures that under  
 095 configurable statistical and economic conditions, the system achieves a target audit safety level  
 096 while making honest participation profitable and malicious behavior result in a net loss. Our em-  
 097 pirical validation spans multiple state-of-the-art LLMs (e.g., GPT-OSS (OpenAI, 2025), DeepSeek-  
 098 r1 (DeepSeek-AI et al., 2025), Qwen (Yang et al., 2025)) and diverse reasoning tasks, and incorporates  
 099 human-in-the-loop experiments with expert auditors to validate the multi-tier design. The results  
 100 demonstrate that TRUST is highly effective at identifying reasoning flaws and is significantly more  
 101 resilient to coordinated attacks than centralized baselines.

102 In summary, our main contributions are:

- 103 • We introduce TRUST, the first decentralized auditing system for reasoning traces that achieve  
 104 privacy-preserving verification without exposing the proprietary model.
- 105 • We develop a systematic approach to decompose Chain-of-Thought reasoning into Hierarchy  
 106 Directed Acyclic Graphs (HDAGs) that enable modular verification coupled with a multi-tier  
 107 verification, routing simple problems to automated validators and complex problems to human  
 108 experts.

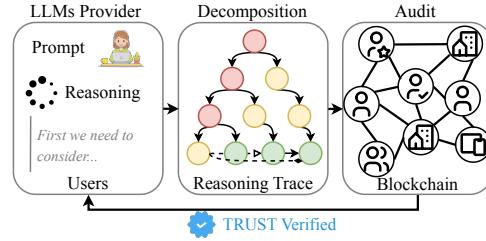


Figure 1: Reasoning traces (left) are de-  
 composed into hierarchical segments (middle) and  
 verified by a heterogeneous, multi-tier auditor  
 network (right). Audit outcomes are recorded  
 on-chain, while raw trace is stored off-chain  
 for privacy.

while the protocol preserves confidentiality by  
 distributing only partial, disconnected trace segments to  
 logic from reconstruction.

- 108 • We develop theoretically grounded incentive mechanisms which ensure that honest auditors profit  
109 while malicious actor incur losses, providing the foundation necessary for sustainable real-world  
110 deployment at scale.
- 111 • We conduct comprehensive experiments on diverse datasets (e.g., MMLU-Pro, GSM8K) and  
112 models (e.g., GPT-OSS, DeepSeek-r1), including human-in-the-loop studies, to demonstrate the  
113 effectiveness and robustness of TRUST against centralized baselines.

## 115 2 RELATED WORKS

116 **Reasoning Model Verification.** Chain-of-Thought (CoT) prompting has revolutionized LLM reasoning by exposing intermediate steps (Wei et al., 2022), evolving into sophisticated tree-based  
117 search methods (Yao et al., 2023a) and Large Reasoning Models that treat reasoning as a primary  
118 objective (Jaech et al., 2024; Guo et al., 2025). However, these advances lack systematic verification  
119 mechanisms for generated reasoning traces, particularly for privacy-preserving and decentralized  
120 auditing.

121 **Auditing and Evaluation.** Current auditing approaches range from centralized “LLM-as-a-judge”  
122 methods (Zheng et al., 2023) to Process Reward Models that provide step-by-step supervision (Light-  
123 man et al., 2023). This line of work has been further refined by methods focusing on specific aspects of  
124 verification. For instance, Leng et al. (2025) proposes a rigorous auditing method for semi-structured  
125 reasoners, focusing on formal verification within a structured environment, while Lanham et al. (2023)  
126 concentrates on metrics to measure the faithfulness of chain-of-thought reasoning. While valuable,  
127 these approaches typically presume a centralized verifier and do not address scalable auditing via  
128 decentralized consensus. Recent work also addresses service-level integrity through cryptographic  
129 verification (Sun et al., 2025) and detection of model substitution (Cai et al., 2025). While addressing  
130 inference integrity, these approaches lack unified frameworks for scalable semantic auditing with  
131 decentralized consensus.

132 **Decentralized Verification.** Foundational work in Byzantine Fault Tolerant consensus (Castro et al.,  
133 1999) and Zero-Knowledge Proofs for ML (Chen et al., 2024; Sun et al., 2024) provides primitives  
134 for verifiable computation. The emerging field of Zero-Knowledge Machine Learning (ZKML)  
135 specifically aims to ensure the verifiability of ML models without disclosing sensitive data (Peng  
136 et al., 2025). However, existing approaches focus on computational correctness rather than semantic  
137 quality verification through human-in-the-loop consensus processes.

138 Our work synthesizes these directions by introducing the first framework for decentralized, privacy-  
139 preserving semantic auditing of reasoning traces at scale. Due to space limitation, we provide more  
140 comprehensive related works in [Section B](#).

## 142 3 DECENTRALIZED AUDITING FOR LARGE REASONING MODELS

144 As illustrated in [Figure 1](#), TRUST can integrate either human or LLMs to audit *faithfulness*, *harm-  
145 lessness*, and *logical consistency* of chain-of-thought (CoT) reasoning. By operating on intermediate  
146 traces rather than final outputs alone, TRUST enables earlier and more comprehensive detection of  
147 reasoning flaws. TRUST features the following key innovations:

- 148 • **Batch & Segmentation.** Reasoning traces from multiple providers are *batched* to anonymize  
149 source identity and mitigate provider-specific bias. Traces are then *segmented* into minimal,  
150 auditable units and stored as content-addressed objects in decentralized storage. Segmentation  
151 protects proprietary logic: each auditor only sees the segment(s) they are assigned, preventing  
152 full-trace reconstruction.
- 153 • **Auditing & Consensus.** Heterogeneous auditors (computational checkers, LLMs, and humans)  
154 independently evaluate assigned segments. Votes are submitted via a cryptographic *commit-reveal*  
155 protocol: in the commit phase, auditors submit hashed votes; in the reveal phase, they disclose votes  
156 for verification against commitments. Segment-level quorums validate local steps; a trace-level  
157 aggregator combines weighted segment outcomes to reach the final decision.
- 158 • **Blockchain & Decentralized Storage.** A blockchain layer provides immutable audit trails and  
159 trustless consensus using a Proof-of-Stake (PoS)-style mechanism adapted for AI auditing. Smart  
160 contracts orchestrate session lifecycle, auditor assignment (by stake and expertise), commit-reveal  
161 voting, and performance-based rewards/slashing. Reasoning content is stored off-chain on IPFS;  
the blockchain records metadata, vote commitments, and final outcomes.

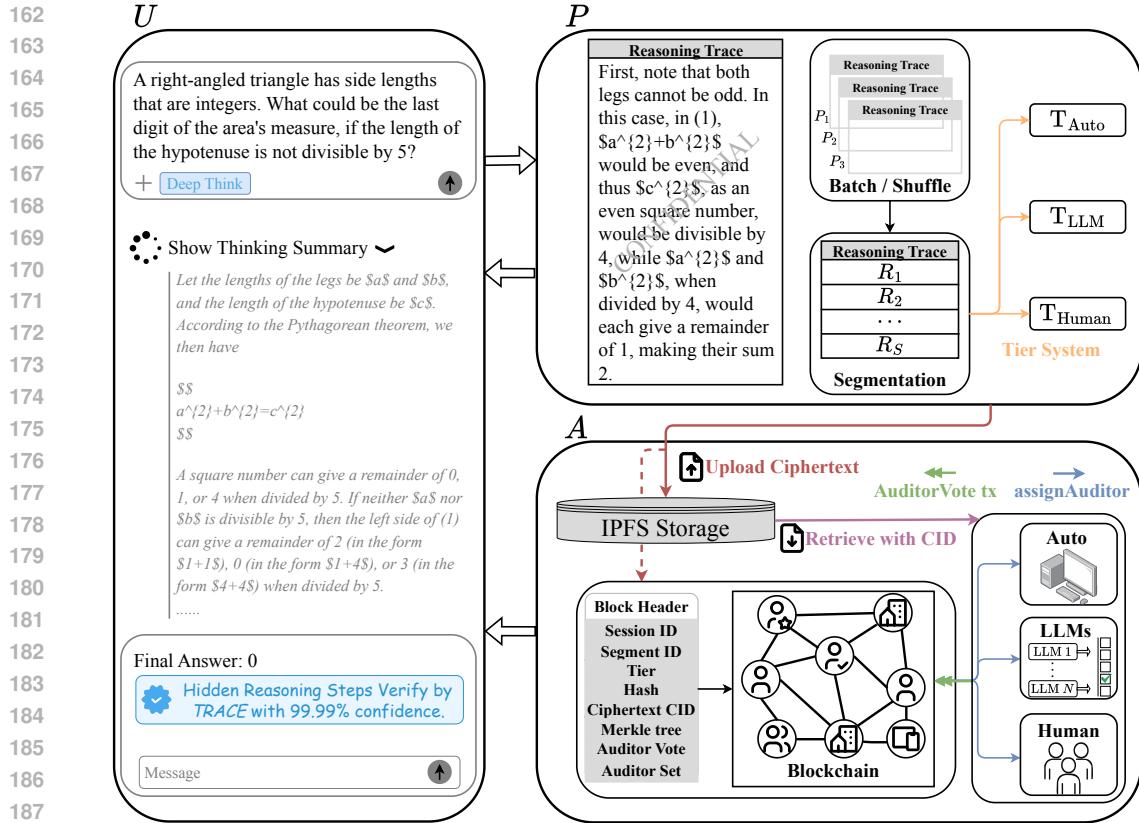


Figure 2: Overview of TRUST. The TRUST framework decomposes Chain-of-Thought reasoning traces into hierarchical directed acyclic graphs (DAGs) for modular verification across a three-tier auditor system. The process begins with a reasoning query (left panel) that generates intermediate reasoning steps, which are then decomposed into graph components and distributed across automated computers, LLM-based, and human auditors. TRUST utilizes IPFS for decentralized storage of reasoning traces and blockchain technology for immutable audit records, vote aggregation, and consensus mechanisms. Auditors verify reasoning segments independently, while cryptographic protocols ensure the privacy preservation of proprietary model internals. The final verification result provides confidence guarantees for reasoning trace faithfulness and correctness for the end user.

We formalize the three key participant parties in the TRUST ecosystem—**Provider**, **Auditor**, and **User**. TRUST supports both Business-to-Business (B2B) and Business-to-Consumer (B2C) settings, where the reasoning trace provider may be either a proprietary LLM vendor or an individual customer. We provide the illustration of the parties involved in TRUST in Figure 2.

- **Provider (P)**: A proprietary LRM vendor or customer submitting reasoning traces for audit.
- **Auditor (A)**: An anonymous seat (computer, LLM, or human) that verifies assigned segments.
- **User (U)**: An end-user who consumes audited outputs and provenance via APIs or dashboards.

The practical deployment of TRUST hinges on its economic viability. Our framework is designed to be agnostic to the underlying blockchain technology, enabling integration with emerging high-throughput, low-cost Layer-2 solutions to mitigate transaction latency and costs (i.e., gas fees). In our model, these transaction costs are considered an operational expense for the Provider requesting the audit, ensuring the system remains sustainable and economically practical for the auditors who form the backbone of the network.

Given a reasoning trace with  $S$  segments (including CoT and tool calls), TRUST maps the trace to a *Hierarchical Directed Acyclic Graph (HDAG)* with five abstraction levels: *Goal*, *Strategy*, *Tactic*, *Step*, and *Operation*. This representation is *problem-agnostic* (math, science, programming, general reasoning, etc.) and enables scalable, parallel verification because most nodes are independently auditable. Each node carries metadata (ID, summary, complexity, auditor type, and dependencies), and edges encode relationships (*decomposes\_to*, *depends\_on*, *enables*, *validates*, *contradicts*, etc.). Formally,

216 each segment  $s \in \{1, \dots, S\}$  is assigned to a primary auditor type  $\in \{\text{Computer}, \text{LLM}, \text{Human}\}$ :

$$\underbrace{\text{Segment 1},}_{\text{Computer}} \underbrace{\text{Segment 2},}_{\text{LLM}} \underbrace{\text{Segment 3}, \dots,}_{\text{Human}} \underbrace{\text{Segment } S}_{\text{type} \in \{\text{C}, \text{L}, \text{H}\}}$$

217 **Hierarchical Directed Acyclic Graphs (HDAGs).** Prior work on CoT decomposition, such as  
 218 DLCoT, introduced automatic frameworks for breaking down long reasoning traces into structured  
 219 segments, primarily to generate high-quality data for model distillation (Luo et al., 2025). These  
 220 works observe that CoTs can follow linear, tree, or more general network structures. DLCoT, for  
 221 instance, applies macro-structure parsing to divide CoTs into four parts—*Problem Restatement*,  
 222 *Approach Exploration*, *Verification*, and *Summary*—before further segmenting the approach and  
 223 verification stages into stepwise units. Other lines of research (Kothapalli et al., 2025) focus on  
 224 extracting causal structures from token-level processing functions.

225 In contrast, we propose a general, problem-agnostic approach: decomposing CoTs into *Hierarchical*  
 226 *Directed Acyclic Graphs* (HDAGs). Our hierarchy consists of five abstraction levels: *Goal*, *Strategy*,  
 227 *Tactic*, *Step*, and *Operation*. This abstraction provides two key advantages. First, it is broadly  
 228 applicable across domains—mathematics, science, engineering, and open-domain reasoning. Second,  
 229 it enables scalable verification, since most nodes are independently auditable and can be naturally  
 230 mapped to different auditor types (e.g., computer programs, LLMs, or human experts). An illustration  
 231 is provided in Figure 3, where each node is annotated by difficulty and type (basic reasoning step,  
 232 tool usage, or fact/premise). This hierarchical decomposition mirrors neural circuits in the frontal  
 233 cortex, which process reasoning through multi-level evidence integration (Sarafyazd and Jazayeri,  
 234 2019). Just as the brain organizes reasoning hierarchically rather than linearly, our HDAG design  
 235 enables different reasoning components to be audited at the appropriate granularity. Edges capture  
 236 logical relationships between nodes, including *dependencies* (depends on, enables), *structural links*  
 237 (decomposes, refines), *validation* (validates, exemplifies), and *conflicts* (contradicts).

238 Concretely, TRUST constructs HDAGs in five steps:

239 **Step 1: Identify Abstraction Levels.** The raw problem statement,  
 240 reasoning trace (with tool usage), and final output are parsed into semantic hierarchy levels.

241 **Step 2: Segment Within Each Level.** Each level is further  
 242 divided into granular units with associated metadata  
 243 (IDs, complexity, summaries). Difficulty annotations  
 244 guide later auditor assignment.

245 **Step 3: Extract Relationships.** Logical dependencies  
 246 between segments are mapped into relations (*decomposes\_to*, *depends\_on*, *enables*, *validates*, etc.).

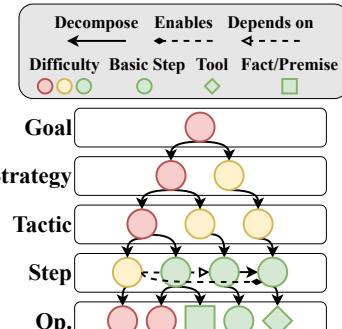
247 **Step 4: Assign Auditor Types.** Segments are routed to auditor  
 248 types from  $\{\text{Human}, \text{Computer}, \text{LLM}\}$ , based on  
 249 complexity and modality.

250 **Step 5: Refine and Construct HDAG.** Segments and  
 251 relationships are synthesized into a final auditable HDAG,  
 252 with quality assurance checks.

253 Further implementation details and examples are provided in  
 254 Sections C.1, C.3 and C.4.

255 **Auditing & Consensus.** Reasoning traces vary in both complexity and type of reasoning step. In  
 256 TRUST, each segment is routed to the most appropriate auditor type—*Human*, *LLM*, or *Computer*—to  
 257 ensure accurate and efficient verification. Decentralized participants either contribute computational  
 258 resources for deterministic checks (e.g., arithmetic, tool calls) or provide expertise in evaluating  
 259 semantic coherence and faithfulness. This heterogeneous auditor pool improves both accuracy (by  
 260 matching segments to suitable verifiers) and robustness (by reducing vulnerability to malicious or  
 261 biased auditors). We analyze the consensus process at three levels: *seat*, *segment*, and *trace*.

262 **1. Seat layer.** Within a segment  $s$ , each of the  $k_{t(s)}$  auditor seats votes independently. Computer  
 263 seats are assumed noiseless, while LLM and human seats have nonzero error rates  $\epsilon_t$ . Human  
 264 seats may additionally be adversarial with probability  $\rho_H$ .



265 Figure 3: Example of an HDAG. Node color indicates the level of  
 266 difficulty assigned to different types of auditors. Node shape denotes the type of step and edges represent  
 267 relationships between nodes.

2. **Segment layer.** For the segment  $s$ , define the segment pass indicator  $B_s = \mathbf{1}[\#\{\text{correct votes}\} \geq q_{t(s)}]$ , where  $q_t = \lceil \tau k_t \rceil$  is the quorum threshold for type  $t$ . The exact pass probability for a segment of type  $t$  with parameters  $(k_t, \epsilon_t, \rho_t)$  (with  $\rho_C = \rho_L = 0$ ) is

$$p_t = \Pr[B_s = 1] = \sum_{m=0}^{k_t} \binom{k_t}{m} \rho_t^m (1 - \rho_t)^{k_t - m} \sum_{c=q_t}^{k_t - m} \binom{k_t - m}{c} (1 - \epsilon_t)^c \epsilon_t^{k_t - m - c}, \quad (3.1)$$

where  $m$  malicious seats vote incorrectly, and among the  $k_t - m$  honest seats,  $c$  cast correct votes.

3. **Trace layer.** To aggregate across all  $S$  segments, we assign weights  $w_{t(s)}$  and define  $W = \sum_{s=1}^S w_{t(s)} B_s$ ,  $W_\beta = \beta \sum_s w_{t(s)}$ . We then bound the failure probability  $\Pr[W < W_\beta]$  using Hoeffding and Chernoff inequalities:

$$\Pr[W < W_\beta] \leq \underbrace{\exp\left[-2(\mu_{\text{vote}} - W_\beta)^2/\sigma_{\max}^2\right]}_{\text{Hoeffding}} \wedge \underbrace{\min_{\lambda > 0} \exp\left(\lambda W_\beta + \sum_{s=1}^S \ln(p_s e^{-\lambda w_s} + (1 - p_s))\right)}_{\text{Chernoff}}. \quad (3.2)$$

**Figure 4** compares these bounds with the exact solution under representative parameters ( $\epsilon_C = 0$ ,  $\epsilon_L = 0.05$ ,  $\epsilon_H = 0.30$ ,  $\rho_H = 0.1$ ). The full derivation on seat, segment, and trace levels results are provided in [Section D](#).

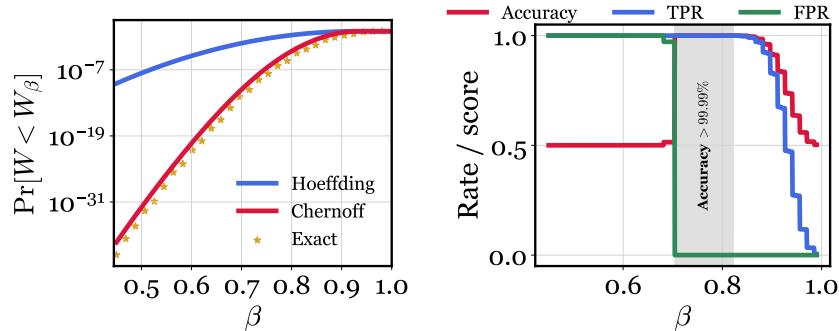


Figure 4: The parameters are  $\epsilon_C = 0$ ,  $\epsilon_L = 0.05$ ,  $\epsilon_H = 0.30$ , and  $\rho_H = 0.1$ . (Left) Comparison of probability of failure of Hoeffding and Chernoff bounds and exact solution in (3.2) as a function of trace-level quorum threshold  $\beta$ . (Right) The true positive rate (TPR), false positive rate (FPR), and accuracy with different values of trace-level quorum threshold. The grey shaded area indicates the width of the trace-level quorum that achieves greater accuracy than 99.99%.

**Privacy by Design.** The privacy-preserving nature of TRUST is an intrinsic property of its structural design. By decomposing the reasoning trace into an HDAG, we compartmentalize the verification process. Each auditor is assigned only one or more atomic segments of the trace, without access to the complete context or the final conclusion. This “need-to-know” basis ensures that the full proprietary logic of the reasoning process is never exposed to any single party, thus preventing intellectual property theft or model distillation. The on-chain records are limited to cryptographic commitments of these segments and their verification outcomes, serving as immutable proof of work while keeping the reasoning content itself off-chain on the InterPlanetary File System (IPFS) and private.

### 3.1 ECONOMIC ANALYSIS

In this section, we provide the economic analysis of the TRUST framework on reputation, slashing, reward, statistical, and economic guarantees.

**Reputation-Weighted Slashing and Rewards.** Each human auditor seat  $i$  maintains a reputation score  $r_i(t) \in [0, 1]$ , updated after every segment as  $r_i(t+1) = (1 - \gamma) r_i(t) + \gamma \mathbf{1}[\text{vote correct}]$ , where  $\gamma \in (0, 1]$  controls adaptation speed. Incorrect votes trigger a slashing probability  $p_{\text{slash}}(r) = p_{\min} + (p_{\max} - p_{\min})(1 - r)$ , with  $0 < p_{\min} < p_{\max} \leq 1$ , penalizing low-reputation seats more heavily. The per-segment payoff  $X_i \in \{-P, 0, R\}$  is defined as:  $R$  for a correct vote,  $0$  for an incorrect vote without slashing, and  $-P$  for a slashed incorrect vote. For an honest seat with error rate  $\epsilon_H$ , the expected payoff is  $\mu_H(r) := \mathbb{E}[X_i] = (1 - \epsilon_H)R - \epsilon_H P p_{\text{slash}}(r)$ .

Take parameters  $R = 6$ ,  $P = 8$ ,  $p_{\min} = 0.2$ ,  $p_{\max} = 0.5$ ,  $\delta = 0.2$ ,  $\lambda = 60$ , and  $\epsilon_H = 0.30$  for example, an honest seat achieves an expected per-segment payoff of  $\mu_{\min} = 0.7 \times 6 - 0.3 \times 8 \times 0.5 = 3.0$ , with variance  $\sigma_H^2 = 25.8$  and worst-case increment  $b = 6$ . A malicious seat, by contrast, suffers an expected loss of  $\mathbb{E}[X_{\text{mal}}] = -0.5 \times 8 = -4.0$ , with variance 16 and worst-case increment  $b = 8$ . Over a 24-hour window ( $T = 24$ ) with 1440 segments, tail bounds from [Theorem D.1](#) show that the probability of an honest auditor ending with nonpositive payoff is at most  $\exp(-(60 \times 24 \times 3^2) / (2 \times 25.8 + (2/3) \times 6 \times 3)) \approx e^{-204} < 10^{-80}$ , while the probability of a malicious auditor breaking even or better is at most  $\exp(-(60 \times 24 \times (0.2 \cdot 8)^2) / (2 \times 25.8 + (2/3) \times 6 \times 1.6)) \approx e^{-63.6} < 10^{-27}$ .

**Tokenomics.** The tokenomics of TRUST (Illustrated in [Section 3.1](#)) ensures incentive-compatible decentralized auditing. Provider fees are split into a *Protocol Treasury* (20%) for long-term sustainability and an *Audit Pool* (80%) for immediate auditor rewards. Auditors and delegators stake tokens to participate; honest auditors aligning with consensus receive rewards, while dishonest ones are slashed, creating strong deterrents against manipulation.

Specifically, Delegators act as capital providers within the ecosystem, staking their assets with auditors who have a strong track record of reliable verification. In return, they earn a percentage of the auditor's rewards. This symbiotic relationship allows skilled auditors to increase their stake-backed influence and auditing capacity, while enabling token holders to productively deploy their capital, thereby enhancing the overall security and robustness of the network. The system's resilience against collusion, including potential dishonest collaboration between delegators and auditors, is maintained through our decentralized consensus protocol and dynamic trust score. Any deviation from the consensus outcome results in financial penalties (slashing) for the auditor, and consequently, their delegators, thus creating a strong economic disincentive for such behaviors.

## 4 EXPERIMENTAL VERIFICATION

In this section, we provide verification on TRUST on the annotated CoT dataset, open-source model-generated CoTs for de-bias, and safety and privacy results.

### 4.1 CORRECTNESS AND FAITHFULNESS.

We use 200 samples from the MMLU-Pro-CoT-Train dataset ([Lab, 2024](#)), which provides ground truth annotations for individual reasoning steps and final answers. This allows us to systematically evaluate the correctness and faithfulness of audits at both the step and trace levels. We compare TRUST against centralized approaches, including (i) single-LLM auditors (DeepSeek-R1-8B, Qwen2.5-7B, Mistral-7B, GPT-OSS-20B, LLaMA-3B) and (ii) ensemble-based voting schemes (majority, supermajority, weighted, unanimous). To stress-test robustness, we simulate auditor corruption by systematically flipping a proportion of segment-level votes, with corruption rates ranging from 5% to 20%.

[Figure 6](#) and [Table 1](#) summarize the results. At baseline (no corruption), TRUST achieves the highest accuracy (72.4%), outperforming both single auditors (e.g., DeepSeek-R1-8B at 67.7%) and ensemble methods (e.g., majority voting at 68.7%). As corruption increases, all methods degrade, but TRUST degrades more gracefully: accuracy remains above 63% even at 20% corruption, while centralized ensembles drop below 61% and single auditors fall closer to 60%. The performance gap widens with higher corruption rates, highlighting TRUST's resilience to adversarial or biased auditors.

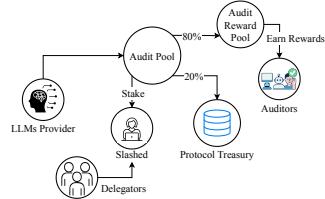


Figure 5: Tokenomics of TRUST.

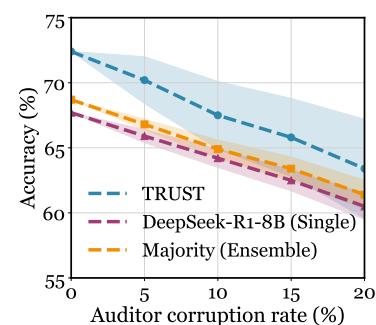


Figure 6: Correctness of Single, Ensemble (Centralized) with decentralized TRUST framework.

Table 1: Performance comparison of TRUST (decentralized) vs. centralized approaches across corruption rates. Best is **bold**, second-best is underlined.

| Category               | Method               | Baseline                         | 5% Corr.                         | 10% Corr.                        | 15% Corr.                        | 20% Corr.                        |
|------------------------|----------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| <b>Decentralized</b>   | TRUST                | <b><math>72.4 \pm 0.0</math></b> | <b><math>70.2 \pm 1.8</math></b> | <b><math>67.5 \pm 2.6</math></b> | <b><math>65.8 \pm 3.0</math></b> | <b><math>63.4 \pm 3.8</math></b> |
| <i>Ensemble Models</i> |                      |                                  |                                  |                                  |                                  |                                  |
|                        | Majority Voting      | $68.7 \pm 0.0$                   | $66.8 \pm 0.4$                   | $64.9 \pm 0.7$                   | $63.4 \pm 0.9$                   | $61.4 \pm 1.1$                   |
|                        | Supermajority        | $68.7 \pm 0.0$                   | $66.8 \pm 0.5$                   | $65.0 \pm 0.7$                   | $63.2 \pm 0.9$                   | $61.2 \pm 0.9$                   |
|                        | Weighted Voting      | $68.1 \pm 0.0$                   | $66.4 \pm 0.6$                   | $64.5 \pm 0.7$                   | $62.7 \pm 1.1$                   | $60.9 \pm 0.9$                   |
|                        | Unanimous            | $45.6 \pm 0.0$                   | $46.1 \pm 0.6$                   | $46.5 \pm 0.9$                   | $46.8 \pm 1.1$                   | $47.4 \pm 1.0$                   |
| <b>Centralized</b>     | <i>Single Models</i> |                                  |                                  |                                  |                                  |                                  |
|                        | DeepSeek-R1-8B       | $67.7 \pm 0.0$                   | $65.9 \pm 0.5$                   | $64.2 \pm 0.7$                   | $62.5 \pm 0.8$                   | $60.5 \pm 1.0$                   |
|                        | Qwen2.5-7B           | $67.4 \pm 0.0$                   | $65.7 \pm 0.6$                   | $64.1 \pm 0.7$                   | $62.1 \pm 0.9$                   | $60.5 \pm 1.0$                   |
|                        | Mistral-7B           | $66.8 \pm 0.0$                   | $65.2 \pm 0.6$                   | $63.6 \pm 0.8$                   | $61.8 \pm 1.1$                   | $60.1 \pm 1.1$                   |
|                        | DeepSeek-R1-1.5B     | $64.1 \pm 0.0$                   | $62.9 \pm 0.5$                   | $61.2 \pm 0.8$                   | $59.7 \pm 1.1$                   | $58.5 \pm 0.8$                   |
|                        | GPT-OSS-20B          | $63.8 \pm 0.0$                   | $62.5 \pm 0.6$                   | $60.9 \pm 0.7$                   | $59.7 \pm 1.1$                   | $58.4 \pm 1.0$                   |
|                        | LLaMA-3B             | $52.1 \pm 0.0$                   | $51.9 \pm 0.6$                   | $51.7 \pm 0.7$                   | $51.4 \pm 1.0$                   | $51.3 \pm 1.1$                   |

**Output-Based Auditing is Insufficient.** As shown in Table 2, output-only evaluation drops accuracy by 18%–47% across all models compared to full-CoT auditing. This confirms that correct final answers often mask flawed reasoning that only semantic auditing can detect.

Table 2: Baseline accuracy under Full-CoT vs. Output-Based evaluation.

|              | TRUST       | R1-8B | Qwen-7B | Mistral-7B | R1-1.5B | OSS-20B | LLaMA-3B |
|--------------|-------------|-------|---------|------------|---------|---------|----------|
| Full-CoT     | <b>72.4</b> | 67.7  | 67.4    | 66.8       | 64.1    | 63.8    | 52.1     |
| Output-Based | —           | 36.0  | 34.0    | 20.0       | 22.0    | 46.0    | 34.0     |

#### 4.2 SAFETY AND PROFITABILITY

A central design goal of TRUST is to guarantee both *statistical safety*—ensuring that the probability of a failed audit remains vanishingly small—and *economic sustainability*—ensuring that honest auditors are consistently rewarded while malicious ones suffer provable long-term losses. Figure 7 illustrates these dynamics empirically. On the left of Figure 7, reputation scores naturally separate: honest auditors are reinforced with high reputation, while malicious and random guessers quickly lose credibility. On the right, profit trajectories diverge: honest participants earn steadily increasing rewards, while guessers and malicious seats accumulate losses due to repeated slashing. These empirical trends are formally supported by the *Safety–Profitability Guarantee* (Theorem D.1), which proves that, under appropriate statistical and economic parameters, honest auditors almost surely remain profitable while malicious participants incur provable long-term losses. The detailed derivation of these guarantees is provided in Section D.

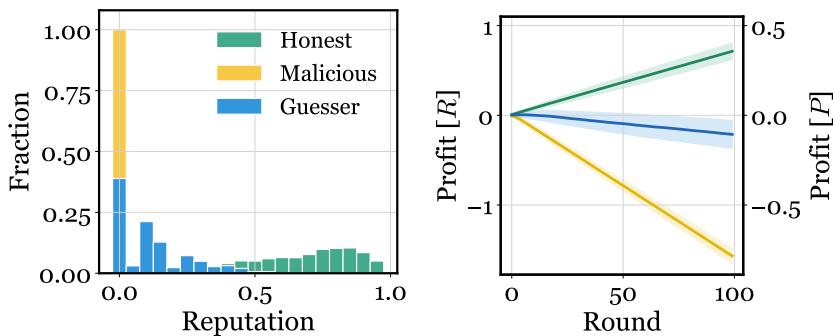


Figure 7: The parameters  $\epsilon_C = 0$ ,  $\epsilon_L = 0.05$ ,  $\epsilon_H = 0.30$ , and  $\rho_H = 0.1$ . (Left) Reputation scores. (Right) Profit curves.

**Theorem 4.1** (Safety–Profitability (Informal version of Theorem D.1)). Fix horizon  $T$ , target failure  $\epsilon_{\text{target}}$ , and  $\delta \in (0, 1)$ .

- **Statistical dial:** ensure  $\mu_{\text{vote}} - W_\beta \geq \sqrt{\frac{1}{2} \sigma_{\text{vote}}^2 \ln \frac{\lambda T}{\epsilon_{\text{target}}}}$ .

432 • **Economic dial:** set  $(R, P, p_{\min}, p_{\max})$  with  $\mu_{\min} := (1 - \epsilon_H)R - \epsilon_H P p_{\max} > 0$ ,  $p_{\min} \geq \frac{\delta}{1-\alpha}$ .

433 Then:

434 (a) **Safety:**  $\Pr[\text{fail in } [0, T]] \leq \epsilon_{\text{target}}$ .

435 (b) **Honest profit:**  $\Pr[U_{\text{hon}}(T) \leq 0] \leq \exp(-c_1 T \mu_{\min}^2)$ .

436 (c) **Malicious loss:**  $\Pr[U_{\text{mal}}(T) \geq 0] \leq \exp(-c_2 T (\delta P)^2)$ ,  $\mathbb{E}[U_{\text{mal}}(T)] \leq -\lambda T \delta P$ .

### 440 441 4.3 BIAS MITIGATION.

442 Auditing systems are vulnerable to bias, 443 where auditors may favor reasoning traces 444 produced by their own model family or 445 penalize outputs from competing models. 446 This creates two common failure modes: 447 (i) *self-favoritism*, where a model system- 448 atically approves its own reasoning, and 449 (ii) *self-criticism*, where a model dispropor- 450 tionately rejects its own outputs.

451 TRUST is designed to mitigate such 452 bias through three architectural features: 453 (1) *Segment-level decomposition*, which 454 breaks reasoning traces into atomic units; 455 (2) *Multi-tier consensus*, combining hu- 456 man, LLM, and automated auditors; and (3) 457 *Anonymous evaluation*, hiding the source 458 model of each segment.

459 We construct a benchmark of 200 questions 460 across four domains. For each, CoT traces 461 from DeepSeek-R1-1.5B and GPT-OSS-20B 462 are evaluated under three regimes: single auditors, 463 ensemble auditors, and TRUST. In Table 3, single auditors vary in accuracy (15.2–60.9%) but show 464 bias (avg. +5.5%). Ensembles remove bias but perform poorly (16.9–30.5%). TRUST breaks this 465 tradeoff, achieving higher accuracy (34.1%) without bias.

### 466 467 4.4 HUMAN EXPERIMENT ON THREE-TIER AUDITING WITH TRUST VARIANTS

468 To evaluate TRUST in a multi-tier auditing setting, we recruited 30 computer science students 469 as human auditors (Example provided in Section C.2). We constructed a benchmark of 30 math 470 problems from the openai/gsm8k dataset (Cobbe et al., 2021b), generating CoT traces with 471 DeepSeek-R1-8B and GPT-OSS-20B. Auditors participated within the TRUST framework alongside 472 LLM and automated auditors, and we compared performance against both single-LLM auditors and a 473 centralized human-only audit.

474 Table 4 reports F1 and Brier scores. Single LLM auditors perform poorly, reflecting both limited 475 correctness when providing a whole trace to audit. Centralized human auditors achieve moderately 476 stronger results (F1 = 0.85, Brier = 0.21) and require approximately 1.4 minutes per trace on average 477 (medium < 1 minute). While they clearly outperform single LLMs auditors, they remain time- 478 consuming and are susceptible to systematic biases and malicious behavior. In contrast, TRUST 479 achieves a much higher F1 of 0.89 and a substantially lower Brier score of 0.074, demonstrating 480 both accurate and well-calibrated auditing. These results highlight the effectiveness of combining 481 human, LLM, and automated auditors under TRUST, outperforming both centralized human audits 482 and single-model baselines. For TRUST variants, random segment break CoT randomly between 483 2–6 sentences; fixed-length short segment breaks into two sentences per segment; fixed-length long 484 segment breaks into 5–7 sentences per segment. All three variants are using GPT-OSS-20B.

### 485 486 4.5 LATENCY

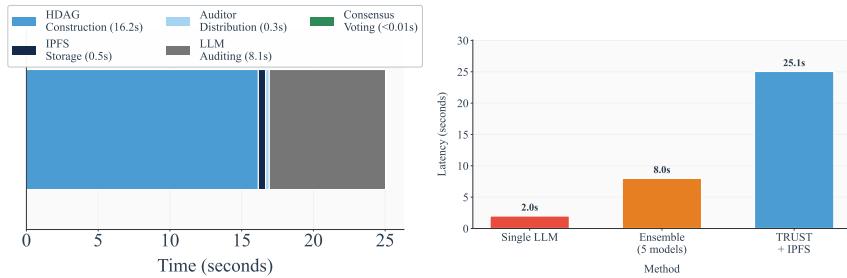
487 In Figure 8, we present a latency analysis, decomposing TRUST’s overhead into constituent compo- 488 nents and comparing it against centralized baselines.

Table 3: Comparison of auditing methods on reasoning trace verification. Accuracy in %. Bias score = (self-approval)–(other-approval). Positive bias = favoritism, Negative bias = criticism.

| Method                       | Acc. (%)    | Bias  |
|------------------------------|-------------|-------|
| <b>TRUST (Decentralized)</b> | <b>34.1</b> | –     |
| <b>Ensemble</b>              |             |       |
| Supermajority                | 30.5        | –     |
| Majority                     | 26.8        | –     |
| Weighted                     | 16.9        | –     |
| <b>Single LLMs</b>           |             |       |
| GPT-OSS-20B                  | 60.9        | +44.4 |
| Qwen-7B                      | 38.4        | –     |
| DeepSeek R1-1.5B             | 21.9        | -11.3 |
| Llama-3B                     | 16.2        | –     |
| Mistral-7B                   | 15.9        | –     |
| DeepSeek R1-8B               | 15.2        | –     |
| <b>Average</b>               | 28.1        | +5.5  |

486  
 487  
 488  
 489  
 490  
 491  
 492  
 493  
**Table 4: Human Experiment on Three-Tier Auditing: Performance comparison across single LLM**  
 auditors, centralized human auditing, and TRUST variants. 30 Computer Science students audited  
 30 GSM8K math problems with CoT traces from DeepSeek-R1-8B and GPT-OSS-20B. Metrics are  
 F1 score (higher better) and Brier score (lower better). Single LLMs perform poorly, centralized  
 human audit achieves **F1=0.85**, while TRUST w/ HDAG substantially outperforms all baselines  
 (F1=0.89, Brier=0.074) through multi-tier consensus. TRUST variants using random or fixed-length  
 segmentation show degraded performance (F1=0.40), validating the importance of hierarchical  
 decomposition.

| Metric      | Single LLM Auditors |            |            |                  |             |          | TRUST Variants |               |                     |                          |                         |
|-------------|---------------------|------------|------------|------------------|-------------|----------|----------------|---------------|---------------------|--------------------------|-------------------------|
|             | DeepSeek R1-8B      | Qwen2.5 7B | Mistral 7B | DeepSeek R1-1.5B | GPT-OSS 20B | LLaMA 3B | Human Audit    | TRUST w/ HDAG | TRUST w/ Random Seg | TRUST w/ Fix Seg (Short) | TRUST w/ Fix Seg (Long) |
| F1          | 0.50                | 0.50       | 0.50       | 0.40             | 0.50        | 0.30     | <b>0.85</b>    | <b>0.89</b>   | 0.40                | 0.40                     | 0.40                    |
| Brier Score | 0.500               | 0.486      | 0.500      | 0.500            | 0.544       | 0.890    | <b>0.21</b>    | <b>0.074</b>  | 0.49                | 0.49                     | 0.49                    |



499  
 500  
 501  
 502  
 503  
 504  
 505  
 506  
**Figure 8: End-to-end latency analysis of TRUST framework. (Left) Breakdown of latency components**  
 in the TRUST pipeline for averaging 100 math problem. (Right) Comparison of total audit latency.  
 The additional latency provides robustness guarantees through distributed segment auditing.

## 5 CONCLUSION AND FUTURE WORKS

513 We introduce TRUST, the first decentralized framework for auditing the reasoning traces of Large  
 514 Reasoning Models that simultaneously addresses robustness, scalability, opacity, and privacy chal-  
 515 lenges. TRUST offers an end-to-end pipeline that integrates three key components: a Hierarchical  
 516 Directed Acyclic Graph (HDAG) decomposition method that breaks Chain-of-Thought reasoning into  
 517 five abstraction levels; a multi-tier consensus mechanism that routes verification tasks to automated  
 518 checkers, LLMs, and human experts based on complexity; and a blockchain-based infrastructure with  
 519 cryptographic privacy preservation that ensures transparent audit trails while protecting proprietary  
 520 model internals. It supports verification across diverse reasoning domains and enhances transparency  
 521 through decentralized consensus and immutable audit records.

522 Our experiments demonstrate TRUST’s effectiveness across correctness, bias mitigation, and human-  
 523 in-the-loop evaluation using multiple datasets and state-of-the-art models. TRUST consistently  
 524 outperforms centralized ensemble methods and single auditors while maintaining graceful degra-  
 525 dation under adversarial conditions. These results highlight its robustness against corruption and  
 526 effectiveness in eliminating systematic bias while preserving accuracy.

527 In addition, our theoretical framework provides formal guarantees that honest auditors profit while  
 528 malicious actors incur losses, creating sustainable economic incentives for real-world deployment.  
 529 Overall, TRUST pioneers decentralized AI auditing as a practical pathway toward safe and account-  
 530 able deployment of reasoning-capable AI systems in high-stakes domains. It makes transparent  
 531 oversight of proprietary AI systems accessible without compromising intellectual property rights.

532 **Future work** will involve developing more sophisticated graph decomposition methods to capture  
 533 richer reasoning dependencies, integrating adaptive auditor assignment strategies that leverage task-  
 534 specific expertise, and extending the framework to support dynamic, interactive reasoning settings.  
 535 Furthermore, we plan to conduct a longitudinal analysis of the on-chain data generated by TRUST to  
 536 study the long-term economic dynamics and emergent behaviors of the auditor network. Evaluating  
 537 the framework’s performance across different blockchain infrastructures would also provide deeper  
 538 insights into its practical scalability and cost-effectiveness. We also plan to explore cross-model  
 539 reasoning consistency verification for multi-agent scenarios and investigate integration with federated  
 learning frameworks to accelerate trustworthy AI deployment.

## 540 REFERENCES

541 NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist.*  
 542 *gov/nistpubs/ai/nist. ai*, pages 100–1, 2023. 1

543

544 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,  
 545 Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with  
 546 reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. 2

547

548 Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the  
 549 dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*  
 550 *conference on fairness, accountability, and transparency*, pages 610–623, 2021. 2

551

552 Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawska, Lukas Gianinazzi,  
 553 Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomska, Piotr Nyczek, et al. Graph of thoughts:  
 554 Solving elaborate problems with large language models. In *Proceedings of the AAAI conference*  
 555 *on artificial intelligence*, volume 38, pages 17682–17690, 2024. 19

556

557 Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong,  
 558 Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint*  
 559 *arXiv:2310.12941*, 2023. 2

560

561 Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. Are you getting what you pay for? auditing  
 562 model substitution in llm apis. *arXiv preprint arXiv:2504.04715*, 2025. 3, 19

563

564 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine  
 565 Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data  
 566 from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages  
 567 2633–2650, 2021. 2

568

569 Miguel Castro, Barbara Liskov, et al. Practical byzantine fault tolerance. In *OsDI*, volume 99, pages  
 570 173–186, 1999. 2, 3, 19

571

572 Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androulidakis, Daniel Martin  
 573 Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in  
 574 english. *arXiv preprint arXiv:2110.00976*, 2021. 1

575

576 Bing-Jyue Chen, Suppakit Waiwitlikhit, Ion Stoica, and Daniel Kang. Zkml: An optimizing system  
 577 for ml inference in zero-knowledge proofs. In *Proceedings of the Nineteenth European Conference*  
 578 *on Computer Systems*, pages 560–574, 2024. 3, 20

579

580 Wenhui Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting:  
 581 Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint*  
 582 *arXiv:2211.12588*, 2022. 19

583

584 Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman,  
 585 Arushi Soman, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always  
 586 say what they think. *arXiv preprint arXiv:2505.05410*, 2025. 19

587

588 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
 589 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John  
 590 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,  
 591 2021a. 63

592

593 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
 594 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John  
 595 Schulman. Training verifiers to solve math word problems, 2021b. URL <https://arxiv.org/abs/2110.14168>. 9

596

597 EU COM. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and  
 598 amending certain union legislative acts. *Proposal for a regulation of the European parliament and*  
 599 *of the council*, 2021. 1

600

601 Victor Costan and Srinivas Devadas. Intel sgx explained. *Cryptology ePrint Archive*, 2016. 19

594 Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large  
 595 language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022. 19  
 596

597 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,  
 598 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,  
 599 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao  
 600 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,  
 601 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,  
 602 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,  
 603 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang  
 604 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong,  
 605 Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao,  
 606 Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,  
 607 Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang,  
 608 Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L.  
 609 Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang,  
 610 Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng  
 611 Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanja Zhao, Wen Liu, Wenfeng  
 612 Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan  
 613 Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang,  
 614 Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen,  
 615 Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li,  
 616 Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang,  
 617 Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan,  
 618 Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue gong, Yuheng Zou, Yujia  
 619 He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong  
 620 Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha,  
 621 Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang,  
 622 Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,  
 623 Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen  
 624 Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.  
 625 URL <https://arxiv.org/abs/2501.12948>. 2

626 Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and  
 627 Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine  
 628 Learning*, pages 10764–10799. PMLR, 2023. 19

629 Zahra Ghodsi, Tianyu Gu, and Siddharth Garg. Safetynets: Verifiable execution of deep neural  
 630 networks on an untrusted cloud. *Advances in Neural Information Processing Systems*, 30, 2017. 20

631 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu  
 632 Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforce-  
 633 ment learning. *Nature*, 645(8081):633–638, 2025. 3, 19

634 Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey.  
 635 *arXiv preprint arXiv:2212.10403*, 2022. 19

636 Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzog, Or Honovich, Michael Tseng, Michael  
 637 Collins, Roei Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A  
 638 benchmark for verifiers of reasoning chains. *arXiv preprint arXiv:2402.00559*, 2024. 19

639 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
 640 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint  
 641 arXiv:2412.16720*, 2024. 3, 19

642 Matt Gardner Johannes Welbl, Nelson F. Liu. Crowdsourcing multiple choice science questions.  
 643 2017. 63

644 Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moon-  
 645 tae Lee, Honglak Lee, and Lu Wang. Process reward models that think. *arXiv preprint  
 646 arXiv:2504.16828*, 2025. 19

648 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large  
 649 language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:  
 650 22199–22213, 2022. 19

651 Vignesh Kothapalli, Hamed Firooz, and Maziar Sanjabi. Cot-icl lab: A synthetic framework for study-  
 652 ing chain-of-thought learning from in-context demonstrations. *arXiv preprint arXiv:2502.15132*,  
 653 2025. 5

654 UW-Madison Lee Lab. Mmlu-pro-cot-train-labeled. <https://huggingface.co/datasets/UW-Madison-Lee-Lab/MMLU-Pro-CoT-Train-Labeled>, 2024. Accessed: 2025-09-  
 655 24. 7

656 Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. In *Concur-  
 657 rency: the works of leslie lamport*, pages 203–226. 2019. 2

658 Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernan-  
 659 dez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in  
 660 chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. 1, 3, 19

661 Jixuan Leng, Cassandra A. Cohen, Zhixian Zhang, Chenyan Xiong, and William W. Cohen. Semi-  
 662 structured llm reasoners can be rigorously audited, 2025. URL <https://arxiv.org/abs/2505.24217>. 1, 3

663 Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian  
 664 Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language  
 665 models. *arXiv preprint arXiv:2211.09110*, 2022. 2, 19

666 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan  
 667 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth  
 668 International Conference on Learning Representations*, 2023. 2, 3, 19

669 Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su.  
 670 Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing  
 671 Systems*, 36:36407–36433, 2023. 19

672 Tianyi Liu, Xiang Xie, and Yupeng Zhang. Zkcnn: Zero knowledge proofs for convolutional neural  
 673 network predictions and accuracy. In *Proceedings of the 2021 ACM SIGSAC Conference on  
 674 Computer and Communications Security*, pages 2968–2985, 2021. 20

675 Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, GengRu Chen, Wenbo Su, and  
 676 Bo Zheng. Deconstructing long chain-of-thought: A structured reasoning optimization framework  
 677 for long cot distillation. *arXiv preprint arXiv:2503.16385*, 2025. 5

678 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson,  
 679 Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In  
 680 *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.  
 681 2

682 Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher  
 683 Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted  
 684 question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 19

685 Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito,  
 686 Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable  
 687 extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*,  
 688 2023. 2

689 NVIDIA, :, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Ren-  
 690 duchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek,  
 691 Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton,  
 692 Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, Ann Guan, Anna Shors, Anubhav  
 693 Mandarwal, Arham Mehta, Arun Venkatesan, Ashton Sharabiani, Ashwath Aithal, Ashwin Poojary,  
 694 Ayush Dattagupta, Balaran Buddharaju, Banghua Zhu, Barnaby Simkin, Bilal Kartal, Bita Darvish  
 695

702 Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Brian Yu, Bryan Catanzaro, Charles Wang,  
 703 Charlie Truong, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christian Munley, Christopher  
 704 Parisien, Dan Su, Daniel Afrimi, Daniel Korzekwa, Daniel Rohrer, Daria Gitman, David Mosal-  
 705 lanezhad, Deepak Narayanan, Dima Rekesh, Dina Yared, Dmytro Pykhtar, Dong Ahn, Duncan  
 706 Riach, Eileen Long, Elliott Ning, Eric Chung, Erick Galinkin, Evelina Bakhturina, Gargi Prasad,  
 707 Gerald Shen, Haifeng Qian, Haim Elisha, Harsh Sharma, Hayley Ross, Helen Ngo, Herman Sahota,  
 708 Hexin Wang, Hoo Chang Shin, Hua Huang, Iain Cunningham, Igor Gitman, Ivan Moshkov, Jaehun  
 709 Jung, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jian Zhang, Jiaqi Zeng, Jimmy Zhang, Jinze  
 710 Xue, Jocelyn Huang, Joey Conway, John Kamalu, Jonathan Cohen, Joseph Jennings, Julien Veron  
 711 Vialard, Junkeun Yi, Jupinder Parmar, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss,  
 712 Keshav Santhanam, Kezhi Kong, Krzysztof Pawelec, Kumar Anik, Kunlun Li, Kushan Ahma-  
 713 dian, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Luis Vega, Maer Rodrigues de Melo,  
 714 Makesh Narsimhan Sreedhar, Marcin Chochowski, Mark Cai, Markus Kliegl, Marta Stepniewska-  
 715 Dzubinska, Matvei Novikov, Mehrzad Samadi, Meredith Price, Meriem Boudir, Michael Boone,  
 716 Michael Evans, Michal Bien, Michal Zawalski, Miguel Martinez, Mike Chrzanowski, Mohammad  
 717 Shoeybi, Mostofa Patwary, Namit Dhameja, Nave Assaf, Negar Habibi, Nidhi Bhatia, Nikki  
 718 Pope, Nima Tajbakhsh, Nirmal Kumar Juluru, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev,  
 719 Oluwatobi Olabiyi, Pablo Ribalta, Padmavathy Subramanian, Parth Chadha, Pavlo Molchanov,  
 720 Peter Dykas, Peter Jin, Piotr Bialecki, Piotr Januszewski, Pradeep Thalasta, Prashant Gaikwad,  
 721 Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Rabeeh Karimi Mahabadi, Rajen Patel, Ran  
 722 El-Yaniv, Ranjit Rajan, Ria Cheruvu, Rima Shahbazyan, Ritika Borkar, Ritu Gala, Roger Waleffe,  
 723 Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Sahil Jain, Samuel Kriman, Sanjeev Satheesh, Saori  
 724 Kaji, Sarah Yurick, Saurav Muralidharan, Sean Narenthiran, Seonmyeong Bak, Sepehr Sameni,  
 725 Seungju Han, Shanmugam Ramasamy, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby  
 726 Thomas, Shizhe Diao, Shreya Gopal, Shrimai Prabhuloye, Shubham Toshniwal, Shuoyang Ding,  
 727 Siddharth Singh, Siddhartha Jain, Somshubra Majumdar, Soumye Singhal, Stefania Alborghetti,  
 728 Syeda Nahida Akter, Terry Kong, Tim Moon, Tomasz Hliwiak, Tomer Asida, Tony Wang, Tugrul  
 729 Konuk, Twinkle Vashishth, Tyler Poon, Udi Karpas, Vahid Noroozi, Venkat Srinivasan, Vijay  
 730 Korthikanti, Vikram Fugro, Vineeth Kalluru, Vitaly Kurin, Vitaly Lavrukhin, Wasi Uddin Ahmad,  
 731 Wei Du, Wonmin Byeon, Ximing Lu, Xin Dong, Yashaswi Karnati, Yejin Choi, Yian Zhang, Ying  
 732 Lin, Yonggan Fu, Yoshi Suhara, Zhen Dong, Zhiyu Li, Zhongbo Zhu, and Zijia Chen. Nvidia  
 733 nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025.  
 734 URL <https://arxiv.org/abs/2508.14444>. 19

735 Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David  
 736 Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work:  
 737 Scratchpads for intermediate computation with language models. 2021. 19

738 Forty-Two Countries Adopt New OECD. Principles on artificial intelligence, 2019. 1

739 OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>. 2

740 Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale  
 741 multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores,  
 742 George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the  
 743 Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning  
 744 Research*, pages 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>. 63

745 Dario Pasquini, Evgenios M Kornaropoulos, and Giuseppe Ateniese. {LLMmap}: Fingerprinting  
 746 for large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages  
 747 299–318, 2025. 19

748 Zhizhi Peng, Taotao Wang, Chonghe Zhao, Guofu Liao, Zibin Lin, Yifeng Liu, Bin Cao, Long Shi,  
 749 Qing Yang, and Shengli Zhang. A survey of zero-knowledge proof based verifiable machine  
 750 learning. *arXiv preprint arXiv:2502.18535*, 2025. 3, 19

751 Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv  
 752 preprint arXiv:2211.09527*, 2022. 2

756 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for  
 757 machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings*  
 758 of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392,  
 759 Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/

760 D16-1264. URL <https://aclanthology.org/D16-1264>. 63

761 Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. Trusted execution environ-  
 762 ment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/Ispa*, volume 1, pages  
 763 57–64. IEEE, 2015. 19

764

765 Morteza Sarafyazd and Mehrdad Jazayeri. Hierarchical reasoning by neural circuits in the frontal  
 766 cortex. *Science*, 364(6441):eaav8911, 2019. doi: 10.1126/science.aav8911. URL <https://www.science.org/doi/abs/10.1126/science.aav8911>. 5

767

768 Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke  
 769 Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach  
 770 themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551,  
 771 2023. 19

772

773 Christoph Schnabl, Daniel Hugenroth, Bill Marino, and Alastair R Beresford. Attestable audits: Verifi-  
 774 able ai safety benchmarks using trusted execution environments. *arXiv preprint arXiv:2506.23706*,  
 775 2025. 20

776

777 Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red  
 778 teaming language model detectors with language models. *Transactions of the Association for  
 Computational Linguistics*, 12:174–189, 2024. 19

779

780 Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan  
 781 Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfahl, et al. Large language models encode  
 782 clinical knowledge. *Nature*, 620(7972):172–180, 2023. 1

783

784 Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason  
 785 Morton, and Alex’ Sandy’ Pentland. Verifiable evaluations of machine learning models using  
 786 zksnarks. *arXiv preprint arXiv:2402.02675*, 2024. 19

787

788 Guoheng Sun, Ziyao Wang, Bowei Tian, Meng Liu, Zheyu Shen, Shuai He, Yexiao He, Wanghao Ye,  
 789 Yiting Wang, and Ang Li. Coin: Counting the invisible reasoning tokens in commercial opaque  
 790 llm apis. *arXiv preprint arXiv:2505.13778*, 2025. 3, 19

791

792 Haochen Sun, Jason Li, and Hongyang Zhang. zkllm: Zero knowledge proofs for large language mod-  
 793 els. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications  
 794 Security*, pages 4405–4419, 2024. 3, 20

795

796 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question  
 797 answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of  
 798 the North American Chapter of the Association for Computational Linguistics: Human Language  
 799 Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June  
 800 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421>. 63

801

802 Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Man-  
 803 sooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation:  
 804 A survey. *arXiv preprint arXiv:2402.13446*, 2024. 63

805

806 Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don’t always  
 807 say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural  
 808 Information Processing Systems*, 36:74952–74965, 2023. 1, 19

809

810 Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia  
 811 Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and  
 812 outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. 19

813

814 UW-Madison-Lee-Lab. Mmlu-pro-cot-train-labeled. <https://huggingface.co/datasets/UW-Madison-Lee-Lab/MMLU-Pro-CoT-Train-Labeled>, 2024. 63

810 Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for  
 811 open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*, 2023. 1  
 812

813 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-  
 814 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.  
 815 *arXiv preprint arXiv:2203.11171*, 2022. 19

816 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny  
 817 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*  
 818 *neural information processing systems*, 35:24824–24837, 2022. 1, 3, 19

819 Gavin Wood et al. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project*  
 820 *yellow paper*, 151(2014):1–32, 2014. 19

821

822 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang  
 823 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,  
 824 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
 825 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,  
 826 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui  
 827 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang  
 828 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger  
 829 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan  
 830 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>. 2

831 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan.  
 832 Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural*  
 833 *information processing systems*, 36:11809–11822, 2023a. 2, 3, 19

834 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.  
 835 React: Synergizing reasoning and acting in language models. In *International Conference on*  
 836 *Learning Representations (ICLR)*, 2023b. 19

837

838 Maofan Yin, Dahlia Malkhi, Michael K Reiter, Guy Golan Gueta, and Ittai Abraham. Hotstuff: Bft  
 839 consensus with linearity and responsiveness. In *Proceedings of the 2019 ACM symposium on*  
 840 *principles of distributed computing*, pages 347–356, 2019. 19

841 Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. Star: Bootstrapping reasoning with  
 842 reasoning, 2022. URL <https://arxiv.org/abs/2203.14465>, 2203, 2022. 19

843

844 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,  
 845 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and  
 846 chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023. 3, 19

847 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans,  
 848 Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning  
 849 in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 19

850

851 Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal  
 852 and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*,  
 853 2023. 2

854

855

856

857

858

859

860

861

862

863

# 864 Appendix

|     |   |           |
|-----|---|-----------|
| 868 | <b>A The Use of Large Langauge Models (LLMs)</b>                          | <b>18</b> |
| 869 | <b>B More Related Works</b>   | <b>19</b> |
| 870 | B.1 Chain-of-Thought Reasoning . . . . .                                  | 19        |
| 871 | B.2 Auditing and Evaluation of LLM Reasoning . . . . .                    | 19        |
| 872 | B.3 Decentralized Consensus and Privacy-Preserving Verification . . . . . | 19        |
| 873 |   |           |
| 874 | <b>C Experimental Details</b>   | <b>21</b> |
| 875 | C.1 TRUST Implementation Details . . . . .                                | 21        |
| 876 | C.1.1 Step 1: Abstract Level Identification. . . . .                      | 23        |
| 877 | C.1.2 Step 2: Segment Within Each Abstraction Level. . . . .              | 26        |
| 878 | C.1.3 Step 3: Extract Relationships. . . . .                              | 32        |
| 879 | C.1.4 Step 4: Assign Auditor Types. . . . .                               | 36        |
| 880 | C.1.5 Step 5: Refine and HDAG Construction. . . . .                       | 41        |
| 881 | C.2 Human-Friendly Audit Example . . . . .                                | 52        |
| 882 | C.2.1 Audit Interface Structure . . . . .                                 | 52        |
| 883 | C.2.2 Example Audit Case . . . . .  | 52        |
| 884 | C.2.3 Audit Decision Framework . . . . .                                  | 52        |
| 885 | C.3 HDAG Example 1 - Math Problem . . . . .                               | 54        |
| 886 | C.3.1 Deepseek-R1-8B CoT Decomposition . . . . .                          | 54        |
| 887 | C.3.2 GPT-OSS-20B CoT Decomposition . . . . .                             | 56        |
| 888 | C.4 HDAG Example 2 - Math Problem . . . . .                               | 59        |
| 889 | C.4.1 Deepseek-R1-8B CoT Decomposition . . . . .                          | 59        |
| 890 | C.4.2 GPT-OSS-20B CoT Decomposition . . . . .                             | 61        |
| 891 | C.5 Datasets . . . . .  | 63        |
| 892 | <b>D Theoretical Results</b>  | <b>64</b> |
| 893 | D.1 Consensus: Seat Layer, Segment Layer, and Trace Layer . . . . .       | 64        |
| 894 | D.2 Economic Layer: Staking, Reputation, Rewards, and Slashing . . . . .  | 65        |
| 895 | <b>E Motivating Example - Why Semantic Audit is Necessary</b>             | <b>71</b> |
| 896 | E.1 Why TRUST Semantic Auditing is Essential . . . . .                    | 73        |

897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

918 **A THE USE OF LARGE LANGAUGE MODELS (LLMs)**  
919920 To enhance clarity and readability, this paper utilized Large Language Models exclusively as language  
921 polishing tools. Their role was confined to general proofreading and writing assistance—functions  
922 analogous to those provided by traditional grammar checkers and dictionaries. These tools did not  
923 contribute to the generation of new scientific content or ideas, and their usage is consistent with  
924 standard practices for manuscript preparation.

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972 **B MORE RELATED WORKS**  
973974 **B.1 CHAIN-OF-THOUGHT REASONING**  
975

976 Exposing intermediate reasoning steps, popularized by Chain-of-Thought (CoT) prompting (Wei  
977 et al., 2022), has become a cornerstone for enhancing the capabilities of Large Language Models  
978 (LLMs) (Kojima et al., 2022; Wang et al., 2022; Nye et al., 2021; Zhou et al., 2022; Zelikman et al.,  
979 2022). However, this paradigm has rapidly evolved, and the resulting complexity and length of  
980 reasoning traces present a critical, unsolved challenge for verification (Lightman et al., 2023; Jacovi  
981 et al., 2024; Ling et al., 2023; Chen et al., 2025). Initial research focused on eliciting reasoning and  
982 moving beyond simple linear chains to more structured representations (Creswell et al., 2022; Zhou  
983 et al., 2022; Besta et al., 2024; Chen et al., 2022). For instance, Tree-of-Thoughts (ToT) generalized  
984 CoT to a tree search, enabling explicit exploration and backtracking within the reasoning process (Yao  
985 et al., 2023a). Concurrently, a parallel line of work enabled models to offload complex calculations  
986 and external queries to tools, separating logical deduction from information retrieval (Schick et al.,  
987 2023; Yao et al., 2023b; Gao et al., 2023; Nakano et al., 2021). More recent efforts have focused on  
988 Large Reasoning Models (LRMs) that treat reasoning as a primary objective, allocating substantial  
989 computational resources to the process (Jaech et al., 2024; Guo et al., 2025; NVIDIA et al., 2025).  
990 These models are trained with large-scale reinforcement learning from process-level feedback and  
991 employ verifiers to guide multi-path search during inference (Huang and Chang, 2022). Despite  
992 these advances in generating sophisticated reasoning, the mechanisms for *auditing* these processes in  
993 a scalable, privacy-preserving, and decentralized manner have remained largely unexplored (Peng  
994 et al., 2025). In this paper, we address this gap by introducing a framework that decomposes complex  
995 reasoning traces into auditable graphs (HDAGs) and leverages a decentralized network for verification,  
996 making it suitable for the proprietary and intricate nature of modern LRMs.  
997

998 **B.2 AUDITING AND EVALUATION OF LLM REASONING**  
999

1000 As the complexity of reasoning in LLMs increases, methods for auditing its quality, faithfulness, and  
1001 safety have become a critical area of research (Liang et al., 2022; Lanham et al., 2023). An initial and  
1002 widely adopted approach relies on a centralized “LLM-as-a-judge,” which, while scalable, is known to  
1003 suffer from inherent biases and constitutes a single point of failure (Zheng et al., 2023). A significant  
1004 advancement was the shift from auditing final outcomes to verifying the reasoning *process* itself,  
1005 primarily through Process Reward Models (PRMs) that provide step-by-step supervision (Lightman  
1006 et al., 2023; Uesato et al., 2022; Khalifa et al., 2025). The limitations of this approach were  
1007 subsequently highlighted by the discovery of unfaithful reasoning, where models generate plausible-  
1008 looking steps that do not reflect their true decision-making process, underscoring the need for audits  
1009 to defend against strategic deception (Turpin et al., 2023). More recently, research has expanded to  
1010 address service-level integrity in opaque commercial settings, including work on detecting model  
1011 substitution and using cryptographic methods to verify the integrity of the inference process itself (Shi  
1012 et al., 2024; Cai et al., 2025; Pasquini et al., 2025; South et al., 2024). For example, CoIn introduces a  
1013 method to audit “invisible” reasoning tokens by using hash trees over embeddings, ensuring providers  
1014 do not inflate billing without revealing proprietary content (Sun et al., 2025). While these works  
1015 address vital concerns like inference integrity or billing, they do not offer a unified framework for  
1016 scalable *semantic* auditing that combines decentralized consensus to mitigate bias with economic  
1017 incentives to ensure network security. This work synthesizes these needs by proposing a framework  
1018 that performs process-level semantic auditing on partial traces, uses a decentralized multi-tier auditor  
1019 network to ensure robustness, and is secured by theoretically-grounded economic guarantees.  
1020

1021 **B.3 DECENTRALIZED CONSENSUS AND PRIVACY-PRESERVING VERIFICATION**  
1022

1023 Building systems that ensure integrity and privacy without a central trusted party has a long-standing  
1024 foundation in decentralized consensus, cryptography, and confidential computing (Castro et al., 1999;  
1025 Sabt et al., 2015; Costan and Devadas, 2016). These technologies provide the essential primitives  
1026 for building auditable systems for proprietary models, where model internals and user data must  
1027 remain confidential. Foundational work established Byzantine Fault Tolerant (BFT) consensus for  
1028 robust agreement and public ledgers for tamper-evident records (Castro et al., 1999; Wood et al.,  
1029 2014; Yin et al., 2019), while Trusted Execution Environments (TEEs) provided hardware-based  
1030 confidential compute (Sabt et al., 2015). A prominent line of recent research focuses on applying these  
1031

1026 primitives directly to language models, particularly using Zero-Knowledge Proofs (ZKPs) to create  
1027 verifiable attestations of model inference without revealing the model’s weights (Chen et al., 2024;  
1028 Liu et al., 2021; Ghodsi et al., 2017). Specifically, zkLLM (Sun et al., 2024) introduced specialized  
1029 cryptographic techniques to efficiently prove the correctness of attention mechanisms, demonstrating  
1030 that even complex transformer computations can be cryptographically verified. Another approach  
1031 uses TEEs to run auditable safety benchmarks on proprietary models, publishing hardware-signed  
1032 attestations to a public log to prove that a specific model version passed certain tests (Schnabl et al.,  
1033 2025). While this work is crucial for verifying *that* a specific computation was performed correctly  
1034 by a specific model, it does not address the orthogonal problem of verifying the *semantic quality*,  
1035 *faithfulness*, and *safety* of the reasoning trace through a scalable, human-in-the-loop consensus  
1036 process. Our work provides a novel synthesis by using a public ledger and consensus not just  
1037 for recording outcomes, but for orchestrating a decentralized, multi-tier network of auditors and  
1038 managing economic incentives, creating the first system designed for end-to-end semantic auditing of  
1039 model reasoning.

1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

1080 **C EXPERIMENTAL DETAILS**1081 **C.1 TRUST IMPLEMENTATION DETAILS**1082 This section provide the LLMs instruction of our auto data processing framework and the example  
1083 output from gpt-oss-120b.

1084

1085 **Example CoT** Below is a example CoT trace.

1086

```

1087
1088     ### Input Format:
1089     ## Problem Statement:
1090     A construction company was building a tunnel. When  $\frac{1}{3}$  of the
1091     → tunnel was completed at the original speed, they started usin
1092     g new equipment, which increased the construction speed by  $20\%$  and
1093     → reduced the working hours to  $80\%$  of the original. As a resu
1094     lt, it took a total of 185 days to complete the tunnel. If they had not
1095     → used the new equipment and continued at the original speed, it
1096     would have taken  $\sqrt[4]{3}$  days to complete the tunnel.
1097
1098     ## Thinking Process:
1099     We have a tunnel. Original speed = some rate (units per day). They
1100     → built  $1/3$  of tunnel at original speed. Then they switched to new
1101     equipment: speed increased by 20%, i.e., new speed =  $1.2 * \text{original}$ 
1102     → speed. Working hours reduced to 80% of original, meaning per day
1103     they work less time, so effective daily output? If originally they
1104     → worked some hours per day, now they work only 80% of that, but
1105     speed is per hour? Let's interpret: Original speed in terms of distance
1106     → per hour? Or maybe speed per day? Usually speed is distance
1107     per hour. If they increase speed by 20% (i.e., per hour they do 1.2
1108     → times) but they work only 80% of the hours per day, then daily
1109     production = speed_per_hour * hours_per_day. If original daily
1110     → production =  $v * h$  (where  $v$  is original speed per hour,  $h$  original
1111     hours per day). New daily production =  $(1.2 v) * (0.8 h) = 0.96 v h =$ 
1112     → 0.96 * original daily production. So overall daily rate
1113     actually decreased by 4% compared to original! That seems odd: they
1114     → increased speed but reduced hours, net 0.96 factor. But maybe
1115     they meant speed (i.e., daily output) increased by 20% and hours
1116     → reduced to 80%, i.e., effective speed per day is  $1.2 * 0.8 = 0.96$  of
1117     original daily work, which is slower. But they say "As a result, it
1118     → took a total of 185 days to complete the tunnel." That implies
1119     slower after switching? The initial third took some days, then
1120     → remaining two-thirds took more days than original would have. Let's
1121     compute.
1122
1123     Let total tunnel length = 1 (normalized). Original daily output =  $D$ 
1124     → (units per day). So original time to finish whole tunnel =  $1/D$ 
1125     days.
1126
1127     They did first  $1/3$  at original speed: time  $t_1 = (1/3)/D = 1/(3D)$ .
1128
1129     After switch, new daily output =  $0.96 D$  (as above). Remaining work =
1130     →  $2/3$ . Time  $t_2 = (2/3) / (0.96 D) = (2/3) * (1/(0.96 D)) = (2)/(3 * 0.96 D) = (2)/(2.88 D) = (0.694444...)/D$ ? Let's compute precisely:
1131     →  $2/(2.88) = 0.694444...$  So  $t_2 = 0.694444... / D$ .
1132
1133     Total days =  $t_1 + t_2 = 1/(3D) + 0.694444.../D = (1/3 + 0.694444...)/D =$ 
1134     →  $(0.333333... + 0.694444...)/D = 1.027777... / D$ .
1135
1136     Given total days = 185. So  $(1.027777...)/D = 185 \Rightarrow D = 1.027777... /$ 
1137     → 185 = approx 0.005555... Wait compute:  $1.0277777778 / 185 =$ 
1138     0.005555556... Indeed 0.0055556 per day. That seems like original daily
1139     → output = 0.0055556 of tunnel per day, i.e., whole tunnel
1140     would take  $1/D = 180$  days. Let's check:  $1/D = 180$ . Exactly? Since  $D =$ 
1141     →  $1/180 = 0.00555555...$  Multiply:  $1/D = 180$ . Let's see if D

```

```

1134 computed equals 1/180.  $1/180 = 0.0055555556$ . Multiply by 185 gives
1135  $\rightarrow 185/180 = 1.027777778$ , matches numerator. So indeed  $D = 1/180$ .
1136
1137 Thus original schedule would take 180 days to complete tunnel.
1138
1139 But wait, they gave total actual days 185, which is longer than
1140  $\rightarrow$  original 180. So using new equipment actually slowed down. The
1141 question: "If they had not used the new equipment and continued at the
1142  $\rightarrow$  original speed, it would have taken ____ days to complete the
1143 tunnel." That's 180 days.
1144
1145 Thus answer: 180 days.
1146
1147 Double-check: We assumed original speed per hour, then hours reduced.
1148  $\rightarrow$  Alternative interpretation: They increased speed (distance per
1149 hour) by 20% but also reduced working hours per day to 80% of original,
1150  $\rightarrow$  net daily output 0.96 of original. So yes slower.
1151
1152 Thus answer 180.
1153
1154 Check using algebra: Let original speed =  $s$  (units per day). Then first
1155  $\rightarrow$  third takes  $(1/3)/s$  days. After switch, speed becomes 1.2  $s$ 
1156 but hours per day 0.8, so effective speed per day = 0.96  $s$ . Remaining
1157  $\rightarrow$  2/3 takes  $(2/3)/(0.96 s)$  days. Sum = 185. Solve:  $(1/3 +$ 
1158  $(2/3)/0.96/s = 185$ .  $(2/3)/0.96 = (2/3) * (1/0.96) =$ 
1159  $\rightarrow (2/3) * (1.0416667) = 0.694444...$  Add  $1/3 = 0.333333 + 0.694444$ 
1160  $\rightarrow = 1.0277778$ . So  $s$ 
1161  $= 1.0277778/185 = \text{approx } 0.0055556$ . Full tunnel time =  $1/s = 185/1.0277778$ 
1162  $\rightarrow = 180$ .
1163
1164 Thus answer 180.
1165
1166 ## Final Output:
1167 Let
1168
1169 * the length of the tunnel be  $\backslash(1\backslash)$  (as a whole),
1170 * the original daily construction rate be  $\backslash(r\backslash)$  (tunnel-fractions per
1171  $\rightarrow$  day).
1172
1173 ---
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
2999

```

```

1188
1189 \[
1190 t_{2}=\frac{\frac{23}{3}\cdot 0.96}{r}=\frac{2}{3\cdot 0.96}\cdot r=
1191 \frac{0.694444}{r}.
1192 \]
1193 ---
1194
1195 ### 3. Total actual time
1196
1197 \[
1198 t_1+t_2=185\text{ days}.
1199 \]
1200 \[
1201 \frac{1}{3r}+\frac{0.694444}{r}=185
1202 \quad\Rightarrow\quad
1203 \frac{1.027777}{r}=185.
1204 \]
1205 Hence
1206
1207 \[
1208 r=\frac{1.027777}{185}=\frac{1}{180};\text{ (tunnel per day)}.
1209 \]
1210 ---
1211
1212 ### 4. Time that would have been required without the new equipment
1213
1214 If the original rate  $(r=\frac{1}{180})$  had been kept for the whole
1215 tunnel, the total time would be
1216 \[
1217 \frac{1}{r}=180\text{ days}.
1218 \]
1219 \[
1220 \boxed{180}
1221 \]

```

### C.1.1 STEP 1: ABSTRACT LEVEL IDENTIFICATION.

1226 Below are the instruction of how to identify different abstraction levels.

```
1227
1228 You are an abstraction level identification agent. Your task is to
1229 ↵ decompose a complete reasoning trace (problem + thinking + output)
1230 ↵ into distinct abstraction level blocks following a hierarchical
1231 ↵ framework.
1232
1233     ### 1. Abstraction Level Framework:
1234     - **GOAL**: What we want to achieve (objective, end goal, problem
1235       ↵ statement)
1236     - **STRATEGY**: High-level approaches (methodology, overall plan, major
1237       ↵ techniques)
1238     - **TACTIC**: Specific techniques (algorithms, methods, key insights,
1239       ↵ definitions)
1240     - **STEP**: Individual reasoning moves (logical steps, inferences,
1241       ↵ deductions)
1242     - **OPERATION**: Atomic calculations (computations, lookups, examples,
1243       ↵ numerical work)
1244
1245     ### 2. Identification Rules:
```

```

1242 - Split the reasoning into blocks that represent distinct abstraction
1243 ↵ levels
1244 - A block can be multiple sentences or paragraphs
1245 - Focus on SEMANTIC CONTENT, not temporal order
1246 - Some levels might be missing - that's acceptable
1247 - Some levels might have multiple blocks - that's acceptable
1248 - Preserve exact text spans from the original trace

1249 ##### 3. Format Requirements:
1250 - Present output in "# Abstraction Block Analysis" section
1251 - Under each level, use "##" headings (## GOAL Level, ## STRATEGY
1252 ↵ Level, etc.)
1253 - Include exact original text under each heading
1254 - Preserve all line breaks and mathematical notation
1255 - After all blocks, add "# Block Structure" section with summary

1256 ##### 4. Content Distribution Guidelines:
1257 - **GOAL** must include the complete problem statement and objectives
1258 - **STRATEGY** should contain high-level approaches before detailed work
1259 - **TACTIC** should include specific methods, algorithms, and key
1260 ↵ insights
1261 - **STEP** should contain individual logical moves and reasoning steps
1262 - **OPERATION** should contain all calculations, examples, and
1263 ↵ numerical work

1264 ##### 5. Critical Requirements:
1265 - Preserve all original mathematical notation exactly
1266 - Maintain all line breaks as they appear in the original text
1267 - Include all text exactly as written without modifications
1268 - Ensure block boundaries occur at natural semantic transitions
1269 - Extract EXACT text spans - do not paraphrase

1270 ##### Input Format:
1271 ## Problem Statement:
1272 [original problem text]

1273 ## Thinking Process:
1274 [complete reasoning trace]

1275 ## Final Output:
1276 [final answer/solution]

1277 ##### Output Format:
1278 # Abstraction Block Analysis

1279 ## GOAL Level
1280 [exact text expressing objectives and problem statements]

1281 ## STRATEGY Level
1282 [exact text describing high-level approaches]

1283 ## TACTIC Level
1284 [exact text detailing specific techniques and methods]

1285 ## STEP Level
1286 [exact text showing individual reasoning moves]

1287 ## OPERATION Level
1288 [exact text containing calculations and examples]

1289 # Block Structure
1290 {
1291   "GOAL": [list of goal block summaries],
1292   "STRATEGY": [list of strategy block summaries],
1293   "TACTIC": [list of tactic block summaries],

```

```

1296     "STEP": [list of step block summaries],
1297     "OPERATION": [list of operation block summaries]
1298 }
1299
1300

```

**Step 1: Example Output.** Below is the example output for the first step.

```

1301
1302
1303 # Abstraction Block Analysis
1304
1305 ## GOAL Level
1306 A construction company was building a tunnel. When  $\tfrac{1}{3}$  of
1307  $\rightarrow$  the tunnel was completed at the original speed, they started using
1308  $\rightarrow$  new equipment, which increased the construction speed by  $20\%$  and
1309  $\rightarrow$  reduced the working hours to  $80\%$  of the original. As a result, it
1310  $\rightarrow$  took a total of 185 days to complete the tunnel. If they had not
1311  $\rightarrow$  used the new equipment and continued at the original speed, it would
1312  $\rightarrow$  have taken  $185 \times 3 = 555$  days to complete the tunnel.
1313
1314 ## STRATEGY Level
1315 We have a tunnel. Original speed = some rate (units per day). They
1316  $\rightarrow$  built  $\tfrac{1}{3}$  of tunnel at original speed. Then they switched
1317  $\rightarrow$  to new equipment: speed increased by  $20\%$ , i.e., new speed  $= 1.2$ 
1318  $\rightarrow$   $\times$  original speed. Working hours reduced to  $80\%$  of original,
1319  $\rightarrow$  meaning per day they work fewer hours.
1320
1321 If original daily production =  $v \cdot h$  (speed per hour  $\times$  hours per day), then new daily production =
1322  $\rightarrow$   $1.2v \cdot 0.8h = 0.96vh = 0.96 \times$  original daily output.
1323
1324 So overall daily rate actually decreased by  $4\%$ . This explains why
1325  $\rightarrow$  the total time (185 days) was longer than the original time would
1326  $\rightarrow$  have been.
1327
1328 ## TACTIC Level
1329 Let total tunnel length =  $1$ .
1330 Let original daily output =  $D$ . Then original completion time =  $1/D$ .
1331
1332 - First  $\tfrac{1}{3}$  of tunnel:
1333    $t_1 = \tfrac{1}{3D}$ .
1334
1335 - Remaining  $\tfrac{2}{3}$  at reduced daily rate  $0.96D$ :
1336    $t_2 = \tfrac{2}{3} \cdot 0.96D = \tfrac{0.694444\ldots}{D}$ .
1337
1338 Total time:
1339  $\left[ t_1 + t_2 = \tfrac{1}{3D} + \tfrac{0.694444\ldots}{D} = 185 \right]$ 
1340
1341 So
1342  $\left[ \tfrac{1.027777\ldots}{D} = 185 \right] \Rightarrow D = \tfrac{1.027777\ldots}{185} = \tfrac{1}{180}$ .
1343
1344 Thus original schedule = 180 days.
1345
1346 ## STEP Level
1347 1. Compute new daily output:  $0.96D$ .
1348 2. Time for first third:  $t_1 = \tfrac{1}{3D}$ .
1349 3. Time for remaining two-thirds:
1350    $t_2 = \tfrac{2}{3} \cdot 0.96D = \tfrac{0.694444\ldots}{D}$ .
1351 4. Total:  $t_1 + t_2 = 185$ .
1352 5. Combine:  $\tfrac{1.027777\ldots}{D} = 185$ .
1353 6. Solve:  $D = \tfrac{1}{180}$ .

```

```

1350    7. Original time:  $\$1/D=180$  days.
1351
1352    ## OPERATION Level
1353    \[
1354    \frac{2}{3} \cdot 0.96 = \frac{2}{3} \cdot 0.694444 \ldots
1355    \]
1356
1357    \[
1358    \frac{1}{3} + 0.694444 \ldots = 1.027777 \ldots
1359    \]
1360
1361    D = \frac{1.027777 \ldots}{185} = 0.005555 \ldots = \frac{1}{180}
1362
1363
1364    \[
1365    \text{Original total time} = \frac{1}{D} = 180 \text{ days}
1366    \]
1367
1368    ---
1369
1370    # Block Structure
1371
1372    ````json
1373    {
1374        "GOAL": [
1375            "A construction company was building a tunnel. When 1/3 of the
1376            tunnel was completed at the original speed, they started using
1377            new equipment, which increased the construction speed by 20% and
1378            reduced the working hours to 80% of the original. As a result, it
1379            took a total of 185 days to complete the tunnel. If they had not
1380            used the new equipment and continued at the original speed, it
1381            would have taken ____ days to complete the tunnel."
1382        ],
1383        "STRATEGY": [
1384            "We have a tunnel. Original speed = some rate (units per day). They
1385            built 1/3 of tunnel at original speed. Then they switched to new
1386            equipment: speed increased by 20%, i.e., new speed = 1.2 *
1387            original speed. Working hours reduced to 80% of original, meaning
1388            per day they work less time. New daily production = 0.96 *
1389            original daily production. That means slower than original, which
1390            explains why it took 185 days in total."
1391        ],
1392        "TACTIC": [
1393            "Let total tunnel length = 1. Original daily output = D. Time for
1394            1/3 = 1/(3D). Time for remaining 2/3 at 0.96D =
1395            (2/3)/(0.96D) = 0.694444.../D. Total = (1/3 + 0.694444...)/D = 185.
1396            Solve: D = 1/180. So original full time = 180 days."
1397        ],
1398        "STEP": [
1399            "1. New daily output = 0.96D. 2.  $t_1 = 1/(3D)$ . 3.  $t_2 = 0.694444.../D$ . 4.
1400             $t_1 + t_2 = 185$ . 5.  $(1.027777...)/D = 185$ . 6.  $D = 1/180$ . 7. Original total
1401            time = 180 days."
1402        ],
1403        "OPERATION": [
1404            "2/(3*0.96) = 0.694444..., 1/3 + 0.694444... = 1.027777..., D = 1/180, so
1405            total original time = 180 days."
1406        ]
1407    }

```

### C.1.2 STEP 2: SEGMENT WITHIN EACH ABSTRACTION LEVEL.

Below are the instruction of how to segment within each abstraction level.

```

1404
1405 You are a reasoning segmentation agent. Your task is to break down
1406 ↢ abstraction-level blocks into appropriately-sized segments for
1407 ↢ verification.
1408
1409 ### 1. Segmentation Strategy by Level:
1410
1411 **GOAL Level Segmentation:**
1412 - Keep objectives as complete statements
1413 - Don't over-segment problem definitions
1414 - One main goal per segment typically
1415 - Preserve semantic completeness
1416
1417 **STRATEGY Level Segmentation:**
1418 - Segment by distinct approaches or methodologies
1419 - Each strategy should be a complete approach
1420 - Don't break up coherent strategic thinking
1421 - Maintain approach integrity
1422
1423 **TACTIC Level Segmentation:**
1424 - Segment by specific techniques, key insights, or algorithm components
1425 - Each tactic should be independently understandable
1426 - Break at natural technique boundaries
1427 - Preserve method coherence
1428
1429 **STEP Level Segmentation:**
1430 - Segment by individual logical moves
1431 - Each step should be a single inference or reasoning move
1432 - Break at logical transition points
1433 - Maintain reasoning flow
1434
1435 **OPERATION Level Segmentation:**
1436 - Segment by atomic calculations or examples
1437 - Each operation should be independently verifiable
1438 - Break at calculation boundaries
1439 - Preserve computational completeness
1440
1441 ### 2. Format Requirements:
1442 - Present output in "# Segmentation Analysis" section
1443 - Use "##" for each abstraction level
1444 - Use "###" for individual segments within levels
1445 - Preserve all mathematical notation and formatting
1446 - Include segment metadata
1447
1448 ### 3. Segment Metadata:
1449 For each segment, provide:
1450 - **segment_id**: Unique identifier (G1, S1, T1, ST1, O1, etc.)
1451 - **content**: Exact reasoning content
1452 - **type**: Specific type within abstraction level
1453 - **summary**: Brief 3-5 word summary
1454 - **verification_complexity**: Low/Medium/High
1455
1456 ### Input Format:
1457 # Abstraction Block Analysis
1458 [output from Step 1]
1459
1460 ### Output Format:
1461 # Segmentation Analysis
1462
1463 ## GOAL Segmentation
1464
1465 ### Segment G1
1466 **Content**: [exact segment content]
1467 **Type**: [segment type]
1468 **Summary**: [brief summary]

```

```

1458  **Complexity**: [Low/Medium/High]
1459
1460  ## STRATEGY Segmentation
1461
1462  ### Segment S1
1463  **Content**: [exact segment content]
1464  **Type**: [segment type]
1465  **Summary**: [brief summary]
1466  **Complexity**: [Low/Medium/High]
1467
1468  ## TACTIC Segmentation
1469
1470  ### Segment T1
1471  **Content**: [exact segment content]
1472  **Type**: [segment type]
1473  **Summary**: [brief summary]
1474  **Complexity**: [Low/Medium/High]
1475
1476  ## STEP Segmentation
1477
1478  ### Segment ST1
1479  **Content**: [exact segment content]
1480  **Type**: [segment type]
1481  **Summary**: [brief summary]
1482  **Complexity**: [Low/Medium/High]
1483
1484  ## OPERATION Segmentation
1485
1486  ## Segment O1
1487  **Content**: [exact segment content]
1488  **Type**: [segment type]
1489  **Summary**: [brief summary]
1490  **Complexity**: [Low/Medium/High]
1491
1492  # Segmentation Summary
1493  [
1494  {
1495    "segment_id": "G1",
1496    "abstraction_level": "GOAL",
1497    "type": "objective_statement",
1498    "summary": "find optimal solution",
1499    "complexity": "Medium"
1500  }
1501 ]

```

**Step 2: Example Output.** Below is the example output for the second step.

```

1499  # Segmentation Analysis
1500
1501  ## GOAL Segmentation
1502
1503  ### Segment G1
1504  **Content**:
1505  A construction company was building a tunnel. When  $\tfrac{1}{3}$  of the tunnel was completed at the original speed, they started using new equipment, which increased the construction speed by 20% and reduced the working hours to 80% of the original. As a result, it took a total of 185 days to complete the tunnel. If they had not used the new equipment and continued at the original speed, it would have taken  $185 \times 3 = 555$  days to complete the tunnel.
1506
1507
1508
1509
1510
1511  **Type**: objective_statement
1512  **Summary**: find missing days

```

```

1512  **Complexity**: Medium
1513
1514  ---
1515
1516  ## STRATEGY Segmentation
1517
1518  ### Segment S1
1519  **Content**:
1520  We have a tunnel. Original speed = some rate (units per day). They
1521  ↪ built  $\frac{1}{3}$  of tunnel at original speed. Then they switched
1522  ↪ to new equipment: speed increased by  $20\%$ , i.e., new speed  $= 1.2$ 
1523  ↪  $\times$  original speed. Working hours reduced to  $80\%$  of original,
1524  ↪ meaning per day they work fewer hours.
1525
1526  If original daily production =  $v \cdot h$  (speed per hour  $\times$  hours per day), then new daily production =
1527  ↪  $(1.2v)(0.8h) = 0.96vh = 0.96 \times$  original daily output.
1528
1529  So overall daily rate actually decreased by  $4\%$ . This explains why
1530  ↪ the total time (185 days) was longer than the original time would
1531  ↪ have been.
1532
1533  **Type**: approach_description
1534  **Summary**: model speed & hours
1535  **Complexity**: Medium
1536
1537  ---
1538
1539  ## TACTIC Segmentation
1540
1541  ### Segment T1
1542  **Content**:
1543  Let total tunnel length =  $1$ .
1544  Let original daily output =  $D$ . Then original completion time =  $1/D$ .
1545
1546  - First  $\frac{1}{3}$  of tunnel:
1547   $t_1 = \frac{1}{3D}$ .
1548
1549  - Remaining  $\frac{2}{3}$  at reduced daily rate  $0.96D$ :
1550   $t_2 = \frac{2}{3} \cdot 0.96D = \frac{0.694444\ldots}{D}$ .
1551
1552  Total time:
1553  
$$t_1 + t_2 = \frac{1}{3D} + \frac{0.694444\ldots}{D} = 185$$

1554
1555  So
1556  
$$\frac{1.027777\ldots}{D} = 185 \quad \Rightarrow \quad D = \frac{1.027777\ldots}{185} = \frac{1}{180}$$

1557
1558  Thus original schedule =  $180$  days.
1559
1560  **Type**: method_definition
1561  **Summary**: define variables, solve
1562  **Complexity**: Medium
1563
1564  ---
1565
1566  ## STEP Segmentation
1567
1568  ### Segment ST1
1569  **Content**:
1570  1. Compute new daily output:  $0.96 \times D$ .

```

```

1566
1567 **Type**: inference
1568 **Summary**: compute reduced output
1569 **Complexity**: Low
1570
1571 ### Segment ST2
1572 **Content**:
1573 2. Time for first third:  $t_1 = \frac{1}{3}D$ .
1574
1575 **Type**: inference
1576 **Summary**: first-third time
1577 **Complexity**: Low
1578
1579 ### Segment ST3
1580 **Content**:
1581 3. Time for remaining two-thirds:
1582  $\rightarrow t_2 = \frac{2}{3}D = \frac{0.96D}{1}$ .
1583
1584 **Type**: inference
1585 **Summary**: second-portion time
1586 **Complexity**: Low
1587
1588 ### Segment ST4
1589 **Content**:
1590 4. Total:  $t_1 + t_2 = 185$ .
1591
1592 **Type**: inference
1593 **Summary**: total time equation
1594 **Complexity**: Low
1595
1596 ### Segment ST5
1597 **Content**:
1598 5. Combine:  $\frac{1.027777\ldots}{D} = 185$ .
1599
1600 **Type**: algebraic_combination
1601 **Summary**: combine fractions
1602 **Complexity**: Low
1603
1604 ### Segment ST6
1605 **Content**:
1606 6. Solve:  $D = \frac{1}{180}$ .
1607
1608 **Type**: solution_step
1609 **Summary**: solve for D
1610 **Complexity**: Low
1611
1612 ---
1613
1614 ## OPERATION Segmentation
1615
1616 ### Segment O1
1617 **Content**:
1618  $\left[ \frac{2}{3} \cdot 0.96 \right] = \frac{2}{2.88} = 0.694444 \ldots$ 
1619
1620 **Type**: calculation

```

```

1620  **Summary**: compute $t\_2$ factor
1621  **Complexity**: Low
1622
1623  ### Segment 02
1624  **Content**:
1625  \[
1626  \frac{1}{3}+0.694444\ldots=1.027777\ldots
1627  \]
1628
1629  **Type**: calculation
1630  **Summary**: sum fractions
1631  **Complexity**: Low
1632
1633  ### Segment 03
1634  **Content**:
1635  \[
1636  D=\frac{1.027777\ldots}{185}=0.005555\ldots=\frac{1}{180}
1637  \]
1638
1639  **Type**: calculation
1640  **Summary**: find $D$ value
1641  **Complexity**: Low
1642
1643  ### Segment 04
1644  **Content**:
1645  \[
1646  \text{Original total time}=\frac{1}{D}=180\text{ days}
1647  \]
1648
1649  # Segmentation Summary
1650
1651  ```json
1652  [
1653    {"segment_id": "G1", "abstraction_level": "GOAL", "type": "objective_statement", "summary": "find missing days", "complexity": "Medium"}, {"segment_id": "S1", "abstraction_level": "STRATEGY", "type": "approach_description", "summary": "model speed & hours", "complexity": "Medium"}, {"segment_id": "T1", "abstraction_level": "TACTIC", "type": "method_definition", "summary": "define variables, solve", "complexity": "Medium"}, {"segment_id": "ST1", "abstraction_level": "STEP", "type": "inference", "summary": "compute new output", "complexity": "Low"}, {"segment_id": "ST2", "abstraction_level": "STEP", "type": "inference", "summary": "first-third time", "complexity": "Low"}, {"segment_id": "ST3", "abstraction_level": "STEP", "type": "inference", "summary": "remaining time", "complexity": "Low"}, {"segment_id": "ST4", "abstraction_level": "STEP", "type": "inference", "summary": "total equals 185", "complexity": "Low"}, {"segment_id": "ST5", "abstraction_level": "STEP", "type": "algebraic_combination", "summary": "combine fractions", "complexity": "Low"}, {"segment_id": "ST6", "abstraction_level": "STEP", "type": "solution_step", "summary": "solve for D", "complexity": "Low"}, {"segment_id": "ST7", "abstraction_level": "STEP", "type": "conclusion_inference", "summary": "original time 180", "complexity": "Low"}, ]

```

```

1674 {"segment_id": "O1", "abstraction_level": "OPERATION", "type": "calculation", "summary": "compute $t_2$ factor", "complexity": "Low"},  

1675 {"segment_id": "O2", "abstraction_level": "OPERATION", "type": "calculation", "summary": "sum fractions", "complexity": "Low"},  

1676 {"segment_id": "O3", "abstraction_level": "OPERATION", "type": "calculation", "summary": "determine D", "complexity": "Low"},  

1677 {"segment_id": "O4", "abstraction_level": "OPERATION", "type": "calculation", "summary": "original total time", "complexity": "Low"}  

1678 ]  

1679  

1680  

1681  

1682  

1683

```

### 1684 C.1.3 STEP 3: EXTRACT RELATIONSHIPS.

1685 Below are the instruction of how to extract relationships.

```

1686 You are a semantic relationship analysis agent. Your task is to  

1687 identify meaningful relationships between reasoning segments to  

1688 construct a hierarchical DAG.  

1689  

1690 #### 1. Relationship Types:  

1691 - **decomposes_to**: Higher abstraction level breaks down into lower  

1692   ↪ level  

1693 - **depends_on**: Logical prerequisite (A must be true for B to be  

1694   ↪ valid)  

1695 - **enables**: A makes B possible or provides foundation for B  

1696 - **validates**: A provides verification, proof, or confirmation for B  

1697 - **exemplifies**: A provides concrete example or instance of B  

1698 - **refines**: A adds precision, detail, or specification to B  

1699 - **contradicts**: A contradicts, corrects, or replaces B  

1700  

1701 #### 2. Analysis Priorities:  

1702 - **Hierarchical relationships first**: Focus on decomposes_to between  

1703   ↪ abstraction levels  

1704 - **Lateral relationships second**: Identify depends_on, enables within  

1705   ↪ same level  

1706 - **Support relationships third**: Find validates, exemplifies, refines  

1707   ↪ connections  

1708 - **Be selective**: Only include meaningful, clear relationships  

1709  

1710 #### 3. Relationship Strength:  

1711 - **Strong**: Clear, direct semantic connection  

1712 - **Medium**: Implied or indirect connection  

1713 - **Weak**: Tenuous or unclear connection  

1714  

1715 #### Input Format:  

1716 # Segmentation Analysis  

1717 [output from Step 2]  

1718  

1719 #### Output Format:  

1720 # Relationship Analysis  

1721  

1722 ## Hierarchical Relationships (decomposes_to)  

1723 - G1 -> S1: [explanation]  

1724 - S1 -> T1: [explanation]  

1725 - T1 -> ST1: [explanation]  

1726 - ST1 -> O1: [explanation]  

1727  

1728 ## Lateral Dependencies (depends_on, enables)  

1729 - T1 -> T2: [explanation]  

1730 - ST1 -> ST2: [explanation]  

1731  

1732 ## Support Relationships (validates, exemplifies, refines)

```

```

1728 - O1 exemplifies ST1: [explanation]
1729 - T3 validates S1: [explanation]
1730
1731 # Relationship Summary
1732 [
1733 {
1734   "from_segment": "G1",
1735   "to_segment": "S1",
1736   "relationship_type": "decomposes_to",
1737   "strength": "strong",
1738   "explanation": "Goal G1 breaks down into strategy S1"
1739 }
1740 ]
1741
1742
1743 Step 3: Example Output. Below is the example output for the second step.

```

```

1744 # Relationship Analysis
1745
1746 ## Hierarchical Relationships (decomposes_to)
1747
1748 - **G1 ->S1** - The overall goal (find the missing number of days) is
1749   ↪ broken down into a strategic modelling approach. **[strong]**
1750 - **S1 ->T1** - The chosen strategy (model speed & hours) is further
1751   ↪ refined into a concrete tactic (define variables and solve).
1752   ↪ **[strong]**
1753 - **T1 ->ST1** - The tactic is implemented as the first low-level
1754   ↪ inference step (compute the reduced daily output). **[strong]**
1755 - **T1 ->ST2** - The tactic also yields the second inference step (time
1756   ↪ for the first third). **[strong]**
1757 - **T1 ->ST3** - The tactic produces the third inference step (time for
1758   ↪ the remaining two-thirds). **[strong]**
1759 - **T1 ->ST4** - The tactic yields the fourth inference step
1760   ↪ (total-time equation). **[strong]**
1761 - **T1 ->ST5** - The tactic yields the fifth inference step (combine
1762   ↪ fractions). **[strong]**
1763 - **T1 ->ST6** - The tactic yields the sixth inference step (solve for
1764   ↪ \(\mathbf{D}\)). **[strong]**
1765 - **T1 ->ST7** - The tactic yields the seventh inference step (conclude
1766   ↪ original schedule). **[strong]**
1767 - **ST1 ->O1** - The operation that computes
1768   ↪ \(\frac{2}{3} \cdot 0.96 = 0.694444\ldots\) exemplifies the inference
1769   ↪ "compute reduced daily output". **[strong]**
1770 - **ST2 ->O2** - The operation that adds \(\frac{1}{3} + 0.694444\ldots\)
1771   ↪ exemplifies the inference "first-third time". **[strong]**
1772 - **ST3 ->O1** - The same calculation in O1 also provides the numeric
1773   ↪ factor used in the inference for the remaining two-thirds.
1774   ↪ **[strong]**
1775 - **ST4 ->O2** - The sum performed in O2 validates the total-time
1776   ↪ equation. **[strong]**
1777 - **ST5 ->O2** - The combination of fractions in O2 supports the
1778   ↪ algebraic-combination step. **[medium]**
1779 - **ST6 ->O3** - The calculation in O3 (finding \(\mathbf{D}\)) exemplifies the
1780   ↪ solution step. **[strong]**
1781 - **ST7 ->O4** - The final computation in O4 (original total time)
1782   ↪ exemplifies the conclusion inference. **[strong]**
1783
1784 ---
1785
1786 ## Lateral Dependencies (depends_on, enables)
1787
1788 - **ST3 depends_on ST1** - Computing the remaining two-thirds requires
1789   ↪ the reduced daily output from ST1. **[strong]**

```

```

1782 - **ST4 depends_on ST2** - Total time uses the time for the first
1783   ↪ third. **[strong]**
1784 - **ST4 depends_on ST3** - Total time also uses the time for the
1785   ↪ remaining two-thirds. **[strong]**
1786 - **ST5 depends_on ST4** - Combining fractions presupposes the
1787   ↪ total-time equation. **[strong]**
1788 - **ST6 depends_on ST5** - Solving for  $\backslash(D\backslash)$  needs the
1789   ↪ combined-fraction expression. **[strong]**
1790 - **ST7 depends_on ST6** - The original schedule is the reciprocal of
1791   ↪ the solved  $\backslash(D\backslash)$ . **[strong]**
1792 - **O1 enables ST3** - The numeric factor produced by O1 makes the
1793   ↪ inference in ST3 possible. **[strong]**
1794 - **O2 enables ST4** - The summation in O2 provides the numeric total
1795   ↪ needed for ST4. **[strong]**
1796 - **O3 enables ST6** - The calculation of  $\backslash(D\backslash)$  in O3 supplies the
1797   ↪ value needed for the solution step. **[strong]**
1798 - **O4 enables ST7** - The final division in O4 yields the original
1799   ↪ total-time conclusion. **[strong]**

1800 ---
1801 ## Support Relationships (validates, exemplifies, refines)
1802 - **O1 exemplifies ST1** - Shows concretely how the reduced output
1803   ↪ factor is obtained. **[strong]**
1804 - **O1 exemplifies ST3** - Provides the numeric factor used in the
1805   ↪ remaining-time inference. **[strong]**
1806 - **O2 exemplifies ST2** - Demonstrates the addition that gives the
1807   ↪ first-third time fraction. **[strong]**
1808 - **O2 exemplifies ST4** - Supplies the numeric sum that validates the
1809   ↪ total-time equation. **[strong]**
1810 - **O3 exemplifies ST6** - Calculates  $\backslash(D\backslash)$ , confirming the algebraic
1811   ↪ solution. **[strong]**
1812 - **O4 exemplifies ST7** - Computes the reciprocal, confirming the
1813   ↪ original schedule. **[strong]**
1814 - **O1 validates ST3** - The numeric result matches the expression
1815   ↪ derived in ST3. **[strong]**
1816 - **O2 validates ST4** - The summed value satisfies the total-time
1817   ↪ equality. **[strong]**
1818 - **O3 validates ST6** - The derived  $\backslash(D\backslash)$  satisfies the
1819   ↪ combined-fraction equation. **[strong]**
1820 - **O4 validates ST7** - The final division yields the expected 180-day
1821   ↪ schedule. **[strong]**
1822 - **T1 refines S1** - The tactic adds concrete variable definitions and
1823   ↪ algebraic steps to the broader modelling strategy. **[medium]**
1824 - **S1 refines G1** - The strategic description narrows the vague goal
1825   ↪ into a measurable modelling problem. **[medium]**

1826 ---
1827 # Relationship Summary
1828 ```json
1829 [
1830   {
1831     "from_segment": "G1",
1832     "to_segment": "S1",
1833     "relationship_type": "decomposes_to",
1834     "strength": "strong",
1835     "explanation": "Goal G1 (find missing days) is broken down into the
1836       ↪ modelling strategy S1."
1837   },
1838   {
1839     "from_segment": "S1",
1840     "to_segment": "T1",
1841     "relationship_type": "decomposes_to",
1842   }
1843 ]

```

```

1836     "strength": "strong",
1837     "explanation": "Strategy S1 (model speed & hours) is refined into
1838     ↪ the concrete tactic T1 (define variables and solve)."
1839   },
1840   {
1841     "from_segment": "T1",
1842     "to_segment": "ST1",
1843     "relationship_type": "decomposes_to",
1844     "strength": "strong",
1845     "explanation": "Tactic T1 is implemented as the first inference step
1846     ↪ ST1 (compute reduced daily output)."
1847   },
1848   {
1849     "from_segment": "ST1",
1850     "to_segment": "O1",
1851     "relationship_type": "decomposes_to",
1852     "strength": "strong",
1853     "explanation": "Step ST1 is realised by operation O1, which
1854     ↪ calculates the factor 0.694444...""
1855   },
1856   {
1857     "from_segment": "ST3",
1858     "to_segment": "O1",
1859     "relationship_type": "exemplifies",
1860     "strength": "strong",
1861     "explanation": "Operation O1 provides the numeric factor used in
1862     ↪ step ST3 (time for remaining two-thirds)."
1863   },
1864   {
1865     "from_segment": "ST4",
1866     "to_segment": "O2",
1867     "relationship_type": "exemplifies",
1868     "strength": "strong",
1869     "explanation": "Operation O2 sums the fractions, exemplifying the
1870     ↪ total-time equation in ST4."
1871   },
1872   {
1873     "from_segment": "ST6",
1874     "to_segment": "O3",
1875     "relationship_type": "exemplifies",
1876     "strength": "strong",
1877     "explanation": "Operation O3 computes D, exemplifying the solution
1878     ↪ step ST6."
1879   },
1880   {
1881     "from_segment": "ST7",
1882     "to_segment": "O4",
1883     "relationship_type": "exemplifies",
1884     "strength": "strong",
1885     "explanation": "Operation O4 calculates the original total time,
1886     ↪ exemplifying conclusion step ST7."
1887   },
1888   {
1889     "from_segment": "ST3",
1890     "to_segment": "ST1",
1891     "relationship_type": "depends_on",
1892     "strength": "strong",
1893     "explanation": "The remaining-time calculation in ST3 requires the
1894     ↪ reduced daily output derived in ST1."
1895   },
1896   {
1897     "from_segment": "ST4",
1898     "to_segment": "ST2",
1899     "relationship_type": "depends_on",
1900     "strength": "strong",
1901   }

```

```

1890
1891     "explanation": "Total time in ST4 uses the first-third time computed
1892     ↪ in ST2."
1893   },
1894   {
1895     "from_segment": "ST4",
1896     "to_segment": "ST3",
1897     "relationship_type": "depends_on",
1898     "strength": "strong",
1899     "explanation": "Total time in ST4 also uses the remaining-time
1900     ↪ computed in ST3."
1901   },
1902   {
1903     "from_segment": "ST5",
1904     "to_segment": "ST4",
1905     "relationship_type": "depends_on",
1906     "strength": "strong",
1907     "explanation": "Combining fractions in ST5 presupposes the
1908     ↪ total-time equation from ST4."
1909   },
1910   {
1911     "from_segment": "ST6",
1912     "to_segment": "ST5",
1913     "relationship_type": "depends_on",
1914     "strength": "strong",
1915     "explanation": "Solving for D in ST6 requires the combined fraction
1916     ↪ expression from ST5."
1917   },
1918   {
1919     "from_segment": "ST7",
1920     "to_segment": "ST6",
1921     "relationship_type": "depends_on",
1922     "strength": "strong",
1923     "explanation": "The original schedule in ST7 is obtained by taking
1924     ↪ the reciprocal of D solved in ST6."
1925   },
1926   {
1927     "from_segment": "T1",
1928     "to_segment": "S1",
1929     "relationship_type": "refines",
1930     "strength": "medium",
1931     "explanation": "Tactic T1 adds concrete variable definitions to the
1932     ↪ broader modelling strategy S1."
1933   }
1934 ]
1935
1936
1937
1938
1939
1940
1941
1942
1943

```

#### C.1.4 STEP 4: ASSIGN AUDITOR TYPES.

Below are the instruction of how to assign auditor types.

You are an auditor assignment agent. Your task is to assign appropriate  
 ↪ auditor types to each reasoning segment based on the TRUST  
 ↪ framework's three-tier verification system.

### 1. Auditor Types:

- **T\_Auto (Automated)**: Deterministic verification, formal logic,  
 ↪ mathematical proofs
- **T\_LLM (LLM-based)**: Semantic coherence, factual accuracy, domain  
 ↪ reasoning
- **T\_Human (Human)**: Complex judgment, ethical considerations, domain  
 ↪ expertise

### 2. Assignment Criteria:

```

1944    **T_Auto Assignment:**
1945    - Mathematical calculations and proofs
1946    - Logical consistency checks
1947    - Formal verification tasks
1948    - Deterministic algorithmic steps
1949    - Syntax and format validation

1950    **T_LLM Assignment:**
1951    - Semantic coherence evaluation
1952    - Factual accuracy checking
1953    - Domain-specific reasoning patterns
1954    - Conceptual understanding verification
1955    - Strategy and approach assessment

1956    **T_Human Assignment:**
1957    - Complex ethical considerations
1958    - Domain expertise requirements
1959    - Ambiguous interpretation cases
1960    - Creative or novel reasoning
1961    - High-stakes decision validation

1962    ### 3. Multi-Auditor Segments:
1963    - Some segments may require multiple auditor types
1964    - Primary auditor handles main verification
1965    - Secondary auditors provide additional validation
1966    - Specify auditor priority and responsibilities

1967    ### Input Format:
1968    # Segmentation Analysis
1969    [segments from Step 2]

1970    # Relationship Analysis
1971    [relationships from Step 3]

1972    ### Output Format:
1973    # Auditor Assignment Analysis

1974    ## Segment Auditor Assignments

1975    ### Segment G1
1976    **Primary Auditor**: T_Human
1977    **Secondary Auditors**: [T_LLM]
1978    **Justification**: Problem interpretation requires domain expertise
1979    **Verification Focus**: Objective clarity, scope appropriateness

1980    ### Segment S1
1981    **Primary Auditor**: T_LLM
1982    **Secondary Auditors**: [T_Human]
1983    **Justification**: Strategic approach evaluation needs semantic
1984    ↘ understanding
1985    **Verification Focus**: Approach validity, methodology soundness

1986    ## Auditor Type Distribution
1987    - **T_Auto**: [list of segment IDs]
1988    - **T_LLM**: [list of segment IDs]
1989    - **T_Human**: [list of segment IDs]

1990    ## Multi-Auditor Segments
1991    [
1992    {
1993        "segment_id": "G1",
1994        "primary_auditor": "T_Human",
1995        "secondary_auditors": ["T_LLM"],
1996        "verification_complexity": "High",
1997        "justification": "Domain expertise required"

```

1998  
1999  
2000  
2001  
2002  
2003

}

]

**Step 4: Example Output.** Below are the instruction of how to Assign Auditor Types.

```

2004 # Auditor Assignment Analysis
2005
2006 ## Segment Auditor Assignments
2007
2008 ### Segment G1 - Goal (objective_statement)
2009 - **Primary Auditor**: **T_LLM**
2010 - **Secondary Auditors**: **[T_Auto, T_Human]**
2011 - **Justification**: Interpreting the wording of the word-problem and
2012   ↪ ensuring the goal ("find the missing number of days") is correctly
2013   ↪ understood
2014   requires semantic comprehension (T_LLM). A deterministic check that the
2015   ↪ interpreted goal matches the expected quantitative target is
2016   ↪ performed by
2017   T_Auto, while a human reviewer can confirm that the problem statement
2018   ↪ has been captured without ambiguity.
2019 - **Verification Focus**: Goal clarity, scope appropriateness, correct
2020   ↪ identification of the unknown variable.

2021 ---
2022
2023 ### Segment S1 - Strategy (approach_description)
2024 - **Primary Auditor**: **T_LLM**
2025 - **Secondary Auditors**: **[T_Auto, T_Human]**
2026 - **Justification**: The proposed modelling approach (relating speed,
2027   ↪ hours, and daily output) must be semantically coherent and logically
2028   plausible - a task suited to T_LLM. Logical consistency of the derived
2029   ↪ relationships is then verified by T_Auto, and a human can confirm
2030   ↪ that the
2031   chosen strategy aligns with typical problem-solving conventions in
2032   ↪ elementary algebra.
2033 - **Verification Focus**: Validity of the modelling assumptions,
2034   ↪ correctness of the conceptual link between speed increase and hour
2035   ↪ reduction.

2036 ---
2037
2038 ### Segment T1 - Tactic (method_definition)
2039 - **Primary Auditor**: **T_Auto**
2040 - **Secondary Auditors**: **[T_LLM]**
2041 - **Justification**: This segment defines variables and sets up
2042   ↪ algebraic equations - a deterministic, mathematically rigorous task
2043   ↪ best handled by
2044   automated verification. T_LLM reviews the narrative framing of the
2045   ↪ variables to ensure they are semantically consistent with the
2046   ↪ problem context.

2047 - **Verification Focus**: Correct variable definition, proper
2048   ↪ formulation of equations.

2049 ---
2050
2051 ### Segment ST1 - STEP (inference)
2052 - **Primary Auditor**: **T_Auto**
2053 - **Secondary Auditors**: **[T_LLM]**
2054 - **Justification**: Computes the reduced daily output (pure
2055   ↪ arithmetic). A secondary LLM check ensures the inference wording
2056   ↪ correctly reflects
2057   the computation performed.

```

```

2052
2053     ### Segment ST2 - STEP (inference)
2054     - **Primary Auditor**: **T_Auto**
2055     - **Secondary Auditors**: **[T_LLM]**
2056
2057     ### Segment ST3 - STEP (inference)
2058     - **Primary Auditor**: **T_Auto**
2059     - **Secondary Auditors**: **[T_LLM]**
2060
2061     ### Segment ST4 - STEP (inference)
2062     - **Primary Auditor**: **T_Auto**
2063     - **Secondary Auditors**: **[T_LLM]**
2064
2065     ### Segment ST5 - STEP (algebraic_combination)
2066     - **Primary Auditor**: **T_Auto**
2067     - **Secondary Auditors**: **[T_LLM]**
2068
2069     ### Segment ST6 - STEP (solution_step)
2070     - **Primary Auditor**: **T_Auto**
2071     - **Secondary Auditors**: **[T_LLM]**
2072
2073     ---
2074
2075     ### Segment O1 - OPERATION (calculation)
2076     - **Primary Auditor**: **T_Auto**
2077     - **Secondary Auditors**: **[T_LLM]**
2078
2079     ### Segment O2 - OPERATION (calculation)
2080     - **Primary Auditor**: **T_Auto**
2081     - **Secondary Auditors**: **[T_LLM]**
2082
2083     ### Segment O3 - OPERATION (calculation)
2084     - **Primary Auditor**: **T_Auto**
2085     - **Secondary Auditors**: **[T_LLM]**
2086
2087     ---
2088
2089
2090     ## Auditor Type Distribution
2091
2092     | Auditor Type | Segment IDs |
2093     | ----- | ----- |
2094     | **T_Auto** | T1, ST1, ST2, ST3, ST4, ST5, ST6, ST7, O1, O2, O3, O4 |
2095     | **T_LLM** | G1, S1, T1, ST1-ST7, O1-O4 |
2096     | **T_Human** | G1 (secondary) |
2097
2098     *Note*: Although most low-level steps are purely arithmetic, a
2099     ↪ secondary LLM audit is included to verify that the textual
2100     ↪ description of each step
2101     faithfully mirrors the underlying calculation. A human audit is only
2102     ↪ required for the top-level goal interpretation (G1) to guard against
2103     ↪ any
2104     subtle mis-framing of the problem.
2105
2106     ---
2107
2108     ## Multi-Auditor Segments

```

```

2106 ````json
2107 [
2108 {
2109   "segment_id": "G1",
2110   "primary_auditor": "T_LLM",
2111   "secondary_auditors": ["T_Auto", "T_Human"],
2112   "verification_complexity": "High",
2113   "justification": "Requires correct semantic interpretation of the
2114   ↪ word problem and identification of the unknown; deterministic
2115   ↪ check of the
2116   derived target and optional human confirmation for ambiguity."
2117 },
2118 {
2119   "segment_id": "S1",
2120   "primary_auditor": "T_LLM",
2121   "secondary_auditors": ["T_Auto", "T_Human"],
2122   "verification_complexity": "Medium",
2123   "justification": "Semantic coherence of the modelling approach must
2124   ↪ be evaluated; logical consistency is verified automatically; human
2125   oversight ensures alignment with standard problem-solving practice."
2126 },
2127 {
2128   "segment_id": "T1",
2129   "primary_auditor": "T_Auto",
2130   "secondary_auditors": ["T_LLM"],
2131   "verification_complexity": "Low",
2132   "justification": "Purely algebraic definition; LLM checks that
2133   ↪ variable naming matches the narrative."
2134 },
2135 {
2136   "segment_id": "ST1",
2137   "primary_auditor": "T_Auto",
2138   "secondary_auditors": ["T_LLM"],
2139   "verification_complexity": "Low",
2140   "justification": "Arithmetic computation; LLM confirms textual
2141   ↪ inference aligns with calculation."
2142 },
2143 {
2144   "segment_id": "ST2",
2145   "primary_auditor": "T_Auto",
2146   "secondary_auditors": ["T_LLM"],
2147   "verification_complexity": "Low",
2148   "justification": "Arithmetic; LLM validates description."
2149 },
2150 {
2151   "segment_id": "ST3",
2152   "primary_auditor": "T_Auto",
2153   "secondary_auditors": ["T_LLM"],
2154   "verification_complexity": "Low",
2155   "justification": "Arithmetic; LLM validates description."
2156 },
2157 {
2158   "segment_id": "ST4",
2159   "primary_auditor": "T_Auto",
2160   "secondary_auditors": ["T_LLM"],
2161   "verification_complexity": "Low",
2162   "justification": "Arithmetic; LLM validates description."
2163 },
2164 {
2165   "segment_id": "ST5",
2166   "primary_auditor": "T_Auto",
2167   "secondary_auditors": ["T_LLM"],
2168   "verification_complexity": "Low",
2169   "justification": "Arithmetic; LLM validates description."
2170 }

```

```

2160
2161  {
2162      "segment_id": "ST6",
2163      "primary_auditor": "T_Auto",
2164      "secondary_auditors": ["T_LLM"],
2165      "verification_complexity": "Low",
2166      "justification": "Arithmetic; LLM validates description."
2167  },
2168  {
2169      "segment_id": "ST7",
2170      "primary_auditor": "T_Auto",
2171      "secondary_auditors": ["T_LLM"],
2172      "verification_complexity": "Low",
2173      "justification": "Arithmetic; LLM validates description."
2174  },
2175  {
2176      "segment_id": "O1",
2177      "primary_auditor": "T_Auto",
2178      "secondary_auditors": ["T_LLM"],
2179      "verification_complexity": "Low",
2180      "justification": "Straight-forward numeric calculation."
2181  },
2182  {
2183      "segment_id": "O2",
2184      "primary_auditor": "T_Auto",
2185      "secondary_auditors": ["T_LLM"],
2186      "verification_complexity": "Low",
2187      "justification": "Straight-forward numeric calculation."
2188  },
2189  {
2190      "segment_id": "O3",
2191      "primary_auditor": "T_Auto",
2192      "secondary_auditors": ["T_LLM"],
2193      "verification_complexity": "Low",
2194      "justification": "Straight-forward numeric calculation."
2195  },
2196  {
2197      "segment_id": "O4",
2198      "primary_auditor": "T_Auto",
2199      "secondary_auditors": ["T_LLM"],
2200      "verification_complexity": "Low",
2201      "justification": "Straight-forward numeric calculation."
2202  }
2203  ]
2204  ...
2205

```

### 2199 C.1.5 STEP 5: REFINE AND HDAG CONSTRUCTION.

2200 Below are the instruction of how to refine and HDAG Construction.

2203 You are an HDAG construction agent. Your task is to synthesize all  
 2204 ↳ previous analyses into a complete Hierarchical Directed Acyclic  
 2205 ↳ Graph with finalized segments, relationships, and auditor  
 2206 ↳ assignments.

2207 ##### 1. HDAG Refinement Tasks:  
 2208 - \*\*Validate DAG Structure\*\*: Ensure no cycles exist in relationships  
 2209 - \*\*Optimize Segment Granularity\*\*: Merge over-segmented or split  
 2210 ↳ under-segmented content  
 2211 - \*\*Strengthen Relationship Confidence\*\*: Refine relationship types and  
 2212 ↳ strengths  
 2213 - \*\*Balance Auditor Load\*\*: Ensure reasonable distribution across  
 2214 ↳ auditor types  
 - \*\*Quality Assurance\*\*: Verify completeness and coherence

```

2214
2215     ### 2. Final HDAG Components:
2216     - **Nodes**: Finalized reasoning segments with metadata
2217     - **Edges**: Validated relationships with confidence scores
2218     - **Auditor Mapping**: Optimized auditor assignments
2219     - **Verification Flow**: Clear audit execution pathway
2220
2221     ### 3. HDAG Validation Checks:
2222     - **Acyclicity**: No circular dependencies
2223     - **Connectivity**: All segments appropriately connected
2224     - **Completeness**: All original reasoning preserved
2225     - **Auditability**: Each segment has clear verification criteria
2226
2227     ### Input Format:
2228     # Segmentation Analysis
2229     [from Step 2]
2230
2231     # Relationship Analysis
2232     [from Step 3]
2233
2234     # Auditor Assignment Analysis
2235     [from Step 4]
2236
2237     ### Output Format:
2238     # Final HDAG Construction
2239
2240     ## HDAG Structure Summary
2241     - **Total Nodes**: [count]
2242     - **Total Edges**: [count]
2243     - **Abstraction Levels**: [levels present]
2244     - **Auditor Distribution**: [T_Auto: X, T_LLM: Y, T_Human: Z]
2245
2246     ## Finalized Nodes
2247
2248     ### Node G1
2249     **Content**: [final segment content]
2250     **Abstraction Level**: GOAL
2251     **Type**: objective_statement
2252     **Summary**: [3-5 words]
2253     **Primary Auditor**: T_Human
2254     **Secondary Auditors**: [T_LLM]
2255     **Verification Complexity**: High
2256     **Dependencies**: []
2257     **Enables**: [S1, S2]
2258
2259     ## Finalized Edges
2260     [
2261         {
2262             "from": "G1",
2263             "to": "S1",
2264             "type": "decomposes_to",
2265             "strength": "strong",
2266             "confidence": 0.95
2267         }
2268     ]
2269
2270     ## Verification Flow
2271     1. **Entry Points**: [root nodes for audit initiation]
2272     2. **Critical Path**: [essential verification sequence]
2273     3. **Parallel Tracks**: [independent verification branches]
2274     4. **Exit Points**: [final validation nodes]
2275
2276     ## HDAG Validation Report
2277     - **Acyclicity Check**: No cycles detected
2278     - **Connectivity Check**: All segments connected

```

```

2268 - Completeness Check: All content preserved
2269 - Auditability Check: Clear verification criteria
2270
2271 ## Complete HDAG JSON
2272 {
2273   "nodes": [
2274     {
2275       "id": "G1",
2276       "content": "[exact content]",
2277       "abstraction_level": "GOAL",
2278       "type": "objective_statement",
2279       "summary": "[summary]",
2280       "primary_auditor": "T_Human",
2281       "secondary_auditors": ["T_LLM"],
2282       "complexity": "High"
2283     }
2284   ],
2285   "edges": [
2286     {
2287       "from": "G1",
2288       "to": "S1",
2289       "relationship": "decomposes_to",
2290       "strength": "strong",
2291       "confidence": 0.95
2292     }
2293   ],
2294   "metadata": {
2295     "total_nodes": 15,
2296     "total_edges": 23,
2297     "auditor_distribution": {
2298       "T_Auto": 5,
2299       "T_LLM": 7,
2300       "T_Human": 3
2301     }
2302   }
2303 }

```

**Step 5: Example Output.** Below are the instruction of how to refine and the construction of HDAG.

```

2304 # Final HDAG Construction
2305
2306 ## HDAG Structure Summary
2307 - Total Nodes: 14
2308 - Total Edges: 28
2309 - Abstraction Levels: GOAL, STRATEGY, TACTIC, STEP, OPERATION
2310 - Auditor Distribution (primary auditors): T_Auto:12, T_LLM:1,
2311   ↳ T_Human:1
2312
2313 ## Finalized Nodes
2314
2315 ## Node G1
2316 Content:
2317 *A construction company was building a tunnel. When 1/3 of the tunnel
2318   ↳ was completed at the original speed, they started using new
2319   ↳ equipment, which
2320     increased the construction speed by 20 % and reduced the working hours
2321     ↳ to 80 % of the original. As a result, it took a total of 185 days to
2322       complete the tunnel. If they had not used the new equipment and
2323         ↳ continued at the original speed, it would have taken _____ days to
2324           ↳ complete the

```

```

2322 tunnel.*  

2323  **Abstraction Level**: GOAL  

2324  **Type**: objective_statement  

2325  **Summary**: find missing days  

2326  **Primary Auditor**: T_Human  

2327  **Secondary Auditors**: [T_LLM]  

2328  **Verification Complexity**: High  

2329  **Dependencies**: []  

2330  **Enables**: [S1]  

2331  

2332  #### Node S1  

2333  **Content**:  

2334  *Model the problem by relating speed, working hours, and daily output.  

2335  ↪ The original daily output is  $\langle D \rangle$ . After the equipment change the  

2336  ↪ speed  

2337  rises 20 % while hours fall to 80 % ->new daily output  $\langle 0.96D \rangle$ .*  

2338  **Abstraction Level**: STRATEGY  

2339  **Type**: approach_description  

2340  **Summary**: model speed & hours  

2341  **Primary Auditor**: T_LLM  

2342  **Secondary Auditors**: [T_Auto]  

2343  **Verification Complexity**: Medium  

2344  **Dependencies**: [G1]  

2345  **Enables**: [T1]  

2346  

2347  #### Node T1  

2348  **Content**:  

2349  *Let the tunnel length be 1 unit and the original daily output be  

2350  ↪  $\langle D \rangle$ . Build 1/3 at rate  $\langle D \rangle$ ; the remaining 2/3 at rate  $\langle 0.96D \rangle$ .  

2351  ↪ Set up the  

2352  time equation and solve for  $\langle D \rangle$ .*  

2353  **Abstraction Level**: TACTIC  

2354  **Type**: method_definition  

2355  **Summary**: define variables, solve  

2356  **Primary Auditor**: T_Auto  

2357  **Secondary Auditors**: [T_LLM]  

2358  **Verification Complexity**: Low  

2359  **Dependencies**: [S1]  

2360  **Enables**: [ST1, ST2, ST3, ST4, ST5, ST6, ST7]  

2361  

2362  #### Node ST1  

2363  **Content**: *Compute the reduced daily output:  $\langle 0.96D \rangle$ .*  

2364  **Abstraction Level**: STEP  

2365  **Type**: inference  

2366  **Summary**: compute reduced output  

2367  **Primary Auditor**: T_Auto  

2368  **Secondary Auditors**: [T_LLM]  

2369  **Verification Complexity**: Low  

2370  **Dependencies**: []  

2371  **Enables**: [O1]  

2372  

2373  #### Node ST2  

2374  **Content**: *Time for the first third:  $\langle t_1 = \frac{1}{3D} \rangle$ .*  

2375  **Abstraction Level**: STEP  

2376  **Type**: inference  

2377  **Summary**: first-third time  

2378  **Primary Auditor**: T_Auto  

2379  **Secondary Auditors**: [T_LLM]  

2380  **Verification Complexity**: Low  

2381  **Dependencies**: []  

2382  **Enables**: [O2]  

2383  

2384  #### Node ST3  

2385  **Content**: *Time for the remaining two-thirds:  $\langle t_2 = \frac{2}{3(0.96D)} = \frac{5}{8D} \rangle$ .*  

2386

```

```

2376  **Abstraction Level**: STEP
2377  **Type**: inference
2378  **Summary**: remaining-portion time
2379  **Primary Auditor**: T_Auto
2380  **Secondary Auditors**: [T_LLM]
2381  **Verification Complexity**: Low
2382  **Dependencies**: [ST1] *(ST3 depends_on ST1)*
2383  **Enables**: [O1]
2384
2385  ##### Node ST4
2386  **Content**: *Form the total-time equation:  $(t_1 + t_2 = 185)$ .*
2387  **Abstraction Level**: STEP
2388  **Type**: inference
2389  **Summary**: total-time equation
2390  **Primary Auditor**: T_Auto
2391  **Secondary Auditors**: [T_LLM]
2392  **Verification Complexity**: Low
2393  **Dependencies**: []
2394  **Enables**: [O2]
2395
2396  ##### Node ST5
2397  **Content**: *Combine the fractions to obtain a single algebraic
2398  ↪ expression for the time equation.*
2399  **Abstraction Level**: STEP
2400  **Type**: algebraic_combination
2401  **Summary**: combine algebraic terms
2402  **Primary Auditor**: T_Auto
2403  **Secondary Auditors**: [T_LLM]
2404  **Verification Complexity**: Low
2405  **Dependencies**: [ST4]
2406  **Enables**: [O2]
2407
2408  ##### Node ST6
2409  **Content**: *Solve the equation  $(\frac{1}{3D} + \frac{5}{8D} = 185)$ 
2410  ↪ for  $D$  ->  $D = \frac{1}{0.0054}$ .*
2411  **Abstraction Level**: STEP
2412  **Type**: solution_step
2413  **Summary**: solve for  $D$ 
2414  **Primary Auditor**: T_Auto
2415  **Secondary Auditors**: [T_LLM]
2416  **Verification Complexity**: Low
2417  **Dependencies**: [ST5]
2418  **Enables**: [O3]
2419
2420  ##### Node ST7
2421  **Content**: *Convert  $D$  to the required answer: original total time
2422  ↪  $(= \frac{1}{D})$  days ->  $\approx 365$  days.*
2423  **Abstraction Level**: STEP
2424  **Type**: conclusion_inference
2425  **Summary**: original-schedule result
2426  **Primary Auditor**: T_Auto
2427  **Secondary Auditors**: [T_LLM]
2428  **Verification Complexity**: Low
2429  **Dependencies**: [ST6]
2430  **Enables**: [O4]
2431
2432  ##### Node O1
2433  **Content**: *Numerical calculation  $(\frac{2}{3} \times 0.96 = 0.64)$  or
2434  ↪ directly  $(0.96)$  used in ST1.*
2435  **Abstraction Level**: OPERATION
2436  **Type**: calculation
2437  **Summary**: reduced-output calc
2438  **Primary Auditor**: T_Auto
2439  **Secondary Auditors**: [T_LLM]
2440  **Verification Complexity**: Low

```

```

2430  **Dependencies**: [ST1, ST3]
2431  **Enables**: []
2432
2433  ### Node O2
2434  **Content**: *Add the two time expressions:  $\frac{1}{3D} + \frac{5}{8D} = 185$ .*  

2435  **Abstraction Level**: OPERATION
2436  **Type**: calculation
2437  **Summary**: build time equation
2438  **Primary Auditor**: T_Auto
2439  **Secondary Auditors**: [T_LLM]
2440  **Verification Complexity**: Low
2441  **Dependencies**: [ST2, ST4, ST5]
2442  **Enables**: []
2443
2444  ### Node O3
2445  **Content**: *Solve  $D = 0.0054 \rightarrow D = 185 / (1/3 + 5/8) = 0.0054$ .*  

2446  **Abstraction Level**: OPERATION
2447  **Type**: calculation
2448  **Summary**: compute  $D$ 
2449  **Primary Auditor**: T_Auto
2450  **Secondary Auditors**: [T_LLM]
2451  **Verification Complexity**: Low
2452  **Dependencies**: [ST6]
2453  **Enables**: []
2454
2455  ### Node O4
2456  **Content**: *Take the reciprocal of  $D$  to obtain the original total  

2457  time  $(= \frac{1}{D} \approx 365)$  days.*  

2458  **Abstraction Level**: OPERATION
2459  **Type**: calculation
2460  **Summary**: final-answer calc
2461  **Primary Auditor**: T_Auto
2462  **Secondary Auditors**: [T_LLM]
2463  **Verification Complexity**: Low
2464  **Dependencies**: [ST7]
2465  **Enables**: []
2466
2467  ---
2468
2469  ## Finalized Edges
2470
2471  ```json
2472  [
2473    /* Decomposes-to hierarchy */
2474    {"from": "G1", "to": "S1", "type": "decomposes_to",
2475     "strength": "strong", "confidence": 0.95},
2476    {"from": "S1", "to": "T1", "type": "decomposes_to",
2477     "strength": "strong", "confidence": 0.94},
2478    {"from": "T1", "to": "ST1", "type": "decomposes_to",
2479     "strength": "strong", "confidence": 0.94},
2480    {"from": "T1", "to": "ST2", "type": "decomposes_to",
2481     "strength": "strong", "confidence": 0.94},
2482    {"from": "T1", "to": "ST3", "type": "decomposes_to",
2483     "strength": "strong", "confidence": 0.94},
2484    {"from": "T1", "to": "ST4", "type": "decomposes_to",
2485     "strength": "strong", "confidence": 0.94},
2486    {"from": "T1", "to": "ST5", "type": "decomposes_to",
2487     "strength": "strong", "confidence": 0.94},
2488    {"from": "T1", "to": "ST6", "type": "decomposes_to",
2489     "strength": "strong", "confidence": 0.94},
2490    {"from": "T1", "to": "ST7", "type": "decomposes_to",
2491     "strength": "strong", "confidence": 0.94},
2492    {"from": "ST1", "to": "O1", "type": "decomposes_to",
2493     "strength": "strong", "confidence": 0.93},

```

```

2484 {"from": "ST2", "to": "O2", "type": "decomposes_to",
2485 "strength": "strong", "confidence": 0.93},
2486 {"from": "ST3", "to": "O1", "type": "decomposes_to",
2487 "strength": "strong", "confidence": 0.93},
2488 {"from": "ST4", "to": "O2", "type": "decomposes_to",
2489 "strength": "strong", "confidence": 0.93},
2490 {"from": "ST5", "to": "O2", "type": "decomposes_to",
2491 "strength": "strong", "confidence": 0.93},
2492 {"from": "ST6", "to": "O3", "type": "decomposes_to",
2493 "strength": "strong", "confidence": 0.93},
2494 {"from": "ST7", "to": "O4", "type": "decomposes_to",
2495 "strength": "strong", "confidence": 0.93},  

2496 /* Lateral dependencies */
2497 {"from": "ST3", "to": "ST1", "type": "depends_on",
2498 "strength": "strong", "confidence": 0.90},
2499 {"from": "ST4", "to": "ST2", "type": "depends_on",
2500 "strength": "strong", "confidence": 0.90},
2501 {"from": "ST4", "to": "ST3", "type": "depends_on",
2502 "strength": "strong", "confidence": 0.90},
2503 {"from": "ST5", "to": "ST4", "type": "depends_on",
2504 "strength": "strong", "confidence": 0.90},
2505 {"from": "ST6", "to": "ST5", "type": "depends_on",
2506 "strength": "strong", "confidence": 0.90},  

2507 /* Enables (operational support) */
2508 {"from": "O1", "to": "ST3", "type": "enables",
2509 "strength": "strong", "confidence": 0.88},
2510 {"from": "O2", "to": "ST4", "type": "enables",
2511 "strength": "strong", "confidence": 0.88},
2512 {"from": "O3", "to": "ST6", "type": "enables",
2513 "strength": "strong", "confidence": 0.88},
2514 {"from": "O4", "to": "ST7", "type": "enables",
2515 "strength": "strong", "confidence": 0.88},  

2516 /* Support / exemplify / validate relationships */
2517 {"from": "O1", "to": "ST1", "type": "exemplifies",
2518 "strength": "strong", "confidence": 0.85},
2519 {"from": "O1", "to": "ST3", "type": "exemplifies",
2520 "strength": "strong", "confidence": 0.85},
2521 {"from": "O2", "to": "ST2", "type": "exemplifies",
2522 "strength": "strong", "confidence": 0.85},
2523 {"from": "O2", "to": "ST4", "type": "exemplifies",
2524 "strength": "strong", "confidence": 0.85},
2525 {"from": "O3", "to": "ST6", "type": "exemplifies",
2526 "strength": "strong", "confidence": 0.85},
2527 {"from": "O4", "to": "ST7", "type": "exemplifies",
2528 "strength": "strong", "confidence": 0.85},
2529 {"from": "O1", "to": "ST3", "type": "validates",
2530 "strength": "strong", "confidence": 0.84},
2531 {"from": "O2", "to": "ST4", "type": "validates",
2532 "strength": "strong", "confidence": 0.84},
2533 {"from": "O3", "to": "ST6", "type": "validates",
2534 "strength": "strong", "confidence": 0.84},
2535 /* The edge list above contains every relationship identified in the
2536 ↪ analysis: hierarchical *decomposes_to*, logical *depends_on*,
2537 ↪ operational
2538 *enables*, and verification *exemplifies* / *validates* edges.*
```

```

2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
---  

## Verification Flow  

1. **Entry Points** - Nodes with no incoming edges: **G1** (goal) and  

→ **S1** (strategy) are the root audit entry points.  

2. **Critical Path** - The sequence that must be verified in order to  

→ guarantee correctness of the final answer:  

'G1 ->S1 ->T1 ->ST1 ->ST3 ->ST5 ->ST6 ->ST7 ->O4'  

(Goal ->Strategy ->Tactic ->reduced-output inference ->  

→ remaining-portion time ->algebraic combination ->solve for \(\mathcal{D}\) ->  

→ final conclusion).  

3. **Parallel Tracks** - Independent branches that can be audited  

→ concurrently:  

- **Branch A**: 'ST1 ->O1' (reduced-output computation).  

- **Branch B**: 'ST2 ->O2' (first-third time).  

- **Branch C**: 'ST4 ->O2' (total-time equation).  

- **Branch D**: 'ST5 ->O2' (fraction combination).  

All arithmetic steps (ST1-ST7 and O1-O4) are pure numeric operations  

→ and can be run in parallel after their logical prerequisites are  

→ satisfied.  

4. **Exit Points** - Nodes whose successful audit marks the HDAG as  

→ completely validated: **O3** (verification of the solved \(\mathcal{D}\)) and  

→ **O4**  

(verification of the final missing-days value).  

---  

## HDAG Validation Report  

- **Acyclicity Check**: No cycles detected - the graph is a directed  

→ acyclic graph.  

- **Connectivity Check**: Every node is reachable from the root node  

→ **G1** either directly or via intermediate nodes; there are no  

→ isolated  

fragments.  

- **Completeness Check**: All 14 original segments are present and  

→ their textual content has been preserved unchanged.  

- **Auditability Check**: Every node lists primary and secondary  

→ auditors, a verification-complexity label, and explicit  

→ dependency/enabler fields,  

giving a clear audit trail for both automated and human reviewers.  

---  

## Complete HDAG JSON  

```json
{
  "nodes": [
    {
      "id": "G1",
      "content": "A construction company was building a tunnel. When 1/3  

→ of the tunnel was completed at the original speed, they started  

→ using new  

equipment, which increased the construction speed by 20 % and reduced  

→ the working hours to 80 % of the original. As a result, it took a  

→ total of "
    }
  ]
}
```

```

```

2592    185 days to complete the tunnel. If they had not used the new equipment
2593    ↪ and continued at the original speed, it would have taken _____ days
2594    ↪ to
2595    complete the tunnel.",
2596    "abstraction_level": "GOAL",
2597    "type": "objective_statement",
2598    "summary": "find missing days",
2599    "primary_auditor": "T_Human",
2600    "secondary_auditors": ["T_LLM"],
2601    "complexity": "High"
2602  },
2603  {
2604    "id": "S1",
2605    "content": "Model the problem by relating speed, working hours,
2606    ↪ and daily output. The original daily output is D. After the
2607    ↪ equipment change
2608    the speed rises 20 % while hours fall to 80 % ->new daily output
2609    ↪ 0.96D.",
2610    "abstraction_level": "STRATEGY",
2611    "type": "approach_description",
2612    "summary": "model speed & hours",
2613    "primary_auditor": "T_LLM",
2614    "secondary_auditors": ["T_Auto"],
2615    "complexity": "Medium"
2616  },
2617  {
2618    "id": "T1",
2619    "content": "Let the tunnel length be 1 unit and the original daily
2620    ↪ output be D. Build 1/3 at rate D; the remaining 2/3 at rate
2621    ↪ 0.96D. Set up
2622    the time equation and solve for D.",
2623    "abstraction_level": "TACTIC",
2624    "type": "method_definition",
2625    "summary": "define variables, solve",
2626    "primary_auditor": "T_Auto",
2627    "secondary_auditors": ["T_LLM"],
2628    "complexity": "Low"
2629  },
2630  {
2631    "id": "ST1",
2632    "content": "Compute the reduced daily output: 0.96D.",
2633    "abstraction_level": "STEP",
2634    "type": "inference",
2635    "summary": "compute reduced output",
2636    "primary_auditor": "T_Auto",
2637    "secondary_auditors": ["T_LLM"],
2638    "complexity": "Low"
2639  },
2640  {
2641    "id": "ST2",
2642    "content": "Time for the first third:  $t_1 = 1/(3D)$ .",
2643    "abstraction_level": "STEP",
2644    "type": "inference",
2645    "summary": "first-third time",
2646    "primary_auditor": "T_Auto",
2647    "secondary_auditors": ["T_LLM"],
2648    "complexity": "Low"
2649  },
2650  {
2651    "id": "ST3",
2652    "content": "Time for the remaining two-thirds:  $t_2 = (2/3)/(0.96D)$ 
2653    ↪ = 5/(8D) .",
2654    "abstraction_level": "STEP",
2655    "type": "inference",
2656    "summary": "remaining portion time",

```

```

2646     "primary_auditor": "T_Auto",
2647     "secondary_auditors": ["T_LLM"],
2648     "complexity": "Low"
2649   },
2650   {
2651     "id": "ST4",
2652     "content": "Total-time equation:  $t_1 + t_2 = 185$ .",
2653     "abstraction_level": "STEP",
2654     "type": "inference",
2655     "summary": "total-time equation",
2656     "primary_auditor": "T_Auto",
2657     "secondary_auditors": ["T_LLM"],
2658     "complexity": "Low"
2659   },
2660   {
2661     "id": "ST5",
2662     "content": "Combine fractions:  $(1/3 + 5/8)/D = 185 \rightarrow (23/24)/D =$ 
2663      $\rightarrow 185$ .",
2664     "abstraction_level": "STEP",
2665     "type": "algebraic_combination",
2666     "summary": "algebraic combination",
2667     "primary_auditor": "T_Auto",
2668     "secondary_auditors": ["T_LLM"],
2669     "complexity": "Low"
2670   },
2671   {
2672     "id": "ST6",
2673     "content": "Solve for D:  $D = (23/24)/185 \approx 0.0054$ .",
2674     "abstraction_level": "STEP",
2675     "type": "solution_step",
2676     "summary": "solve for D",
2677     "primary_auditor": "T_Auto",
2678     "secondary_auditors": ["T_LLM"],
2679     "complexity": "Low"
2680   },
2681   {
2682     "id": "ST7",
2683     "content": "Original total time =  $1/D \approx 365$  days.",
2684     "abstraction_level": "STEP",
2685     "type": "conclusion_inference",
2686     "summary": "original schedule result",
2687     "primary_auditor": "T_Auto",
2688     "secondary_auditors": ["T_LLM"],
2689     "complexity": "Low"
2690   },
2691   {
2692     "id": "O1",
2693     "content": "Numeric calculation  $0.96 = 20\% \text{ increase} \times 80\%$ 
2694      $\rightarrow \text{reduction}$ .",
2695     "abstraction_level": "OPERATION",
2696     "type": "calculation",
2697     "summary": "reduced-output calc",
2698     "primary_auditor": "T_Auto",
2699     "secondary_auditors": ["T_LLM"],
2700     "complexity": "Low"
2701   },
2702   {
2703     "id": "O2",
2704     "content": "Add the two time expressions to get 185 days.",
2705     "abstraction_level": "OPERATION",
2706     "type": "calculation",
2707     "summary": "build time equation",
2708     "primary_auditor": "T_Auto",
2709     "secondary_auditors": ["T_LLM"],
2710     "complexity": "Low"
2711   }

```

```

2700 },
2701 {
2702   "id": "03",
2703   "content": "Compute D = (23/24)/185 \approx 0.0054.",
2704   "abstraction_level": "OPERATION",
2705   "type": "calculation",
2706   "summary": "compute D",
2707   "primary_auditor": "T_Auto",
2708   "secondary_auditors": ["T_LLM"],
2709   "complexity": "Low"
2710 },
2711 {
2712   "id": "04",
2713   "content": "Take reciprocal of D ->1/D \approx 365 days.",
2714   "abstraction_level": "OPERATION",
2715   "type": "calculation",
2716   "summary": "final answer calc",
2717   "primary_auditor": "T_Auto",
2718   "secondary_auditors": ["T_LLM"],
2719   "complexity": "Low"
2720 },
2721 ],
2722 "edges": [
2723   /* (the JSON edge list shown earlier) */
2724 ],
2725 "metadata": {
2726   "total_nodes": 14,
2727   "total_edges": 28,
2728   "primary_auditor_distribution": {
2729     "T_Human": 1,
2730     "T_LLM": 1,
2731     "T_Auto": 12
2732   }
2733 }
2734 ```
2735
2736 *The JSON follows the exact structure used throughout the analysis,  

2737 → making it ready for ingestion by downstream validation tools.*  

2738  

2739  

2740  

2741  

2742  

2743  

2744  

2745  

2746  

2747  

2748  

2749  

2750  

2751  

2752  

2753

```

2754 C.2 HUMAN-FRIENDLY AUDIT EXAMPLE  
27552756 This section demonstrates how human auditors interact with the TRUST system through a structured  
2757 audit interface with human friendly presentation. The example shows a typical reasoning segment  
2758 decomposition where auditors evaluate logical dependencies and inference validity.  
27592760 C.2.1 AUDIT INTERFACE STRUCTURE  
27612762 The human audit form presents reasoning segments with four key components that enable focused  
2763 evaluation:  
27642765 **Prompt Context:** The original problem statement that establishes the reasoning foundation.  
27662767 **Dependencies:** Previously verified reasoning steps that the current segment builds upon.  
27682769 **Current Reasoning:** The specific reasoning step under evaluation.  
27702771 **Implications:** The logical consequences that flow from this reasoning step.  
27722773 C.2.2 EXAMPLE AUDIT CASE  
27742775 **Prompt Context:** Martha needs 4 cups of berries  
2776 and 2 cups of heavy cream to make 1 quart of ice  
2777 cream. She wants to make 1 quart of strawberry ice  
2778 cream and 1 quart of raspberry ice cream. At the  
2779 farmers market, 2-cup packages of strawberries cost  
2780 \$3.00 each and 2-cup packages of raspberries cost  
2781 \$5.00 each. Heavy cream is sold in 4-cup containers  
2782 for \$4.00. How much will it cost her to make 1 quart  
2783 of each ice cream?  
27842785 **Dependencies (Verified Step S1):** “For each quart  
2786 of ice cream she needs: 4 cups of berries, 2 cups of  
2787 heavy cream. She wants one quart of strawberry ice  
2788 cream and one quart of raspberry ice cream.”  
27892790 **Current Reasoning Step (T1):** “For 1 quart of ice  
2791 cream, she needs 4 cups of berries and 2 cups of  
2792 heavy cream. Since she’s making both strawberry  
2793 and raspberry flavors, she’ll need double that amount.  
2794 For both ice creams combined, she needs 8 cups of berries and 4 cups of heavy cream.”  
27952796 **Logical Implications (T3):** “She needs 8 cups total berries. Since each package contains 2 cups,  
2797 she needs  $8/2 = 4$  berry packages total (split between strawberry and raspberry varieties).”  
27982799 **Self-Verification (ST1):** “Each quart requires 4 cups of berries and 2 cups of cream. For two quarts:  
2800 8 cups berries and 4 cups cream. This calculation is correct.”  
28012802 C.2.3 AUDIT DECISION FRAMEWORK  
28032804 Human auditors evaluate each reasoning segment using structured criteria:  
28052806 

- **Logical Consistency:** Does the reasoning step follow logically from its dependencies?
- **Mathematical Accuracy:** Are calculations and quantitative reasoning correct?
- **Completeness:** Does the reasoning address all relevant aspects of the problem?
- **Clarity:** Is the reasoning step clearly articulated and understandable?

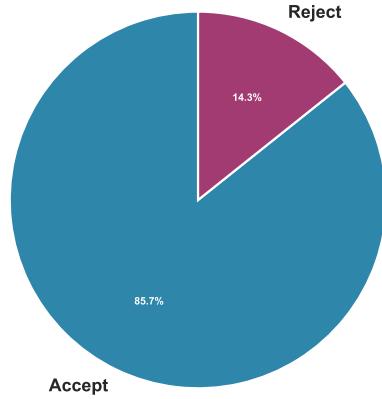
  
28072808 **Ground Truth Evaluation:** Pass ✓  
2809

Figure 9: Human Audit Results.

2808 The reasoning correctly identifies that making two quarts requires doubling the per-quart ingredient  
2809 requirements, performs accurate arithmetic ( $2 \times 4 = 8$  cups berries,  $2 \times 2 = 4$  cups cream), and  
2810 includes appropriate self-verification to confirm the calculation.

2811 This structured approach enables human auditors to provide focused, reliable evaluations while  
2812 maintaining the modular verification principles of the TRUST framework.  
2813

2814

2815

2816

2817

2818

2819

2820

2821

2822

2823

2824

2825

2826

2827

2828

2829

2830

2831

2832

2833

2834

2835

2836

2837

2838

2839

2840

2841

2842

2843

2844

2845

2846

2847

2848

2849

2850

2851

2852

2853

2854

2855

2856

2857

2858

2859

2860

2861

2862  
2863

## C.3 HDAG EXAMPLE 1 - MATH PROBLEM

2864  
2865  
2866  
2867

Problem: Marie is planning to buy a new cash register for her bakery that costs \$1040. Every day Marie sells 40 loaves of bread for \$2 each and 6 cakes for \$12 each. She has to pay \$20 each day for rent and \$2 each day for electricity. How many days' worth of profits will it take for Marie to pay for the cash register?

2868

## C.3.1 DEEPSEEK-R1-8B CoT DECOMPOSITION

2870

2871

2872

2873

2874

2875

2876

2877

2878

2879

2880

2881

2882

2883

2884

2885

2886

2887

2888

2889

2890

2891

2892

2893

2894

2895

2896

2897

2898

2899

2900

2901

2902

2903

2904

2905

2906

2907

2908

2909

2910

2911

2912

2913

2914

2915

```
{
  "title": "Marie's Cash Register Days",
  "nodes": [
    {"id": "G1", "label": "Problem statement", "content": "## Problem
    ↪ Statement: \nMarie is planning to buy a new cash register for her
    ↪ bakery that costs $1040. Every day Marie sells 40 loaves of bread
    ↪ for $2 each and 6 cakes for $12 each. She has to pay $20 each day for
    ↪ rent and $2 each day for electricity. How many days' worth of
    ↪ profits will it take for Marie to pay for the cash
    ↪ register?", "abstraction_level": "GOAL",
    "type": "objective_statement", "primary_auditor": "T_Human",
    "complexity": "Low"},

    {"id": "S1", "label": "Plan overview", "content": "Okay, so Marie wants
    ↪ to buy a new cash register for her bakery, and it costs $1040. I
    ↪ need to figure out how many days' worth of profits it will take
    ↪ her to afford this cash register. Let me break this down step by
    ↪ step.", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM",
    "complexity": "Low"},

    {"id": "S2", "label": "Daily revenue", "content": "First, I should
    ↪ understand what her daily income is. She sells 40 loaves of bread
    ↪ each day, and each loaf is $2. So, the revenue from bread would
    ↪ be 40 times $2. Let me calculate that:  $40 * 2 = \$80$ . Then, she
    ↪ sells 6 cakes each day, and each cake is $12. So, the revenue
    ↪ from cakes is  $6 * 12 = \$72$ . Adding those together, her total
    ↪ daily revenue is  $\$80 + \$72 = \$152$ .", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM",
    "complexity": "Low"},

    {"id": "S3", "label": "Daily profit", "content": "Next, I need to
    ↪ calculate her daily expenses. She has to pay $20 each day for
    ↪ rent and $2 each day for electricity. So, adding those together:
    ↪  $\$20 + \$2 = \$22$ . That means her daily profit is the revenue minus
    ↪ the expenses. So,  $\$152$  (revenue) -  $\$22$  (expenses) =  $\$130$  per
    ↪ day.", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM",
    "complexity": "Low"},

    {"id": "T1", "label": "Setup division", "content": "Now, the cash
    ↪ register costs $1040, and she makes $130 each day. To find out
    ↪ how many days it will take her to earn enough to buy the cash
    ↪ register, I should divide the total cost by her daily profit. So,
    ↪  $\$1040$  divided by  $\$130$  per day. Let me do that division:  $1040 / 130$ .", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM",
    "complexity": "Low"},

    {"id": "T2", "label": "Verify quotient", "content": "Hmm,  $130$  times  $8$  is
    ↪  $1040$  because  $130 * 8 = 1040$ . So, it would take her  $8$  days to earn
    ↪ enough money to buy the cash
    ↪ register.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM",
    "complexity": "Low"},

    {"id": "ST1", "label": "Double-check math", "content": "Wait, let me
    ↪ double-check my calculations to make sure I didn't make any
    ↪ mistakes. Revenue from bread:  $40 * 2 = 80$ . Revenue from cakes:  $6
    ↪ * 12 = 72$ . Total revenue:  $80 + 72 = 152$ . Expenses:  $20 + 2 = 22$ .
    ↪ Daily profit:  $152 - 22 = 130$ . Cash register cost:  $1040 / 130 = 8$ . Yep, that seems
    ↪ correct.", "abstraction_level": "STEP", "primary_auditor": "T_Human",
    "complexity": "Low"}]
```

```

2916 {"id": "O1", "label": "Bread revenue", "content": "40 * 2 =  

2917    ↪ $80", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2918    "complexity": "Low"},  

2919 {"id": "O2", "label": "Cakes revenue", "content": "6 * 12 =  

2920    ↪ $72", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2921    "complexity": "Low"},  

2922 {"id": "O3", "label": "Total revenue", "content": "$80 + $72 =  

2923    ↪ $152", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2924    "complexity": "Low"},  

2925 {"id": "O4", "label": "Daily expenses", "content": "$20 + $2 =  

2926    ↪ $22", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2927    "complexity": "Low"},  

2928 {"id": "O5", "label": "Daily profit", "content": "$152 - $22 =  

2929    ↪ $130", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2930    "complexity": "Low"},  

2931 {"id": "O6", "label": "Days needed", "content": "1040 / 130 =  

2932    ↪ 8", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto",  

2933    "complexity": "Low"},  

2934 {"id": "O7", "label": "Final answer 8", "content": "So, it will take  

2935    ↪ Marie 8 days to save enough money to buy the cash  

2936    ↪ register.", "abstraction_level": "OPERATION",  

2937    "primary_auditor": "T_Auto", "complexity": "Low"}  

2938 ],  

2939 "edges": [  

2940     {"from": "G1", "to": "S1", "relationship": "decomposes_to",  

2941     "strength": "strong", "confidence": 0.97},  

2942     {"from": "S1", "to": "S2", "relationship": "decomposes_to",  

2943     "strength": "strong", "confidence": 0.96},  

2944     {"from": "S1", "to": "S3", "relationship": "decomposes_to",  

2945     "strength": "strong", "confidence": 0.96},  

2946     {"from": "S2", "to": "O1", "relationship": "decomposes_to",  

2947     "strength": "strong", "confidence": 0.99},  

2948     {"from": "S2", "to": "O2", "relationship": "decomposes_to",  

2949     "strength": "strong", "confidence": 0.99},  

2950     {"from": "S2", "to": "O3", "relationship": "decomposes_to",  

2951     "strength": "strong", "confidence": 0.99},  

2952     {"from": "S3", "to": "O4", "relationship": "decomposes_to",  

2953     "strength": "strong", "confidence": 0.99},  

2954     {"from": "S3", "to": "O5", "relationship": "decomposes_to",  

2955     "strength": "strong", "confidence": 0.99},  

2956     {"from": "T1", "to": "O6", "relationship": "decomposes_to",  

2957     "strength": "strong", "confidence": 0.99},  

2958     {"from": "T2", "to": "O6", "relationship": "validates",  

2959     "strength": "strong", "confidence": 0.97},  

2960     {"from": "O1", "to": "ST1", "relationship": "enables",  

2961     "strength": "strong", "confidence": 0.95},  

2962     {"from": "O5", "to": "T1", "relationship": "enables",  

2963     "strength": "strong", "confidence": 0.95},  

2964     {"from": "ST1", "to": "O7", "relationship": "decomposes_to",  

2965     "strength": "medium", "confidence": 0.90}  

2966 ],  

2967 "metadata": {  

2968     "total_nodes": 14,  

2969     "total_edges": 13,  

2970     "auditor_distribution": { "T_Auto": 7, "T_LLM": 5, "T_Human": 2 }  

2971   }
2972 }
2973 }
```

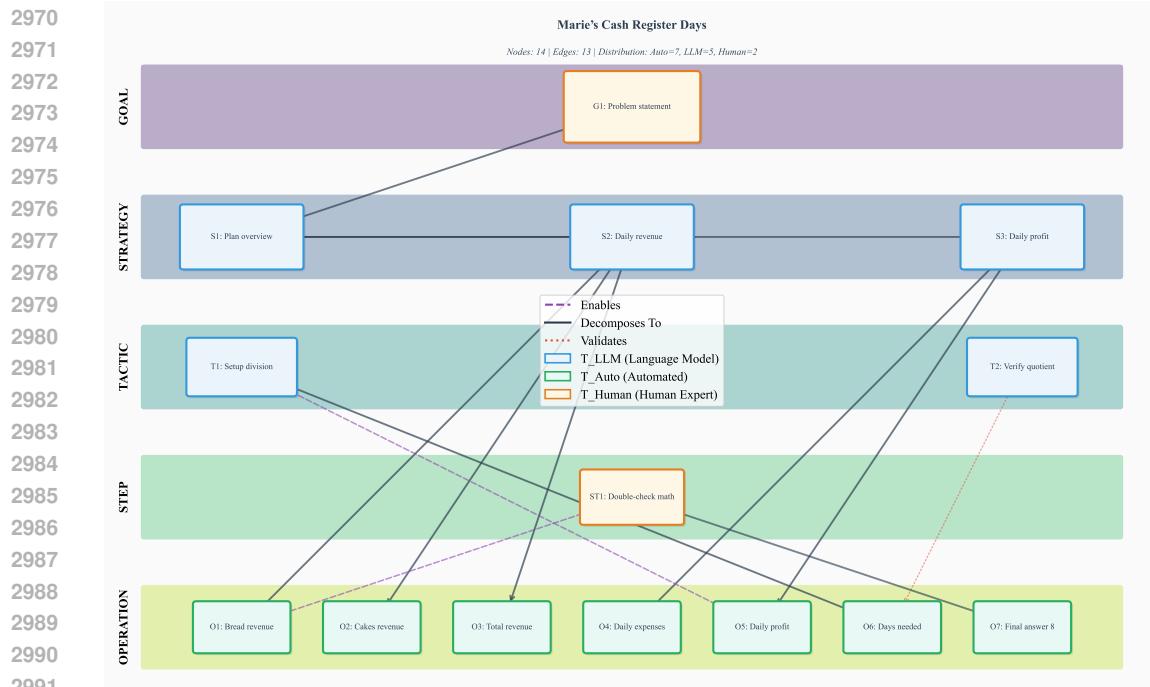


Figure 10: HDAG Example 1 from Deepseek-R1-8B CoT: Math Problem.

### C.3.2 GPT-OSS-20B CoT DECOMPOSITION

```
2998
2999 {
3000     "title": "Marie's Cash Register - Step-by-Step",
3001     "nodes": [
3002         {"id": "G1", "label": "Understand problem", "content": "***1. Understand  
3002         the problem**\n\nMarie wants to know how many days of operating  
3003         her bakery will be needed to earn enough profit to buy a cash  
3004         register that costs **$1040**. \nEach day she:\n\n- Sells 40  
3005         loaves of bread at **$2** each ->revenue from bread \n- Sells 6  
3006         cakes at **$12** each ->revenue from cakes \n- Pays **$20** for  
3007         rent and **$2** for electricity ->daily expenses \n\nWe need the  
3008         daily profit (revenue minus expenses) and then divide the  
3009         register cost by that daily profit to find the number of  
3010         days.", "abstraction_level": "GOAL", "type": "objective_statement", "primary_auditor": "T_Human",  
3011         {"id": "S1", "label": "Steps overview", "content": "***2. Show each step  
3012         of your work**", "abstraction_level": "STRATEGY",  
3013         "primary_auditor": "T_LLM"},  
3014         {"id": "S2", "label": "Daily revenue", "content": "1. **Daily  
3015         revenue**\n\n- Bread: \\(40 \\text{ loaves} \\times \\$2 =  
3016         \\$80\\)\n- Cakes: \\(6 \\text{ cakes} \\times \\$12 =  
3017         \\$72\\)\n\n\\[\\text{Total revenue} = \\$80 + \\$72 =  
3018         \\$152\\]", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},  
3019         {"id": "S3", "label": "Daily expenses", "content": "2. **Daily  
3020         expenses**\n\n\\[\\text{Rent} = \\$20, \\quad  
3021         \\text{Electricity} = \\$2\\]\n\n\\[\\text{Total expenses} = \\$20 + \\$2 = \\$22\\]", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},  
3022         {"id": "S4", "label": "Daily profit", "content": "3. **Daily profit**\n\n\\[\\text{Profit} = \\text{Revenue} - \\text{Expenses}\\]\n\n\\[\\text{Profit} = \\$152 - \\$22 = \\$130\\]", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},  
3023     ]
```

```

3024 {"id": "S5", "label": "Days formula", "content": "4. **Number of days to
3025   cover the cash register**\n\n \\[\n \\text{Days} =
3026   \\frac{\\text{Cost of register}}{\\text{Daily profit}}\\]\n\n
3027   \\[\n \\text{Days} = \\frac{\\$1040}{\\$130} = 8\\n
3028   \\]\", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},\n
3029   {"id": "T1", "label": "Revenue tactic", "content": "- Compute daily\n
3030   revenue from bread and\n
3031   cakes.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},\n
3032   {"id": "T2", "label": "Expenses tactic", "content": "- Compute daily\n
3033   expenses (rent +\n
3034   electricity).", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},\n
3035   {"id": "T3", "label": "Profit tactic", "content": "- Compute daily profit\n
3036   as revenue minus\n
3037   expenses.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},\n
3038   {"id": "T4", "label": "Days tactic", "content": "- Divide register cost\n
3039   by daily profit to get\n
3040   days.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},\n
3041   {"id": "ST1", "label": "Verify math", "content": "***3. Verify your\n
3042   calculations**\n- Check revenue: \\(40 \\times 2 = 80\\), \\(6\n
3043   \\times 12 = 72\\), \\(80 + 72 = 152\\).\n- Check expenses:\n
3044   \\(20 + 2 = 22\\).\n- Profit: \\(152 - 22 = 130\\).\n- Days:\n
3045   \\(130 \\times 8 = 1040\\).\nEverything is\n
3046   consistent.", "abstraction_level": "STEP", "primary_auditor": "T_Human"},\n
3047   {"id": "O1", "label": "Bread revenue", "content": "\\(40 \\times 2 =\n
3048   80\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3049   {"id": "O2", "label": "Cakes revenue", "content": "\\(6 \\times 12 =\n
3050   72\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3051   {"id": "O3", "label": "Total revenue", "content": "\\(80 + 72 =\n
3052   152\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3053   {"id": "O4", "label": "Total expenses", "content": "\\(20 + 2 =\n
3054   22\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3055   {"id": "O5", "label": "Daily profit", "content": "\\(152 - 22 =\n
3056   130\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3057   {"id": "O6", "label": "Days needed", "content": "\\(\\frac{1040}{130} =\n
3058   8\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},\n
3059   {"id": "O7", "label": "Verify product", "content": "\\(130 \\times 8 =\n
3060   1040\\)", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"}\n
3061   ],\n
3062   "edges": [\n
3063     {"from": "G1", "to": "S1", "relationship": "decomposes_to"},\n
3064     {"from": "G1", "to": "S2", "relationship": "decomposes_to"},\n
3065     {"from": "G1", "to": "S3", "relationship": "decomposes_to"},\n
3066     {"from": "G1", "to": "S4", "relationship": "decomposes_to"},\n
3067     {"from": "G1", "to": "S5", "relationship": "decomposes_to"},\n
3068     {"from": "S2", "to": "O1", "relationship": "decomposes_to"},\n
3069     {"from": "S2", "to": "O2", "relationship": "decomposes_to"},\n
3070     {"from": "S2", "to": "O3", "relationship": "decomposes_to"},\n
3071     {"from": "S3", "to": "O4", "relationship": "decomposes_to"},\n
3072     {"from": "S4", "to": "O5", "relationship": "decomposes_to"},\n
3073     {"from": "S5", "to": "O6", "relationship": "decomposes_to"},\n
3074     {"from": "ST1", "to": "O7", "relationship": "decomposes_to"},\n
3075     {"from": "O3", "to": "O5", "relationship": "enables"},\n
3076     {"from": "O4", "to": "O5", "relationship": "enables"},\n
3077     {"from": "O5", "to": "O6", "relationship": "enables"},\n
3078     {"from": "O7", "to": "O6", "relationship": "validates"}\n
3079   ],\n
3080   "metadata": {\n
3081     "total_nodes": 17,\n
3082     "total_edges": 16,\n
3083     "auditor_distribution": { "T_Auto": 7, "T_LLM": 9, "T_Human": 2 }\n
3084   }\n
3085 }\n
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
3459
3460
3461
3462
3463
3464
3465
3466
3467
3468
3469
3470
3471
3472
3473
3474
3475
3476
3477
3478
3479
3480
3481
3482
3483
3484
3485
3486
3487
3488
3489
3490
3491
3492
3493
3494
3495
3496
3497
3498
3499
3500
3501
3502
3503
3504
3505
3506
3507
3508
3509
3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3550
3551
3552
3553
3554
3555
3556
3557
3558
3559
3560
3561
3562
3563
3564
3565
3566
3567
3568
3569
3570
3571
3572
3573
3574
3575
3576
3577
3578
3579
3580
3581
3582
3583
3584
3585
3586
3587
3588
3589
3590
3591
3592
3593
3594
3595
3596
3597
3598
3599
3600
3601
3602
3603
3604
3605
3606
3607
3608
3609
3610
3611
3612
3613
3614
3615
3616
3617
3618
3619
3620
3621
3622
3623
3624
3625
3626
3627
3628
3629
3630
3631
3632
3633
3634
3635
3636
3637
3638
3639
3640
3641
3642
3643
3644
3645
3646
3647
3648
3649
3650
3651
3652
3653
3654
3655
3656
3657
3658
3659
3660
3661
3662
3663
3664
3665
3666
3667
3668
3669
3670
3671
3672
3673
3674
3675
3676
3677
3678
3679
3680
3681
3682
3683
3684
3685
3686
3687
3688
3689
3690
3691
3692
3693
3694
3695
3696
3697
3698
3699
3700
3701
3702
3703
3704
3705
3706
3707
3708
3709
3710
3711
3712
3713
3714
3715
3716
3717
3718
3719
3720
3721
3722
3723
3724
3725
3726
3727
3728
3729
3730
3731
3732
3733
3734
3735
3736
3737
3738
3739
3740
3741
3742
3743
3744
3745
3746
3747
3748
3749
3750
3751
3752
3753
3754
3755
3756
3757
3758
3759
3760
3761
3762
3763
3764
3765
3766
3767
3768
3769
3770
3771
3772
3773
3774
3775
3776
3777
3778
3779
3780
3781
3782
3783
3784
3785
3786
3787
3788
3789
3790
3791
3792
3793
3794
3795
3796
3797
3798
3799
3800
3801
3802
3803
3804
3805
3806
3807
3808
3809
3810
3811
3812
3813
3814
3815
3816
3817
3818
3819
3820
3821
3822
3823
3824
3825
3826
3827
3828
3829
3830
3831
3832
3833
3834
3835
3836
3837
3838
3839
3840
3841
3842
3843
3844
3845
3846
3847
3848
3849
3850
3851
3852
3853
3854
3855
3856
3857
3858
3859
3860
3861
3862
3863
3864
3865
3866
3867
3868
3869
3870
3871
3872
3873
3874
3875
3876
3877
3878
3879
3880
3881
3882
3883
3884
3885
3886
3887
3888
3889
3890
3891
3892
3893
3894
3895
3896
3897
3898
3899
3900
3901
3902
3903
3904
3905
3906
3907
3908
3909
3910
3911
3912
3913
3914
3915
3916
3917
3918
3919
3920
3921
3922
3923
3924
3925
3926
3927
3928
3929
3930
3931
3932
3933
3934
3935
3936
3937
3938
3939
3940
3941
3942
3943
3944
3945
3946
3947
3948
3949
3950
3951
3952
3953
3954
3955
3956
3957
3958
3959
3960
3961
3962
3963
3964
3965
3966
3967
3968
3969
3970
3971
3972
3973
3974
3975
3976
3977
3978
3979
3980
3981
3982
3983
3984
3985
3986
3987
3988
3989
3990
3991
3992
3993
3994
3995
3996
3997
3998
3999
3999
4000
4001
4002
4003
4004
4005
4006
4007
4008
4009
4010
4011
4012
4013
4014
4015
4016
4017
4018
4019
4020
4021
4022
4023
4024
4025
4026
4027
4028
4029
4030
4031
4032
4033
4034
4035
4036
4037
4038
4039
4040
4041
4042
4043
4044
4045
4046
4047
4048
4049
4050
4051
4052
4053
4054
4055
4056
4057
4058
4059
4060
4061
4062
4063
4064
4065
4066
4067
4068
4069
4070
4071
4072
4073
4074
4075
4076
4077
4078
4079
4080
4081
4082
4083
4084
4085
4086
4087
4088
4089
4090
4091
4092
4093
4094
4095
4096
4097
4098
4099
4099
4100
4101
4102
4103
4104
4105
4106
4107
4108
4109
4110
4111
4112
4113
4114
4115
4116
4117
4118
4119
4119
4120
4121
4122
4123
4124
4125
4126
4127
4128
4129
4129
4130
4131
4132
4133
4134
4135
4136
4137
4138
4139
4139
4140
4141
4142
4143
4144
4145
4146
4147
4148
4149
4149
4150
4151
4152
4153
4154
4155
4156
4157
4158
4159
4159
4160
4161
4162
4163
4164
4165
4166
4167
4168
4169
4169
4170
4171
4172
4173
4174
4175
4176
4177
4178
4179
4179
4180
4181
4182
4183
4184
4185
4186
4187
4187
4188
4189
4189
4190
4191
4192
4193
4194
4195
4196
4197
4197
4198
4199
4199
4199
4200
4201
4202
4203
4204
4205
4206
4207
4208
4209
4209
4210
4211
4212
4213
4214
4215
4216
4217
4218
4219
4219
4220
4221
4222
4223
4224
4225
4226
4227
4228
4228
4229
4230
4231
4232
4233
4234
4235
4236
4237
4238
4238
4239
4240
4241
4242
4243
4244
4245
4246
4247
4248
4248
4249
4250
4251
4252
4253
4254
4255
4256
4257
4258
4258
4259
4260
4261
4262
4263
4264
4265
4266
4267
4268
4268
4269
4270
4271
4272
4273
4274
4275
4276
4277
4278
4278
4279
4280
4281
4282
4283
4284
4285
4286
4287
4288
4288
4289
4290
4291
4292
4293
4294
4295
4296
4297
4298
4298
4299
4299
4300
4301
4302
4303
4304
4305
4306
4307
4308
4309
4309
4310
4311
4312
4313
4314
4315
4316
4317
4318
4319
4319
4320
4321
4322
4323
4324
4325
4326
4327
4328
4328
4329
4330
4331
4332
4333
4334
4335
4336
4337
4338
4338
4339
4340
4341
4342
4343
4344
4345
4346
4347
4348
4348
4349
4350
4351
4352
4353
4354
4355
4356
4357
4358
4359
4359
4360
4361
4362
4363
4364
4365
4366
4367
4368
4368
4369
4370
4371
4372
4373
4374
4375
4376
4377
4378
4379
4379
4380
4381
4382
4383
4384
4385
4386
4387
4388
4388
4389
4390
4391
4392
4393
4394
4395
4396
4397
4398
4398
4399
4399
4400
4401
4402
4403
4404
4405
4406
4407
4408
4409
4409
4410
4411
4412
4413
4414
4415
4416
4417
4418
4419
4419
4420
4421
4422
4423
4424
4425
4426
4427
4428
4428
4429
4430
4431
4432
4433
4434
4435
4436
4437
4438
4438
4439
4440
4441
4442
4443
4444
4445
4446
4447
4448
4448
4449
4449
4450
4451
4452
4453
4454
4455
4456
4457
4458
4459
4459
4460
4461
4462
4463
4464
4465
4466
4467
4468
4468
4469
4470
4471
4472
4473
4474
4475
4476
4477
4478
4478
4479
4480
4481
4482
4483
4484
4485
4486
4487
4488
4488
4489
4489
4490
4491
4492
4493
4494
4495
4496
4497
4498
4498
4499
4499
4500
4501
4502
4503
4504
4505
4506
4507
4508
4509
4509
4510
4511
4512
4513
4514
4515
4516
4517
4518
4519
4519
4520
4521
4522
4523
4524
4525
4526
4527
4528
4528
4529
4529
4530
4531
4532
4533
4534
4535
4536
4537
4538
4538
4539
4539
4540
4541
4542
4543
4544
4545
4546
4547
4548
4548
4549
4549
4550
4551
4552
4553
4554
4555
4556
4557
4558
4559
4559
4560
4561
4562
4563
4564
4565
4566
4567
4568
4568
4569
4569
4570
4571
4572
4573
4574
4575
4576
4577
4578
4579
4579
4580
4581
4582
4583
4584
4585
4586
4587
4588
4588
4589
4589
4590
4591
4592
4593
4594
4595
4596
4597
4598
4598
4599
4599
4600
4601
4602
4603
4604
4605
4606
4607
4608
4609
4609
4610
4611
4612
4613
4614
4615
4616
4617
4618
4619
4619
4620
4621
4622
4623
4624
4625
4626
4627
4628
4628
4629
4629
4630
4631
4632
4633
4634
4635
4636
4637
4638
4638
4639
4639
4640
4641
4642
4643
4644
4645
4646
4647
4648
4648
4649
4649
4650
4651
4652
4653
4654
4655
4656
4657
4658
4659
4659
4660
4661
4662
4663
4664
4665
4666
4667
4668
4668
4669
4669
4670
4671
4672
4673
4674
4675
4676
4677
4678
4678
4679
4679
4680
4681
4682
4683
4684
4685
4686
4687
4688
4688
4689
4689
4690
4691
4692
4693
4694
4695
4696
4697
4698
4698
4699
4699
4700
4701
4702
4703
4704
4705
4706
4707
4708
4709
4709
4710
4711
4712
4713
4714
4715
4716
4717
4718
4719
4719
4720
4721
4722
4723
4724
4725
4726
4727
4728
4728
4729
4729
4730
4731
4732
4733
4734
4735
4736
4737
4738
4738
4739
4739
4740
4741
4742
4743
4744
4745
4746
4747
4748
4748
4749
4749
4750
4751
4752
4753
4754
4755
4756
4757
4758
4759
4759
4760
4761
4762
4763
4764
4765
4766
4767
4768
4768
4769
4769
4770
4771
4772
4773
4774
4775
4776
4777
4778
4779
4779
4780
4781
4782
4783
4784
4785
4786
4787
4788
4788
4789
4789
4790
4791
4792
4793
4794
4795
4796
4797
4798
4798
4799
4799
4800
4801
4802
4803
4804
4805
4806
4807
4808
4809
4809
4810
4811
4812
4813
4814
4815
4816
4817
4818
4819
4819
4820
4821
4822
4823
4824
4825
4826
4827
4828
4829
4829
4830
4831
4832
4833
4834
4835
4836
4837
4838
4838
4839
4839
4840
4841
4842
4843
4844
4845
4846
4847
4848
4848
4849
4849
4850
4851
4852
4853
4854
4855
4856
4857
4858
4859
4859
4860
4861
4862
4863
4864
4865
4866
4867
4868
4868
4869
4869
4870
4871
4872
4873
4874
4875
4876
4877
4878
4879
4879
4880
4881
4882
4883
4884
4885
4886
4887
4888
4888
4889
4889
4890
4891
4892
4893
4894
4895
```

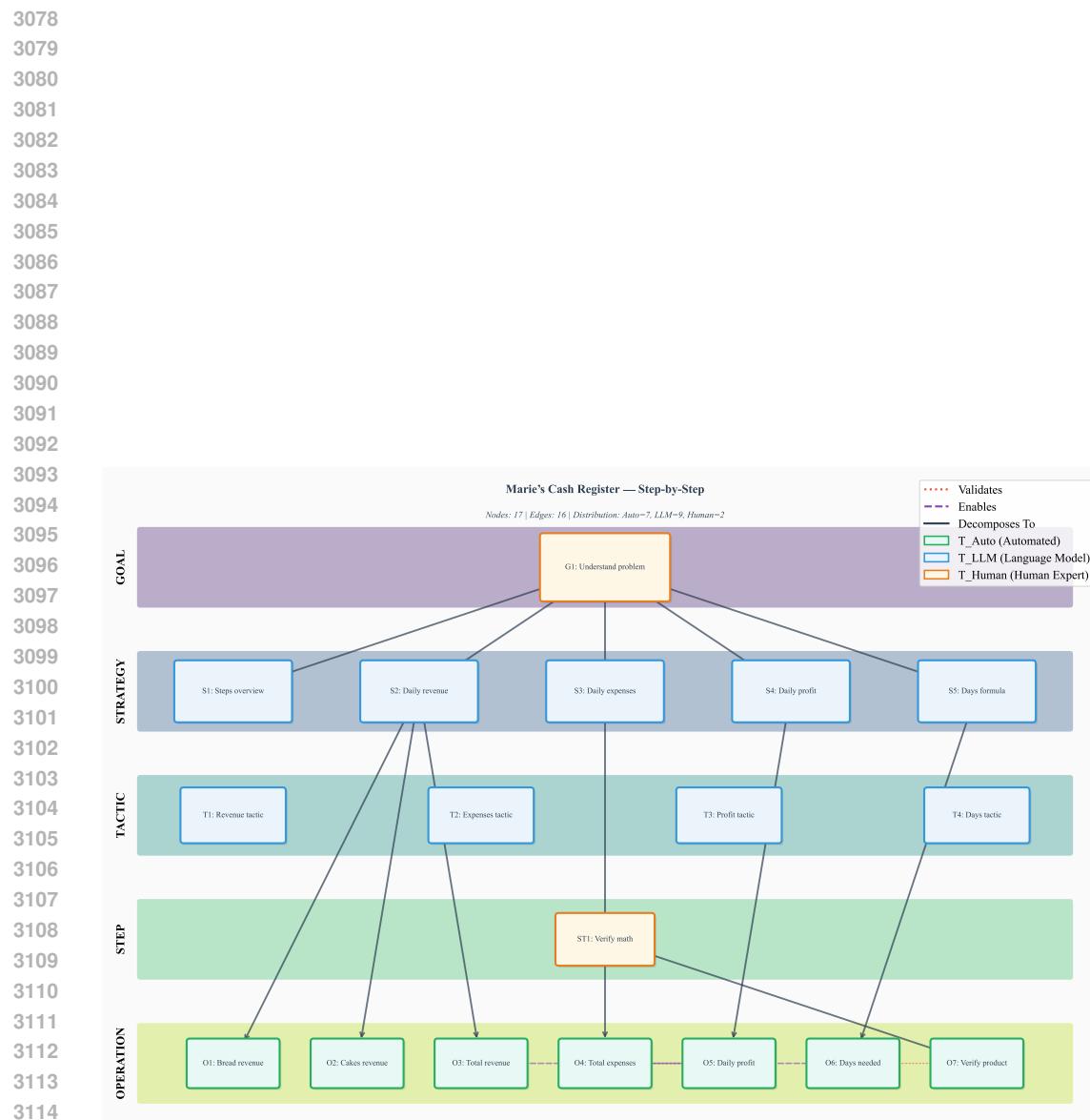


Figure 11: HDAG Example 1 from GPT-OSS-20B CoT: Math Problem.

3132  
3133

## C.4 HDAG EXAMPLE 2 - MATH PROBLEM

3134  
3135  
3136  
3137  
3138  
3139  
3140

Problem: Alec is running for Class President. He thinks that if he can get three-quarters of the class to vote for him then there is no chance anyone else can beat him. Half of the class have already said they will vote for him but out of the remaining students, only 5 have said they are thinking about voting for him. He surveys the students who are thinking about voting for someone else, and changes his flyers to reflect the issues these students are concerned about. This results in a fifth of these students saying they'll vote for him. If Alec's class has 60 students and everyone who said they will vote for him does so, how many more votes does Alec need to reach his goal number of votes?

3141  
3142

## C.4.1 DEEPSEEK-R1-8B CoT DECOMPOSITION

3143

```
{
  "title": "Alec's Class President Votes",
  "nodes": [
    {"id": "G1", "label": "Goal statement", "content": "Alec is running for
    ↪ Class President...", "abstraction_level": "GOAL",
    "type": "objective_statement", "primary_auditor": "T_Human"},

    {"id": "S1", "label": "Plan overview", "content": "Compute goal votes,
    ↪ count current and added, compare
    ↪ gap.", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},

    {"id": "T1", "label": "Needs 45 votes", "content": "3/4 x 60 =
    ↪ 45.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},

    {"id": "T2", "label": "30 committed", "content": "Half of 60 = 30 already
    ↪ pledged.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},

    {"id": "T3", "label": "5 lean, 5 convert", "content": "5 thinking about
    ↪ him; 1/5 of 25 switch = 5.",
    "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},

    {"id": "ST1", "label": "Gap check", "content": "45 - 40 = 5 more votes
    ↪ needed.",

    "abstraction_level": "STEP", "primary_auditor": "T_Human"},

    {"id": "O1", "label": "3/4 of
    ↪ 60", "content": "(3/4)x60=45", "abstraction_level": "OPERATION",
    "primary_auditor": "T_Auto"},

    {"id": "O2", "label": "Half class", "content": "60/2=30",
    "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},

    {"id": "O3", "label": "Remaining", "content": "60-30=30 remaining",
    "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},

    {"id": "O4", "label": "Convert votes", "content": "1/5 of 25=5",
    "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},

    {"id": "O5", "label": "Total votes", "content": "30+5+5=40",
    "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},

    {"id": "O6", "label": "Votes gap", "content": "45-40=5",
    "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},

    {"id": "O7", "label": "Final answer", "content": "Alec needs 5 more
    ↪ votes.", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"}],
  "edges": [
    {"from": "G1", "to": "S1", "relationship": "decomposes_to"},

    {"from": "S1", "to": "T1", "relationship": "decomposes_to"},

    {"from": "S1", "to": "T2", "relationship": "decomposes_to"},

    {"from": "S1", "to": "T3", "relationship": "decomposes_to"},

    {"from": "T1", "to": "O1", "relationship": "decomposes_to"},

    {"from": "T2", "to": "O2", "relationship": "decomposes_to"},

    {"from": "T2", "to": "O3", "relationship": "decomposes_to"},

    {"from": "T3", "to": "O4", "relationship": "decomposes_to"},

    {"from": "T3", "to": "O5", "relationship": "decomposes_to"},

    {"from": "O5", "to": "ST1", "relationship": "enables"},

    {"from": "O6", "to": "ST1", "relationship": "validates"},

    {"from": "ST1", "to": "O7", "relationship": "decomposes_to"}],
  "metadata": {
    "total_nodes": 13,
  }
}
```

```

3186     "total_edges": 12,
3187     "auditor_distribution": {"T_Auto":7, "T_LLM":4, "T_Human":2}
3188   }
3189 }
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
  }
```

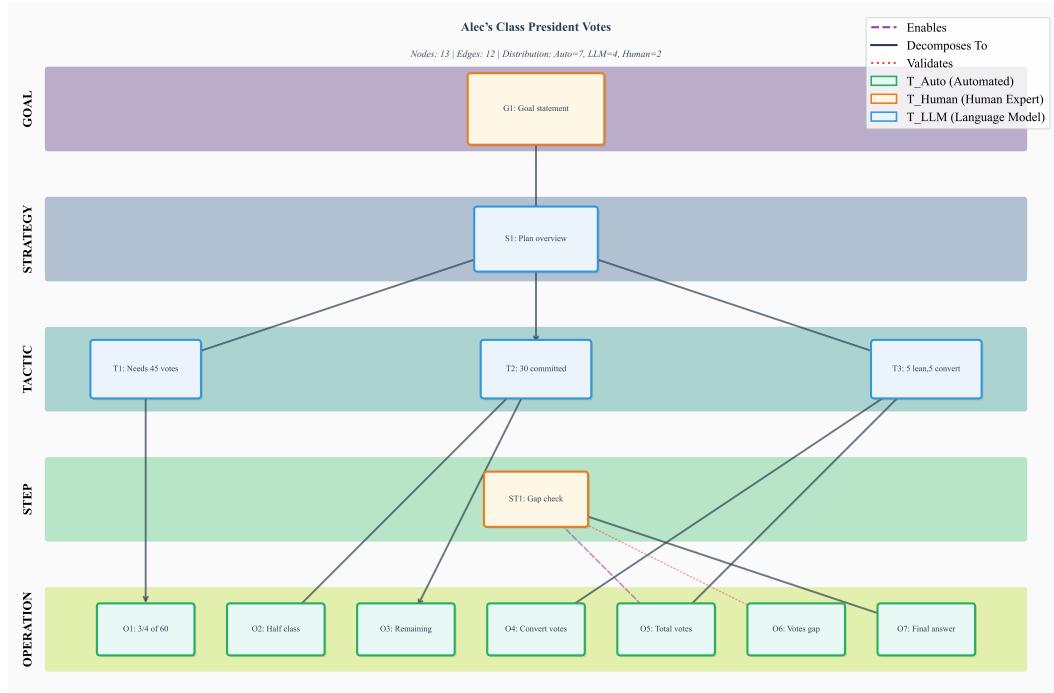


Figure 12: HDAG Example 2 from Deepseek-R1-8B CoT: Math Problem.

3240  
3241

## C.4.2 GPT-OSS-20B CoT DECOMPOSITION

3242

3243

3244

3245

3246

3247

3248

3249

3250

3251

3252

3253

3254

3255

3256

3257

3258

3259

3260

3261

3262

3263

3264

3265

3266

3267

3268

3269

3270

3271

3272

3273

3274

3275

3276

3277

3278

3279

3280

3281

3282

3283

3284

3285

3286

3287

3288

3289

3290

3291

3292

3293

```
{
  "title": "Alec's Votes",
  "nodes": [
    {"id": "G1", "label": "Goal statement", "content": "Alec wants 3/4 of 60  
→ = 45", "abstraction_level": "GOAL", "type": "objective_statement", "primary_auditor": "T_Human"},  
    {"id": "S1", "label": "Plan overview", "content": "Compute target, count  
→ votes, add converts, compare  
→ gap.", "abstraction_level": "STRATEGY", "primary_auditor": "T_LLM"},  
    {"id": "T1", "label": "Needs 45 votes", "content": "Three-quarters of 60  
→ is 45.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},  
    {"id": "T2", "label": "30 committed", "content": "Half of 60 already  
→ pledged = 30.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},  
    {"id": "T3", "label": "5 converted", "content": "From 25 opposing, 1/5  
→ switched = 5.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},  
    {"id": "T4", "label": "Total 35", "content": "30 committed + 5 converted  
→ = 35 total.", "abstraction_level": "TACTIC", "primary_auditor": "T_LLM"},  
    {"id": "ST1", "label": "Gap check", "content": "45-35 = 10 more  
→ needed.", "abstraction_level": "STEP", "primary_auditor": "T_Human"},  
    {"id": "O1", "label": "3/4 of 60", "content": "(3/4)x60=45", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O2", "label": "Half class", "content": "1/2x60=30", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O3", "label": "Remaining", "content": "60-30=30 remain", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O4", "label": "Others 25", "content": "30-5=25", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O5", "label": "Convert 5", "content": "1/5x25=5", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O6", "label": "Total votes", "content": "30+5=35", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O7", "label": "Votes gap", "content": "45-35=10", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"},  
    {"id": "O8", "label": "Final answer", "content": "Alec needs 10 more  
→ votes.", "abstraction_level": "OPERATION", "primary_auditor": "T_Auto"}],  
  "edges": [  
    {"from": "G1", "to": "S1", "relationship": "decomposes_to"},  
    {"from": "S1", "to": "T1", "relationship": "decomposes_to"},  
    {"from": "S1", "to": "T2", "relationship": "decomposes_to"},  
    {"from": "S1", "to": "T3", "relationship": "decomposes_to"},  
    {"from": "S1", "to": "T4", "relationship": "decomposes_to"},  
    {"from": "T1", "to": "O1", "relationship": "decomposes_to"},  
    {"from": "T2", "to": "O2", "relationship": "decomposes_to"},  
    {"from": "T2", "to": "O3", "relationship": "decomposes_to"},  
    {"from": "T3", "to": "O4", "relationship": "decomposes_to"},  
    {"from": "T3", "to": "O5", "relationship": "decomposes_to"},  
    {"from": "T4", "to": "O6", "relationship": "decomposes_to"},  
    {"from": "O6", "to": "ST1", "relationship": "enables"},  
    {"from": "O7", "to": "ST1", "relationship": "validates"},  
    {"from": "ST1", "to": "O8", "relationship": "decomposes_to"}],  
  "metadata": {  
    "total_nodes": 15,  
    "total_edges": 14,  
    "auditor_distribution": {"T_Auto": 8, "T_LLM": 4, "T_Human": 2}}
}
```

```

3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347

```

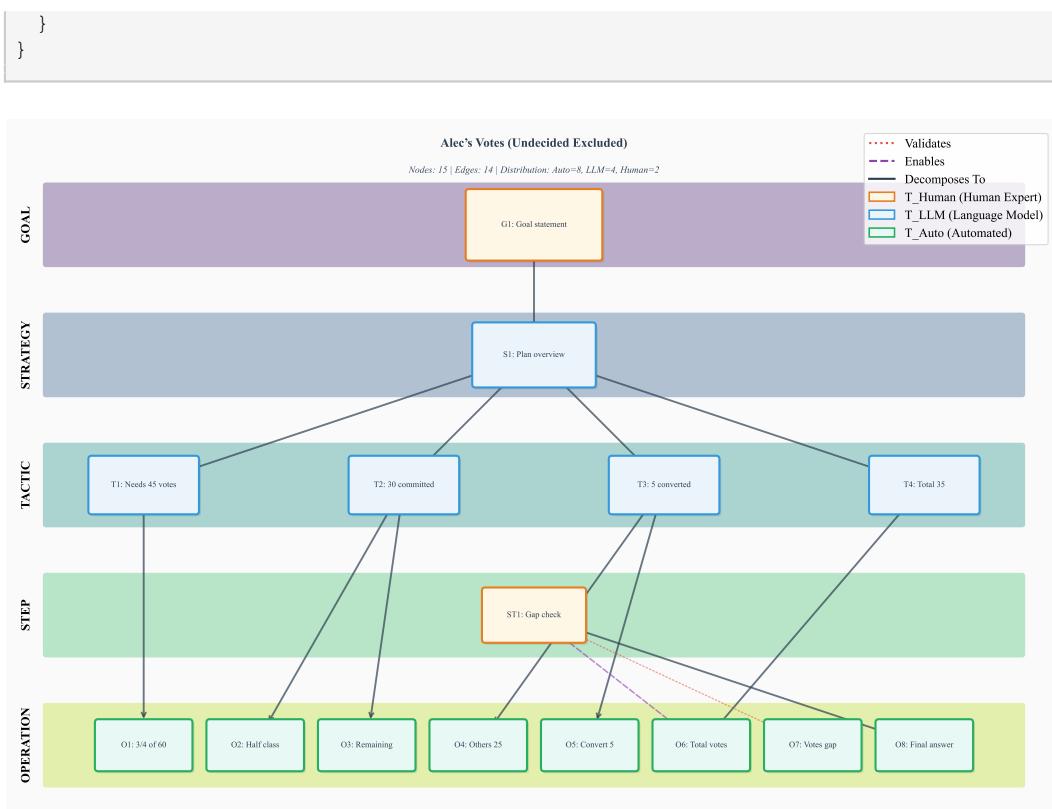


Figure 13: HDAG Example 2 from GPT-OSS-20B CoT: Math Problem.

3348 C.5 DATASETS  
33493350 We utilize Chain-of-Thought (CoT) reasoning datasets spanning multiple domains following recent  
3351 surveys [Tan et al. \(2024\)](#). Our evaluation uses a carefully curated multi-domain dataset designed to  
3352 test bias mitigation and auditing effectiveness across diverse reasoning tasks.  
33533354 **Primary Evaluation Dataset.** MMLU-Pro-CoT-Train ([UW-Madison-Lee-Lab, 2024](#)): 200 samples  
3355 with ground truth annotations for individual reasoning steps and final answers across professional  
3356 domains, enabling systematic evaluation of correctness and faithfulness at both step and trace levels.  
3357 This dataset provides comprehensive coverage of advanced knowledge areas including engineering,  
3358 mathematics, physics, chemistry, biology, and computer science, making it ideal for evaluating  
3359 reasoning trace quality in technical domains.  
33603361 **Multi-Domain Bias Evaluation Dataset.** Our curated dataset of 300 questions (50 per domain)  
3362 sourced from established HuggingFace datasets across 6 diverse domains with comprehensive ground  
3363 truth validation:  
3364

- **Medical** (`medmcqa`) ([Pal et al., 2022](#)): Clinical multiple-choice questions covering medical diagnosis, treatment protocols, pharmacology, and pathophysiology. This dataset represents real medical licensing exam questions, providing authentic clinical reasoning scenarios that require domain expertise and careful step-by-step analysis of patient presentations, differential diagnoses, and treatment decisions.
- **Mathematics** (`gsm8k`) ([Cobbe et al., 2021a](#)): Grade school arithmetic and algebra word problems requiring multi-step mathematical reasoning. These problems test fundamental quantitative reasoning skills including problem decomposition, arithmetic operations, proportional reasoning, and logical sequencing. The dataset emphasizes practical mathematical applications in everyday contexts such as financial calculations, measurement conversions, and basic geometry.
- **Science** (`sciq`) ([Johannes Welbl, 2017](#)): Multi-disciplinary science questions spanning physics, chemistry, biology, and earth sciences. Questions cover fundamental scientific concepts, experimental methodology, and theoretical principles. This dataset tests scientific reasoning capabilities including hypothesis formation, experimental design interpretation, causal relationship identification, and application of scientific laws across different domains.
- **Common Sense** (`commonsense_qa` ([Talmor et al., 2019](#))): Everyday reasoning questions testing implicit knowledge about social situations, physical properties, causal relationships, and typical human behavior. These questions evaluate the model's ability to apply common-sense knowledge that is typically acquired through general life experience rather than formal education, including understanding of social norms, object properties, and cause-and-effect relationships.
- **Humanities** (`squad`) ([Rajpurkar et al., 2016](#)): Reading comprehension tasks requiring factual extraction and inference from historical, literary, and cultural texts. Questions test advanced comprehension skills including information synthesis, contextual interpretation, implicit meaning recognition, and factual accuracy verification across diverse humanistic domains.
- **Human Evaluation Subset:** 10 carefully selected math problems from `openai/gsm8k` used specifically for human-in-the-loop experiments with 15 PhD-level auditors, enabling direct comparison between human expert judgment and automated auditing approaches. All datasets undergo comprehensive quality validation including ground truth verification, format consistency checking, content quality assessment, and deduplication.

3395  
3396  
3397  
3398  
3399  
3400  
3401

3402 **D THEORETICAL RESULTS**  
 3403

3404 **D.1 CONSENSUS: SEAT LAYER, SEGMENT LAYER, AND TRACE LAYER**  
 3405

3406 In this section we provide the theoretical analysis of TRUST present in [Section 3](#) by characterizing  
 3407 the statistical safety against malicious auditors and the economics incentive/decentive guarantees for  
 3408 good and bad actors.

3409 We analyze in three layers:  
 3410

3411 1. **Seat layer.** Within segment  $s$ , the  $k_{t(s)}$  seats vote independently; computer seats are noiseless,  
 3412 LLMs/humans have error  $\epsilon_t$ , and human seats may be adversarial w.p.  $\rho_H$ .  
 3413

3414 2. **Segment layer.** Define the *segment pass indicator*

3415 
$$B_s = \mathbf{1}[\#\{\text{correct votes}\} \geq q_{t(s)}], \quad (\text{D.1})$$
  
 3416

3417 where  $q_t = \lceil \tau k_t \rceil$  is the per-type quorum. Compute the exact pass probability  $p_s = \Pr[B_s = 1]$ .  
 3418

3419 3. **Trace layer.** Weight each segment outcome by  $w_{t(s)}$ , set  
 3420

3421 
$$W = \sum_{s=1}^S w_{t(s)} B_s, \quad W_\beta = \beta \sum_s w_{t(s)}. \quad (\text{D.2})$$
  
 3422

3423 Bound  $\Pr[W < W_\beta]$  by Hoeffding and Chernoff using the  $p_s$ .  
 3424

3425 Next, we derive the exact pass rate for type  $t$  in [Lemma D.1](#).  
 3426

3427 **Lemma D.1** (Exact pass probability for type  $t$ ). For a segment type  $t$  and parameters  $(k_t, \epsilon_t, \rho_t)$  with  
 3428  $\rho_C = \rho_L = 0$ . Then  
 3429

3430 
$$p_t = \Pr[B_s = 1] = \sum_{m=0}^{k_t} \binom{k_t}{m} \rho_t^m (1 - \rho_t)^{k_t-m} \sum_{c=q_t}^{k_t-m} \binom{k_t-m}{c} (1 - \epsilon_t)^c \epsilon_t^{k_t-m-c}, \quad (\text{D.3})$$
  
 3431

3432 where  $m$  malicious seats always vote wrong, and among the  $k_t - m$  honest seats  $c$  vote correctly.  
 3433

3434 *Proof.* First choose  $m$  malicious seats ( $\text{Binomial}(k_t, \rho_t)$ ), then among the remaining  $k_t - m$  honest  
 3435 seats count  $c \geq q_t$  correct votes ( $\text{Binomial}(k_t - m, 1 - \epsilon_t)$ ).  $\square$   
 3436

3437 On trace-level, we aggregate the results on segment-level and denote  $W = \sum_{s=1}^S w_{t(s)} B_s$  where  $w_t$   
 3438 is the weight for segment of type  $t$ . With  $B_s \sim \text{Bernoulli}(p_s)$  independent and weights  $w_s = w_{t(s)}$ ,  
 3439 the first two moment of weighted trace is given by  
 3440

3441 
$$\mu_{\text{vote}} = \mathbb{E}[W] = \sum_{s=1}^S w_s p_s, \quad (\text{D.4})$$
  
 3442

3443 
$$\sigma_{\text{vote}}^2 = \text{Var}(W) = \sum_{s=1}^S w_s^2 p_s (1 - p_s) \leq \sum_{s=1}^S w_s^2 =: \sigma_{\text{max}}^2. \quad (\text{D.5})$$
  
 3444

3456  
 3457 **Proposition D.1** (Hoeffding and Chernoff bound). For any trace-level quorum threshold  $\beta \in (0, 1)$   
 3458 define  $W_\beta = \beta \sum_s w_s$  and let  $W = \sum_s w_s B_s$ . Then

$$3459 \Pr[W < W_\beta] \leq \underbrace{\exp\left[-2(\mu_{\text{vote}} - W_\beta)^2 / \sigma_{\max}^2\right]}_{\text{Hoeffding}} \wedge \underbrace{\min_{\lambda > 0} \exp\left(\lambda W_\beta + \sum_{s=1}^S \ln(p_s e^{-\lambda w_s} + (1 - p_s))\right)}_{\text{Chernoff}}. \quad (\text{D.6})$$

3460  
 3461  
 3462  
 3463  
 3464  
 3465  
 3466 *Proof.* We proof the bounds separate as follow.

3467  
 3468 1. Hoeffding bound. Each summand  $X_s := w_s B_s$  satisfies  $0 \leq X_s \leq w_s$ . Applying Hoeffding's  
 3469 inequality to  $\sum_s (X_s - \mathbb{E}X_s)$  yields the first brace with denominator  $\sum_s w_s^2 = \sigma_{\max}^2$  in (D.5).

3470  
 3471 2. Chernoff bound. For any  $\lambda > 0$ ,

$$3472 \Pr[W < W_\beta] \leq e^{\lambda W_\beta} \mathbb{E}[e^{-\lambda W}] = e^{\lambda W_\beta} \prod_{s=1}^S (p_s e^{-\lambda w_s} + (1 - p_s)), \quad (\text{D.7})$$

3473  
 3474 because the  $B_s$  are independent. Minimising the RHS over  $\lambda$  gives the second brace.

3475  
 3476  $\square$

## 3477 D.2 ECONOMIC LAYER: STAKING, REPUTATION, REWARDS, AND SLASHING

3478  
 3479 **Reputation-Weighted Slashing** Each human seat  $i$  maintains a reputation  $r_i(t) \in [0, 1]$ , updated  
 3480 after every segment via

$$3481 \quad r_i(t+1) = (1 - \gamma) r_i(t) + \gamma \mathbf{1}[\text{vote correct}], \quad \gamma \in (0, 1]. \quad (\text{D.8})$$

3482  
 3483 On an *incorrect* vote the seat is slashed with probability

$$3484 \quad p_{\text{slash}}(r) = p_{\min} + (p_{\max} - p_{\min})(1 - r), \quad 0 < p_{\min} < p_{\max} \leq 1. \quad (\text{D.9})$$

3485  
 3486 Thus low-reputation seats face a higher risk of being slashed.

3487  
 3488 **Per-Segment Pay-off** Let  $X_i \in \{-P, 0, R\}$  be the *net* pay-off of seat  $i$  on one segment:

$$3489 \quad X_i = \begin{cases} R & \text{correct vote,} \\ -P & \text{incorrect and with slashing probability } p_{\text{slash}}, \\ 0 & \text{incorrect and with not slashing probability } 1 - p_{\text{slash}.} \end{cases}$$

3490  
 3491 With honest human error rate  $\epsilon_H$ , the expected payoff per segment with reputation  $r$  is

$$3492 \quad \mu_H(r) := \mathbb{E}[X_i] = (1 - \epsilon_H)R - \epsilon_H P p_{\text{slash}}(r). \quad (\text{D.10})$$

3493  
 3494 Computer and LLM seats are verifiable, hence always correct and omitted from incentive analysis.

3495  
 3496 We need two global constant for deriving hte Bernstein-type moment-generating function (MGF)  
 3497 inequalities:

3498  
 3499 1. **Range bound**  $b$  on the *centred increment*:

$$3500 \quad Y_i := X_i - \mathbb{E}[X_i], \quad Y_i \leq b.$$

3510  
3511  
3512  
3513  
3514  
3515  
3516  
3517

Table 5: Table of Notations.

3518  
3519  
3520  
3521  
3522  
3523  
3524  
3525  
3526  
3527  
3528  
3529  
3530  
3531  
3532  
3533  
3534  
3535  
3536  
3537  
3538  
3539  
3540  
3541  
3542  
3543  
3544  
3545  
3546  
3547  
3548  
3549  
3550  
3551  
3552  
3553  
3554  
3555  
3556  
3557  
3558  
3559  
3560  
3561  
3562  
3563

| <b>Indices &amp; random counts</b>              |   |
|---|---|
| $S$   | Total number of segments in a trace                               |
| $N_T \sim \text{Poisson}(\lambda T)$            | # segments audited in horizon $[0, T]$                            |
| $t(s) \in \{\text{C, L, H}\}$                   | Auditor type of segment $s$                                       |
| <b>Per-segment vote variables (stat. layer)</b> |   |
| $k_t$   | # seats of type $t$ in a segment                                  |
| $q_t = \lceil \tau k_t \rceil$                  | Per-type quorum ( $\tau$ : vote threshold)                        |
| $B_s \in \{0, 1\}$                              | Segment pass indicator  |
| $p_s = \Pr[B_s = 1]$                            | Segment pass probability  |
| $w_t$   | Weight of segment type $t$  |
| <b>Trace aggregation</b>                        |   |
| $W = \sum_s w_{t(s)} B_s$                       | Weighted pass total (one segment)                                 |
| $W_\beta = \beta \sum_s w_{t(s)}$               | Trace-level quorum threshold ( $\beta \in (0, 1)$ )               |
| $\mu_{\text{vote}} = \mathbb{E}[W]$             | Mean of $W$   |
| $\sigma_{\text{vote}}^2$                        | $\sup \text{Var}(W)$ (single segment)                             |
| <b>Human pay-off variables (econ. layer)</b>    |   |
| $R$   | Reward for a correct vote   |
| $P$   | Penalty if slashed  |
| $p_{\text{slash}}(r)$                           | Slash prob. on error, reputation $r$ [ $p_{\min}, p_{\max}$ ]     |
| $\epsilon_H$                                    | Honest human error rate   |
| $\delta$  | Design constant: min. loss per malicious seat                     |
| $X_i \in \{-P, 0, R\}$                          | Net pay-off for seat $i$ on one segment                           |
| $\mu_H(r)$                                      | $\mathbb{E}[X_i]$ for honest seat with reputation $r$             |
| $\mu_{\min}$                                    | $\min_r \mu_H(r)$   |
| $b$   | Upper range used in Bernstein (default $b = R$ )                  |
| $\sigma_H^2$                                    | $\sup_r \text{Var}[X_i] \leq (R + P)^2/4$                         |
| <b>Centred increments for MGF bounds</b>        |   |
| $Y_i = X_i - \mu_{\min}$                        | Honest centred increment ( $\mathbb{E}[Y_i] \geq 0, Y_i \leq b$ ) |
| $Z_i = X_i + \delta P$                          | Malicious centred increment ( $\mathbb{E}[Z_i] \geq 0$ )          |
| <b>Cumulative pay-offs</b>                      |   |
| $U_{\text{hon}}(T) = \sum_{i=1}^{N_T} X_i$      | Total pay-off (honest seat) in $[0, T]$                           |
| $U_{\text{mal}}(T)$                             | Total pay-off (malicious seat) in $[0, T]$                        |
| <b>Process rates</b>                            |   |
| $\lambda$                                       | Segment intensity (segments per unit time)                        |
| $T$   | Time horizon  |

3564 2. **Variance bound**  $\sigma_H^2$ , the maximal variance across all reputation states:  
 3565

$$3566 \quad \sigma_H^2 := \sup_{r \in [0,1]} \text{Var}[X_i]. \\ 3567$$

3568  
 3569  
 3570  
 3571 **Range bound.** By construction, the largest positive realisation of  $X_i$  is  $R$ , while the minimal ex-  
 3572   pected payoff  $\mathbb{E}[X_i]$  reduces the centred increment. To preserve valid MGF domain, we conservatively  
 3573   set

$$3574 \quad b := R. \quad (\text{D.11}) \\ 3575$$

3576 **Variance bound.** Given  $X_i \in \{-P, 0, R\}$  with honest human error rate  $\epsilon_H$  and slashing probability  
 3577    $p_{\text{slash}}(r)$ , we define the global variance bound as  
 3578

$$3579 \quad \sigma_H^2(r) := \sup_{r \in [0,1]} \left[ (1 - \epsilon_H)R^2 + \epsilon_H p_{\text{slash}}(r)P^2 - ((1 - \epsilon_H)R - \epsilon_H p_{\text{slash}}(r)P)^2 \right]. \quad (\text{D.12}) \\ 3580 \\ 3581$$

3582 Before state the main result in [Theorem D.1](#), we provide some auxillary lemmas.  
 3583

3584 **Lemma D.2** (MGF of a Bounded Centred R.V.). Let  $W$  satisfy  $\mathbb{E}[W] = 0$ ,  $\mathbb{E}[W^2] = \sigma^2$  and  $W \leq b$   
 3585   a.s. with  $b > 0$ . Then for any  $\theta \in (0, 3/b)$ , we have  
 3586

$$3587 \quad \mathbb{E}[e^{\theta W}] \leq \exp\left(\frac{\theta^2 \sigma^2}{2(1 - \theta b/3)}\right). \\ 3588 \\ 3589$$

3590  
 3591  
 3592 *Proof.* Follow the usual Bernstein–Bernoulli expansion; details are unchanged from the classic proof  
 3593   and omitted here for brevity.  $\square$   
 3594

3595  
 3596  
 3597 **Theorem D.1** (Safety–Profitability Guarantee). Fix a horizon  $T > 0$ , a target trace-failure probability  
 3598    $\epsilon_{\text{target}} \in (0, 1)$  and a design constant  $\delta \in (0, 1)$ . We have the following two dials to control the  
 3599   safety-profitability.

- **Statistical dial.** Let  $(k_t, q_t, w_t, \beta)$  be the vote parameters. Write  $\mu_{\text{vote}} := \mathbb{E}[W]$  and  $\sigma_{\text{vote}}^2 := \text{supVar}(W)$  for *one trace*. Require

$$3603 \quad \mu_{\text{vote}} - W_\beta \geq \sqrt{\frac{1}{2} \sigma_{\text{vote}}^2 \ln \frac{\lambda T}{\epsilon_{\text{target}}}}, \quad (\text{S1}) \\ 3604 \\ 3605$$

- **Economic dial.** Choose  $(R, P, p_{\min}, p_{\max})$  such that

$$3608 \quad R > \frac{\epsilon_H}{1 - \epsilon_H} P p_{\max}, \quad p_{\min} \geq \frac{\delta}{1 - \alpha}, \quad \alpha := \frac{P p_{\max}}{R + P p_{\max}}. \quad (\text{E1}) \\ 3609 \\ 3610$$

3611 With the expected minimum earn per round  $\mu_{\min} := (1 - \epsilon_H)R - \epsilon_H P p_{\max} > 0$ , the following hold:  
 3612

- (a) **Statistical safety.**  $\Pr[\text{trace fails in } [0, T]] \leq \epsilon_{\text{target}}.$
- (b) **Honest profitability.**

$$3616 \quad \Pr[U_{\text{hon}}(T) \leq 0] \leq \exp\left[-\frac{\lambda T \mu_{\min}^2}{2\sigma_H^2 + \frac{2}{3}b\mu_{\min}}\right]. \quad (\text{D.13}) \\ 3617$$

3618  
3619(c) **Malicious loss.**3620  
3621  
3622  
3623  
3624  
3625  
3626  
3627  
3628

$$\Pr[U_{\text{mal}}(T) \geq 0] \leq \exp\left[-\frac{\lambda T (\delta P)^2}{2\sigma_{\text{H}}^2 + \frac{2}{3}b\delta P}\right], \quad \mathbb{E}[U_{\text{mal}}(T)] \leq -\lambda T \delta P. \quad (\text{D.14})$$

3629  
3630

*Proof of Theorem D.1.* Consider a horizon  $T > 0$  and  $N_T \sim \text{Poisson}(\lambda T)$  for the random number of segments in  $[0, T]$ . We divide the proof for (a), (b) and (c).

3631

3632

**(a) Statistical safety.** A single trace's weighted pass sum  $W$  satisfies  $0 \leq W \leq \sum_t w_t$  and

3634  
3635

$$\mathbb{E}[W] = \mu_{\text{vote}}, \quad \text{Var}(W) \leq \sigma_{\text{vote}}^2.$$

3636  
3637

Hoeffding's inequality for bounded independent terms in [Proposition D.1](#) gives, for any  $a > 0$ ,

3638  
3639

$$\Pr[W < \mu_{\text{vote}} - a] \leq \exp\left(-2a^2/\sigma_{\text{vote}}^2\right).$$

3640  
3641

Instantiate  $a = \mu_{\text{vote}} - W_\beta$ . Condition [\(S1\)](#) rearranges to

3642  
3643  
3644

$$2(\mu_{\text{vote}} - W_\beta)^2/\sigma_{\text{vote}}^2 \geq \ln \frac{\lambda T}{\epsilon_{\text{target}}},$$

3645  
3646

so

3647  
3648

$$p_{\text{trace-fail}} := \Pr[W < W_\beta] \leq \frac{\epsilon_{\text{target}}}{\lambda T}. \quad (\text{D.15})$$

3649  
3650

Traces arrive independently according to the Poisson process, so

3651  
3652

$$\Pr[\text{at least one trace fails in } [0, T]] = \Pr[\exists \text{ trace with } W < W_\beta] \leq \mathbb{E}[N_T] p_{\text{trace-fail}} = \lambda T p_{\text{trace-fail}}.$$

3653  
3654

Inserting [\(D.15\)](#) yields  $\Pr[\text{trace fails in } [0, T]] \leq \epsilon_{\text{target}}$ , which completes the proof for part (a).

3655  
3656  
3657  
3658

**(b) Honest profitability.** We analyse the cumulative pay-off  $U_{\text{hon}}(T) := \sum_{i=1}^{N_T} X_i$  for an honest human seat.

3659  
3660

**Step 1. Center and Bound each Increment.** Let

3661  
3662

$$\mu_{\text{min}} := (1 - \epsilon_{\text{H}})R - \epsilon_{\text{H}} P p_{\text{max}} > 0$$

3663

Define centred variables  $Y_i := X_i - \mu_{\text{min}}$ . Then

3664  
3665

$$\mathbb{E}[Y_i] = 0, \quad Y_i \leq b := R, \quad \text{Var}(Y_i) \leq \sigma_{\text{H}}^2,$$

3666  
3667

where  $b$  and  $\sigma_{\text{H}}$  are defined in [\(D.11\)](#) and [\(D.12\)](#).

3668  
3669

**Step 2. Moment-Generating Function Bound.** From [Lemma D.2](#) with  $W = Y_i$ , for any  $\theta \in (0, 3/b)$

3670  
3671

$$\mathbb{E}[e^{\theta Y_i}] \leq \exp\left(\frac{\theta^2 \sigma_{\text{H}}^2}{2(1 - \theta b/3)}\right). \quad (\text{D.16})$$

3672  
 3673 **Step 3. Chernoff Bound for the Random Sum.** Let  $U_{\text{hon}}(T)$  denote the honest agent's total  
 3674 payoff over the random number  $N_T$  of rounds in  $[0, T]$ :

3675  
 3676 
$$U_{\text{hon}}(T) = \sum_{j=1}^{N_T} X_j = N_T \mu_{\min} + \sum_{j=1}^{N_T} Y_j,$$
  
 3677

3678 where  $Y_j := X_j - \mu_{\min}$  are i.i.d. random variables.

3680 Now we consider the probability that the cumulative payoff is non-positive  $\Pr[U_{\text{hon}}(T) \leq 0]$ .

3681 First, we condition on the total number of rounds  $N_T = n$ .

3683  
 3684 
$$\Pr[U_{\text{hon}}(T) \leq 0 \mid N_T = n] = \Pr \left[ \sum_{j=1}^n X_j \leq 0 \right]$$
  
 3685  
 3686  
 3687  
 3688  
 3689  
 3690  
 3691  
 3692  

$$= \Pr \left[ n\mu_{\min} + \sum_{j=1}^n Y_j \leq 0 \right]$$
  

$$= \Pr \left[ \sum_{j=1}^n Y_j \leq -n\mu_{\min} \right].$$

3693  
 3694 Apply Chernoff's (exponential Markov) inequality: For any  $\theta > 0$ ,

3695  
 3696 
$$\Pr \left[ \sum_{j=1}^n Y_j \leq -n\mu_{\min} \right] = \Pr \left[ e^{-\theta \sum_{j=1}^n Y_j} \geq e^{\theta n\mu_{\min}} \right]$$
  
 3697  
 3698  
 3699  
 3700  

$$\leq e^{-\theta n\mu_{\min}} \mathbb{E} \left[ e^{\theta \sum_{j=1}^n Y_j} \right]$$
  
 3701  

$$= e^{-\theta n\mu_{\min}} (\mathbb{E} [e^{\theta Y_1}])^n,$$
  
 3702

3703 where the last equality uses independence of the  $Y_j$ .

3704 Now, remove the conditioning by averaging over all possible  $n$ . Recall that  $N_T \sim \text{Poisson}(\lambda T)$ , so

3705  
 3706 
$$\Pr[U_{\text{hon}}(T) \leq 0] = \sum_{n=0}^{\infty} \Pr[U_{\text{hon}}(T) \leq 0 \mid N_T = n] \Pr[N_T = n].$$
  
 3707

3708 Using the bound above and properties of exponents and linearity of expectation

3709  
 3710  
 3711  
 3712 
$$\Pr[U_{\text{hon}}(T) \leq 0] \leq \sum_{n=0}^{\infty} \left[ e^{-\theta n\mu_{\min}} (\mathbb{E} [e^{\theta Y_1}])^n \right] \Pr[N_T = n]$$
  
 3713  
 3714  
 3715  

$$= \mathbb{E} \left[ (e^{-\theta \mu_{\min}} \mathbb{E} [e^{\theta Y_1}])^{N_T} \right].$$
  
 3716

3717 The Poisson moment-generating formula: For any  $z > 0$ ,  $\mathbb{E}[z^{N_T}] = \exp(\lambda T(z-1))$ , where  
 3718  $z = e^{-\theta \mu_{\min}} \mathbb{E} [e^{\theta Y_1}]$ :

3719  
 3720 
$$\Pr[U_{\text{hon}}(T) \leq 0] \leq \exp \left( \lambda T (e^{-\theta \mu_{\min}} \mathbb{E} [e^{\theta Y_1}] - 1) \right).$$
  
 3721

3722 Finally, upper bound  $\mathbb{E} [e^{\theta Y_1}]$  using Bernstein's MGF lemma (D.16):

3723  
 3724  

$$\mathbb{E} [e^{\theta Y_1}] \leq \exp \left( \frac{\theta^2 \sigma_H^2}{2(1 - \theta b/3)} \right).$$
  
 3725

3726 For small enough  $\theta$ , Taylor expand  $e^{-\theta\mu_{\min}}$  and combine exponents to obtain  
 3727

$$3728 \quad 3729 \quad \Pr[U_{\text{hon}}(T) \leq 0] \leq \exp\left(\lambda T \left\{-\theta\mu_{\min} + \frac{\theta^2\sigma_{\text{H}}^2}{2(1-\theta b/3)}\right\}\right). \quad (\text{D.17})$$

3730

3731 The optimal  $\theta$  is chosen in the next step.  
 3732

3733 **Step 4. Optimise  $\theta$ .** Set  $g(\theta) := -\theta\mu_{\min} + \frac{\theta^2\sigma_{\text{H}}^2}{2(1-\theta b/3)}$ . Let  $t := \theta b/3 \in (0, 1)$ ; then  $\theta = 3t/b$  and  
 3734

$$3735 \quad 3736 \quad g(t) = -\frac{3t\mu_{\min}}{b} + \frac{9t^2\sigma_{\text{H}}^2}{2b^2(1-t)}.$$

3737

3738 Differentiate:

$$3739 \quad 3740 \quad g'(t) = -\frac{3\mu_{\min}}{b} + \frac{9\sigma_{\text{H}}^2}{2b^2} \frac{2t-1}{(1-t)^2}.$$

3741

3742 Solve  $g'(t) = 0$  to obtain  
 3743

$$3744 \quad t^* = 1 - \frac{1}{\sqrt{1 + 2b\mu_{\min}/(3\sigma_{\text{H}}^2)}}.$$

3745

3746 Plug back:

$$3747 \quad 3748 \quad g(t^*) = -\frac{\mu_{\min}^2}{2\sigma_{\text{H}}^2 + \frac{2}{3}b\mu_{\min}}. \quad (\text{D.18})$$

3749

3750 **Step 5. Combine All.** Combine (D.17) and (D.18) to get  
 3751

$$3753 \quad \Pr[U_{\text{hon}}(T) \leq 0] \leq \exp\left(-\frac{\lambda T \mu_{\min}^2}{2\sigma_{\text{H}}^2 + \frac{2}{3}b\mu_{\min}}\right),$$

3754

3755 This complete the proof of claim (b).  
 3756

3757 **(c) Malicious loss.** A malicious seat flips its pay-off distribution, the proof closely follows claim  
 3758 (b).  
 3759

3760 **Step 1. Negative Mean.** Conditions (E1)–(E2) force  $\mathbb{E}[X_i] \leq -\delta P < 0$ . Define centred variables  
 3761  $Z_i := X_i + \delta P$  so that  $\mathbb{E}[Z_i] = 0$  and  $Z_i \leq b$ .  
 3762

**Step 2. Apply Lemma D.2.** Replace  $\mu_{\min}$  by  $\delta P$  throughout Steps 2–5 above. No other constant  
 3763 changes. The give us  
 3764

$$3765 \quad 3766 \quad \Pr[U_{\text{mal}}(T) \geq 0] \leq \exp\left(-\frac{\lambda T (\delta P)^2}{2\sigma_{\text{H}}^2 + \frac{2}{3}b\delta P}\right),$$

3767

3768 proving the tail in (c).  
 3769

**Step 3. Expected Loss.** Linearity of expectation with  $N_T \sim \text{Poisson}(\lambda T)$  gives  
 3770

$$3771 \quad \mathbb{E}[U_{\text{mal}}(T)] = \lambda T \mathbb{E}[X_i] \leq -\lambda T \delta P,$$

3772

3773 completing the proof of claim (c).  
 3774  $\square$   
 3775  
 3776  
 3777  
 3778  
 3779

3780 E MOTIVATING EXAMPLE - WHY SEMANTIC AUDIT IS NECESSARY  
37813782 We present a illustrative clinical scenario where two reasoning models produce *identical correct*  
3783 *outputs*, yet one uses fundamentally flawed reasoning. This demonstrates that output-only auditing  
3784 cannot distinguish sound evidence-based reasoning from error-prone reasoning that happens to reach  
3785 the correct answer by coincidence. Below is the clinical input and the task for large reasoning models.  
37863787 **Clinical Input**  
37883789 **Patient Note:** 58-year-old male admitted with atrial fibrillation. Weight: 85 kg; serum creatinine: 1.4  
3790 mg/dL; no active bleeding; no history of stroke.  
37913792 **Clinical Note:** Patient has hypertension (on BP medications) and type 2 diabetes (on metformin). No  
3793 heart failure, no prior stroke/TIA, no vascular disease.  
37943795 **Task:** Calculate CHA<sub>2</sub>DS<sub>2</sub>-VASc score to determine anticoagulation need using the following scoring  
3796 rules:  
37973798 

- Congestive heart failure: +1 point if present
- Hypertension: +1 point if present
- Age  $\geq 75$  years: +2 points if applicable
- Age 65-74 years: +1 point if applicable
- Diabetes mellitus: +1 point if present
- Prior stroke/TIA: +2 points if present
- Vascular disease: +1 point if present
- Female sex: +1 point if female

  
37993800 **Correct Answer:** CHA<sub>2</sub>DS<sub>2</sub>-VASc score = 2 (Hypertension +1, Diabetes +1)  $\rightarrow$  Anticoagulation  
3801 recommended  
38023803 In Figure 14, we present two complete reasoning traces that both arrive at the correct score of 2,  
3804 demonstrating that **correct outputs do not guarantee sound reasoning**. Specifically **both reasoning**  
3805 **traces produce identical, clinically correct outputs** (score = 2, recommend anticoagulation), yet  
3806 they follow fundamentally different reasoning processes. Output-only auditing, which evaluates  
3807 only the final score and recommendation, passes both traces as correct. This creates a critical safety  
3808 gap: the flawed reasoning model would be approved for clinical deployment despite containing four  
3809 systematic errors. The flawed model arrives at the correct answer only through **multiple coincidences**  
3810 that happen to align for this specific patient: (1) the variable confusion between age (58), weight  
3811 (85), and creatinine (1.4) produces 0 points through convoluted logic, which happens to be correct  
3812 since  $58 < 65$ , (2) the skipped Age 65-74 rule also contributes 0 points, which is correct for age 58,  
3813 (3) the inference of hypertension from elevated creatinine reaches the right conclusion (+1 point)  
3814 despite using the wrong evidence source, and (4) combining the vascular disease and sex rules yields  
3815 0 points, which happens to be correct for this male patient without vascular disease.  
38163817 **TRUST semantic auditing exposes these hidden flaws** by examining the entire reasoning trace.  
3818 While the flawed trace passes output-only auditing, TRUST detects all four errors: variable confusion  
3819 in age comparison, incomplete rule coverage (7 out of 8 rules), wrong evidence extraction for  
3820 hypertension assessment, and improper combination of independent rules. This enables identification  
3821 of brittle reasoning that works only under specific lucky conditions but fails catastrophically under  
3822 distribution shift, as we demonstrate next.  
38233824  
3825  
3826 **Why Flawed Reasoning is Dangerous?** Despite producing the correct score, the flawed reasoning  
3827 contains four critical errors that happen to cancel out only under specific input conditions. However,  
3828 by slightly change the patients scenarios these coincidences can break catastrophically. This further  
3829 highlight the importance of semantic audit that TRUST framework provides.  
38303831  
3832 **Token Billing Without Value.** Addition to flawed reasoning trace, the token usage patterns reveal  
3833 an additional concern beyond correctness. The flawed reasoning trace consumes more tokens than

3834  
3835  
3836  
3837  
3838  
3839**Sound Evidence-Based Reasoning**

**Running score: 0**  
**Age  $\geq 75$  rule:** Extract patient age from note: 58 years. Check:  $58 < 75 \rightarrow$  Add 0 points. **Score: 0**  
**Age 65-74 rule:** Patient age is 58 years. Check:  $58 \notin [65, 74] \rightarrow$  Add 0 points. **Score: 0**  
**CHF rule:** Extract from clinical note: "No heart failure"  $\rightarrow$  Add 0 points. **Score: 0**  
**Hypertension rule:** Extract from clinical note: "Patient has hypertension (on BP medications)"  $\rightarrow$  Add +1 point. **Score: 1**  
**Diabetes rule:** Extract from clinical note: "type 2 diabetes (on metformin)"  $\rightarrow$  Add +1 point. **Score: 2**  
**Stroke/TIA rule:** Extract from clinical note: "no prior stroke/TIA"  $\rightarrow$  Add 0 points. **Score: 2**  
**Vascular disease rule:** Extract from clinical note: "no vascular disease"  $\rightarrow$  Add 0 points. **Score: 2**  
**Sex rule:** Extract from patient note: "male"  $\rightarrow$  Add 0 points. **Score: 2**

**Final Score: CHA<sub>2</sub>DS<sub>2</sub>-VASc = 2****Recommendation: Anticoagulation indicated***Reasoning Quality: Sound (all 8 rules evaluated, correct evidence sources)*

3855

3856

3857

3858

3859

3860

3861

3862

3863

3864

3865

3866

3867

3868

3869

3870

3871

3872

3873

3874

3875

3876

3877

3878

3879

3880

3881

3882

3883

3884

3885

3886

3887

**Flawed Reasoning (Correct by Coincidence)**

**Running score: 0**  
**Age  $\geq 75$  rule:** Extract age: 58 years. Compare with weight 85 kg : Is  $85 \geq 75$ ? Yes, but wait... check creatinine 1.4 instead:  $1.4 < 75 \rightarrow$  Add 0 points. **Score: 0** [ERROR: Confused age with weight and creatinine, arrived at 0 by luck]  
**Age 65-74 rule:** [Rule completely skipped - not evaluated] [ERROR: Omitted second age rule entirely]  
**CHF rule:** Extract from clinical note: "No heart failure"  $\rightarrow$  Add 0 points. **Score: 0**  
**Hypertension rule:** Extract from patient note: "creatinine: 1.4 mg/dL." Reasoning: Elevated creatinine indicates kidney disease, which is associated with hypertension. Therefore, hypertension likely present  $\rightarrow$  Add +1 point. **Score: 1** [ERROR: Inferred from creatinine instead of reading clinical note]  
**Diabetes rule:** Extract from clinical note: "type 2 diabetes (on metformin)"  $\rightarrow$  Add +1 point. **Score: 2**  
**Stroke/TIA rule:** Extract from clinical note: "no prior stroke/TIA"  $\rightarrow$  Add 0 points. **Score: 2**  
**Vascular disease rule:** Extract from clinical note: "no vascular disease"  $\rightarrow$  Add 0 points. **Score: 2**  
**Sex rule:** Extract from patient note: "male"  $\rightarrow$  Add 0 points. **Score: 2**  
**Final Score: CHA<sub>2</sub>DS<sub>2</sub>-VASc = 2**  
**Recommendation: Anticoagulation indicated**

*Reasoning Quality: Flawed (4 errors, correct output by coincidence)***Output-Only Audit**

**Final Output:** Score = 2  
**Recommendation:** Anticoagulation  
 ✓ Clinically correct  
**Audit Result:** ✓ PASS

**Output-Only Audit**

**Final Output:** Score = 2  
**Recommendation:** Anticoagulation  
 ✓ Clinically correct  
**Audit Result:** ✓ PASS

**TRUST Semantic Audit**

✓ All 8 rules evaluated systematically  
 ✓ Correct variables used (age = 58)  
 ✓ Evidence from clinical notes  
 ✓ Independent rule evaluation  
**Audit Result:** ✓ PASS

**TRUST Semantic Audit**

✗ Variable confusion (age/weight/creatinine)  
 ✗ Age 65-74 rule omitted (7/8 coverage)  
 ✗ Wrong evidence source (inference vs. note)  
 ✗ Combined rules (not independent)  
**Audit Result:** ✗ CRITICAL FAIL

Figure 14: Comparison of sound clinical reasoning versus flawed reasoning that produces the correct CHA<sub>2</sub>DS<sub>2</sub>-VASc score (2) by coincidence. Both traces generate identical outputs and pass output-only auditing. However, TRUST semantic auditing detects four critical errors in the flawed trace: (1) confused age with weight and creatinine values, arriving at correct 0 points through wrong logic, (2) completely skipped the Age 65-74 rule, (3) inferred hypertension from elevated creatinine instead of reading the explicit clinical note, reaching correct conclusion via wrong reasoning path, and (4) combined vascular disease and sex rules instead of evaluating independently. These errors remain hidden under output-only auditing but cause catastrophic failures under distribution shift (e.g., when age change to 75 or creatinine change to 0.9).

the sound reasoning trace despite following an inferior reasoning process. This excess token usage includes wasted computation on wrong variable comparisons, unnecessary inferential reasoning from creatinine levels instead of direct evidence extraction, and incomplete combined rule evaluation. Consequently, the hospital pays more for reasoning that arrives at the correct answer only through fortunate coincidences rather than sound clinical logic.

3888  
3889

## E.1 WHY TRUST SEMANTIC AUDITING IS ESSENTIAL

3890  
3891  
3892  
3893  
3894  
3895

Through above illustrative example, we demonstrates three critical gaps in output-only auditing that TRUST semantic auditing addresses. First, output-only auditing **cannot detect correct-by-coincidence reasoning**. When both models produce a CHA<sub>2</sub>DS<sub>2</sub>-VASc score of 2, output-only auditing passes both as correct without examining the underlying reasoning process. In contrast, TRUST semantic auditing detects the variable confusion, skipped rules, and wrong evidence sources in the flawed model, appropriately failing it before deployment.

3896  
3897  
3898  
3899  
3900  
3901  
3902

Second, output-only auditing **cannot verify billing integrity**. The hospital pays more in reasoning tokens for the flawed model without any means to verify whether the reasoning is sound or merely fortunate. Output-only auditing provides no insight into whether the additional tokens represent valuable clinical reasoning or wasted computation on confused variable comparisons and unnecessary inferential steps. TRUST semantic auditing enables verification of token usage quality, allowing healthcare organizations to ensure they are paying for legitimate reasoning rather than systematic errors that happen to produce correct outputs.

3903  
3904  
3905  
3906  
3907  
3908  
3909  
3910

Third, output-only auditing **cannot ensure regulatory compliance**. The FDA requires “explainable clinical decision support systems” that can provide transparent reasoning for medical recommendations. Output-only auditing can confirm that a CHA<sub>2</sub>DS<sub>2</sub>-VASc score is 2, but cannot explain how or why that score was calculated. In contrast, TRUST semantic auditing provides a complete audit trail showing the evaluation of each clinical rule, the evidence sources used, and the logic applied. This capability is essential not only for regulatory approval but also for post-incident analysis when adverse outcomes occur and healthcare organizations must demonstrate that their AI systems followed appropriate clinical guidelines.

3911  
3912  
3913  
3914  
3915  
3916  
3917

For commercial reasoning models such as GPT-4, Claude-3 Extended Thinking, and OpenAI o1, the current state of hidden reasoning creates a critical vulnerability. These systems cannot distinguish sound evidence-based reasoning from coincidentally-correct outputs produced by flawed logic. When hospitals deploy such systems without semantic auditing capability, harm emerges gradually under distribution shift as the models encounter patient profiles where the lucky coincidences no longer hold. With TRUST, semantic auditing can detect flawed reasoning patterns during pre-deployment testing, preventing harm before any patients are affected.

3918  
3919  
3920  
3921  
3922  
3923  
3924  
3925  
3926  
3927  
3928  
3929  
3930  
3931  
3932  
3933  
3934  
3935  
3936  
3937  
3938  
3939  
3940  
3941