BEYOND CROWD LABELS: LEARNING PROCESS SUPERVISION FROM CROWD REASONING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Collecting high-quality supervision data is critical to the development of modern AI systems, and crowdsourcing provides a scalable way to obtain such data from multiple workers. Classical learning-from-crowds methods recover reliable labels from noisy worker annotations over predefined response options. Recent advances in large language models have increased demand for supervision that captures not only *what* the correct answer is but also *how* it is derived, which motivates the crowdsourced collection of free-form reasoning trajectories. In this paper, we ask whether, when crowd workers provide free-form reasoning trajectories instead of annotations over predefined options, reliable process supervision can still be recovered directly from noisy and heterogeneous worker contributions. To answer this question, we propose a structured inference framework that leverages shared reasoning structure across heterogeneous worker trajectories. We first construct a task-specific latent consensus process and softly align each worker trajectory to this process to establish cross-trajectory correspondences. We then perform structured variational inference over the aligned trajectories and observed final outcomes to recover the validity of individual reasoning steps. Evaluations with both LLM and human crowdworkers show that reliable step-level validity signals can be recovered from heterogeneous crowd reasoning trajectories and verified final outcomes without external step-level evaluators. The recovered signals capture meaningful variation in intermediate reasoning validity and provide useful supervision for downstream LLM fine-tuning across multiple reasoning domains.

1 INTRODUCTION

Collecting high-quality supervision data plays a critical role in the development of modern AI systems, as the quality of the collected data directly affects what models can learn and how well they perform on downstream tasks (Whang et al., 2023). Crowdsourcing provides a scalable way to collect such supervision data from a large and diverse pool of crowd workers (Snow et al., 2008; Zhang et al., 2016). In typical crowdsourcing, the same annotation task is assigned to multiple workers, who independently provide a label or answer from the same predefined set of response options. Since workers may differ in expertise and make mistakes, the annotations can be noisy and inconsistent, which motivates a large body of research on learning from crowds that seeks to recover reliable latent ground-truth labels from multiple noisy worker annotations (Dawid & Skene, 1979).

Recent advances in large language models (LLMs) have increased demand for supervision that captures not only *what* the correct answer is but also *how* it is derived, with high-quality reasoning trajectories shown to improve LLM reasoning on mathematical problem solving (Singh et al., 2024) and logical reasoning (Hsieh et al., 2023; Dalvi et al., 2021). When collected through crowdsourcing, reasoning trajectories from different workers solving the same question can vary substantially in length, reasoning granularity, and solution strategy. For many reasoning tasks, final outcomes can be readily verified, but such outcome-level feedback is too coarse to characterize the intermediate reasoning process, as a trajectory may reach the correct answer despite containing flawed steps or end with an incorrect answer despite containing valid intermediate reasoning (Uesato et al., 2022; Xia et al., 2025). Obtaining reliable step-level feedback therefore rely on either human step-level annotations, which require substantial annotation effort (Lightman et al., 2024), or model-based evaluators such as LLM judges, whose reliability depends on access to a model already capable of identifying reasoning errors in the target tasks (Jacovi et al., 2024), even though such a model may

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

not yet be available when collecting data to develop those capabilities (Zheng et al., 2025). In this paper, we ask whether, when crowdworkers provide free-form reasoning trajectories instead of predefined response options, reliable process supervision can still be recovered directly from noisy and heterogeneous worker contributions without relying on an external step-level evaluator.

Answering this question is challenging because, unlike annotations over predefined options, intermediate reasoning steps from different workers have no predefined correspondence, as related reasoning steps may appear at different positions and granularities or follow different solution strategies. In addition, step validity may change as reasoning progresses, since intermediate errors can arise or be corrected at later steps (Yang et al., 2025). To address these challenges, we propose a structured inference framework that leverages shared reasoning structure across heterogeneous worker trajectories. We first represent each worker trajectory through its intermediate reasoning content and local transition structure, and construct a task-specific latent consensus process that serves as a shared reference across heterogeneous trajectories. We then softly align each worker trajectory to this latent process, allowing related reasoning to be matched across different positions and granularities. Based on these cross-trajectory correspondences, we then develop a structured variational inference framework to recover the validity of each reasoning step, which jointly estimates the shared latent validity landscape and the validity sequence of each worker trajectory from the aligned trajectories and observed final outcomes. To evaluate the effectiveness of the proposed framework, we first conduct experiments with heterogeneous LLM crowdworkers across five reasoning task sources. Our results show that reliable step-level validity signals can be recovered directly from crowd reasoning trajectories and verified final outcomes without relying on external step-level supervision. We then use the recovered step-validity signals as supervision for LLM fine-tuning and find that they consistently improve downstream reasoning performance. Finally, we evaluate the framework on reasoning trajectories collected from real human crowdworkers, where the inferred signals recover expert-annotated step validity both overall and within individual workers and capture meaningful variation in workers' reasoning dynamics.

2 METHOD

2.1 PROBLEM FORMULATION

Classical crowdsourcing infers a latent task label from multiple worker-provided categorical labels while accounting for worker reliability (Zheng et al., 2017; Raykar et al., 2010; Whitehill et al., 2009). In contrast, in our setting, each worker contributes a free-form reasoning trajectory together with a committed final answer. Our goal is to infer the latent validity of each reasoning step from these noisy, heterogeneous trajectories without relying on step-level expert annotation. Let $\{x_i\}_{i=1}^N$ denote N tasks and let $\mathcal{W}_i \subseteq \{1, \dots, M\}$ denote the workers contributing to task x_i. For each task x_i, worker $j \in \mathcal{W}_i$ provides a reasoning trajectory $\tau_{ij} = (s_{ij}^1, \dots, s_{ij}^{T_{ij}})$ and a committed final answer y_{ij}, where $s_{ij}^t \in \mathcal{S}$ is the t-th reasoning step, $\mathcal{S}$ denotes the space of free-form textual reasoning steps, and T_{ij} is the trajectory length. Each reasoning step s_{ij}^t is associated with a latent validity variable $g_{ij}^t \in \{0, 1\}$, where $g_{ij}^t = 1$ indicates that, after step t, the current reasoning state is logically consistent with at least one correct solution for x_i. Here, the current reasoning state reflects the claims and intermediate conclusions that remain in effect after incorporating the reasoning steps $s_{ij}^1, \dots, s_{ij}^t$. Since a subsequent step may correct or replace an earlier erroneous statement, g_{ij}^t may transit from $0 \to 1$ at a later step. The final answer y_{ij} is evaluated by an outcome verifier V_i, which returns $o_{ij} = V_i(y_{ij}) \in \{0, 1\}$, where $o_{ij} = 1$ indicates the final answer is correct. Let $\mathcal{D}_{\text{proc}} = \{(x_i, \{\tau_{ij}, o_{ij}\}_{j \in \mathcal{W}_i})\}_{i=1}^N$ denote the collected dataset and $\boldsymbol{g}_{ij} = (g_{ij}^1, \dots, g_{ij}^{T_{ij}})$ the latent validity sequence. With $\boldsymbol{g}_i = \{\boldsymbol{g}_{ij}\}_{j \in \mathcal{W}_i}$ and $\boldsymbol{o}_i = \{o_{ij}\}_{j \in \mathcal{W}_i}$, by Bayes' rule and assuming task-wise conditional independence, we formulate the joint posterior of the latent validity sequences $\{\boldsymbol{g}_i\}_{i=1}^N$ and worker reliability parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j\}_{j=1}^M$ as:

$$\mathbb{P}\Big(\{\boldsymbol{g}_i\}_{i=1}^N, \boldsymbol{\theta} \mid \mathcal{D}_{\text{proc}}\Big) \propto \mathbb{P}(\boldsymbol{\theta}) \prod_{i=1}^N \mathbb{P}(\boldsymbol{g}_i, \boldsymbol{o}_i \mid x_i, \{\tau_{ij}\}_{j \in \mathcal{W}_i}, \boldsymbol{\theta}) \tag{1}$$

Our objective is to infer the posterior marginal probability $\mathbb{P}(g_{ij}^t = 1 \mid \mathcal{D}_{\text{proc}})$ associated with each reasoning step. Sections 2.2–2.3 specify the joint term in Equation 1 by introducing a shared latent validity landscape that couples the validity sequences across workers, and Section 2.4 derives a tractable variational approximation to these posterior marginals.

2.2 Representation of Reasoning Trajectories and Latent Consensus

Directly optimizing Equation 1 requires comparing free-form reasoning steps drawn from heterogeneous trajectories that differ in length, granularity, and solution strategy. Because these steps lie in an unstructured textual space $\mathcal{S}$, no explicit correspondence between steps from different workers is available. To enable such comparisons, we first use a frozen pretrained text encoder f_ϕ [1] to construct a claim embedding $\boldsymbol{e}_{ij}^{\text{claim},t} = f_\phi(x_i, s_{ij}^t) \in \mathbb{R}^d$, which represents the semantic claim expressed by step s_{ij}^t. To distinguish semantically similar claims that play different roles in the reasoning process, following Wang et al. (2025), we then construct a transition embedding $\boldsymbol{e}_{ij}^{\text{trans},t} = f_\phi\big(x_i,\ [\text{prompt}],\ s_{ij}^{t-1},\ s_{ij}^t\big) \in \mathbb{R}^d$ [2]. We concatenate the two embeddings to obtain the step representation $\boldsymbol{r}_{ij}^t = [\boldsymbol{e}_{ij}^{\text{claim},t}; \boldsymbol{e}_{ij}^{\text{trans},t}] \in \mathbb{R}^{2d}$ and stack all step representations into $\boldsymbol{R}_{ij} = [\boldsymbol{r}_{ij}^1, \ldots, \boldsymbol{r}_{ij}^{T_{ij}}]^\top \in \mathbb{R}^{T_{ij} \times 2d}$. We further encode the directed transition structure of each trajectory using $\boldsymbol{\Delta}_{ij} \in \{0,1\}^{T_{ij} \times T_{ij}}$, where $[\boldsymbol{\Delta}_{ij}]_{tt'} = \mathbb{I}[t' = t+1]$. Although $(\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij})$ provides a structured representation of each individual trajectory, it does not establish cross-worker correspondence because related steps may occur at different positions or granularities under different solution strategies. To establish such correspondence, we introduce a task-specific *latent consensus process structure* $\Psi_i = (\boldsymbol{C}_i, \boldsymbol{\Delta}_i)$, where $\boldsymbol{C}_i = [\boldsymbol{c}_i^1, \ldots, \boldsymbol{c}_i^K]^\top \in \mathbb{R}^{K \times 2d}$ contains K [3] latent reasoning states and $\boldsymbol{\Delta}_i \in [0,1]^{K \times K}$ encodes their relational structures, with $[\boldsymbol{\Delta}_i]_{kl}$ denoting the transition strength from node k to node l. Here, both $\boldsymbol{C}_i$ and $\boldsymbol{\Delta}_i$ are latent and estimated from the trajectories. We further associate Ψ_i with a latent validity landscape $\boldsymbol{v}_i = (v_i^1, \ldots, v_i^K) \in \{0,1\}^K$, where $v_i^k = 1$ indicates that latent node k represents a region of the latent reasoning space containing valid trajectory states. Under $\Psi_i \perp\!\!\!\perp \boldsymbol{\theta} \mid x_i, \{(\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij})\}_{j \in \mathcal{W}_i}$ and conditional independence across worker validity sequences given $\Psi_i, \boldsymbol{v}_i$, and their respective trajectory representations and reliability parameters, the per-task factor in Equation 1 decomposes as:

$$\begin{aligned}&\mathbb{P}\big(\boldsymbol{g}_i, \boldsymbol{o}_i \mid x_i, \{\tau_{ij}\}_{j \in \mathcal{W}_i}, \boldsymbol{\theta}\big) \\ &= \int \mathbb{P}\big(\Psi_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j \in \mathcal{W}_i}\big)\, \mathbb{E}_{\boldsymbol{v}_i \sim \mathbb{P}(\cdot \mid \Psi_i)} \left[\prod_{j \in \mathcal{W}_i} \mathbb{P}(\boldsymbol{g}_{ij} \mid \boldsymbol{v}_i, \Psi_i, \boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}, \boldsymbol{\theta}_j)\, \mathbb{P}(o_{ij} \mid \boldsymbol{g}_{ij}, \boldsymbol{\theta}_j) \right] d\Psi_i\end{aligned} \tag{2}$$

Here, $\mathbb{P}(\boldsymbol{v}_i \mid \Psi_i)$ is the structural prior over the validity landscape.

2.3 Latent Consensus Estimation and Step Alignment

Given the structured trajectory representations $\{(\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij})\}_{j \in \mathcal{W}_i}$, we jointly estimate the latent consensus structure $\Psi_i = (\boldsymbol{C}_i, \boldsymbol{\Delta}_i)$ and a soft correspondence between observed steps $\boldsymbol{r}_{ij}^t$ and latent nodes $\boldsymbol{c}_i^k$, thereby enabling evidence aggregation across worker steps aligned with the same latent node. Because workers may express equivalent reasoning content at different granularities, the alignment must support one-to-many and many-to-one correspondences. Moreover, different solution strategies may cover different subsets of the latent structure. Matching on semantic similarity alone would ignore dependencies among reasoning transitions, whereas searching for an exact structural match is intractable. Therefore, we formulate the alignment as a fused unbalanced Gromov–Wasserstein (FUGW) transport problem (Vayer et al., 2019; Séjourné et al., 2021) that jointly matches step semantics and relational structure. Specifically, we assign uniform source and target masses $\boldsymbol{\mu}_{ij}^{\text{src}} = (1/T_{ij}, \ldots, 1/T_{ij}) \in \mathbb{R}^{T_{ij}}$ and $\boldsymbol{\mu}_i^{\text{tgt}} = (1/K, \ldots, 1/K) \in \mathbb{R}^K$ to $\boldsymbol{R}_{ij}$ and $\boldsymbol{C}_i$, respectively. Given a trade-off parameter $\alpha \in [0,1]$ and a marginal relaxation strength $\rho > 0$, the transport plan $\boldsymbol{\Gamma}_{ij} \in \mathbb{R}_+^{T_{ij} \times K}$ can be obtained by minimizing the FUGW loss below, where $[\boldsymbol{\Gamma}_{ij}]_{tk}$ denotes the mass transported from the t-th reasoning step of worker j to the k-th latent node:

$$\begin{aligned}\mathcal{L}_{\text{FUGW}}(\boldsymbol{\Gamma}; \Psi_i) \;=\; &(1-\alpha) \int 1 - \frac{\langle \boldsymbol{r}_{ij}^t, \boldsymbol{c}_i^k \rangle}{(\|\boldsymbol{r}_{ij}^t\|_2 \, \|\boldsymbol{c}_i^k\|_2)}\, d[\boldsymbol{\Gamma}]_{tk} \;+\; \alpha \int \int |[\boldsymbol{\Delta}_{ij}]_{tt'} - [\boldsymbol{\Delta}_i]_{k\ell}|^2 \, d[\boldsymbol{\Gamma}]_{tk}\, d[\boldsymbol{\Gamma}]_{t'\ell} \\ &+\; \rho \Big[\mathrm{KL}\big(\boldsymbol{\Gamma}_{ij} \mathbf{1}_K \,\big\|\, \boldsymbol{\mu}_{ij}^{\text{src}}\big) + \mathrm{KL}\Big(\boldsymbol{\Gamma}_{ij}^\top \mathbf{1}_{T_{ij}} \,\Big\|\, \boldsymbol{\mu}_i^{\text{tgt}}\Big) \Big]\end{aligned} \tag{3}$$

Here, the first two terms are the semantic transport cost and the structure transport cost, and the third KL term softly controls the transported mass on both sides rather than requiring every step and node to be fully matched. Following Thual et al. (2022), Ψ_i and the trajectory-specific transport plans $\{\boldsymbol{\Gamma}_{ij}\}_{j \in \mathcal{W}_i}$ are jointly estimated by solving $\min_{\Psi_i, \{\boldsymbol{\Gamma}_{ij}\}_{j \in \mathcal{W}_i}} \sum_{j \in \mathcal{W}_i} w_{ij}\, \mathcal{L}_{\text{FUGW}}(\boldsymbol{\Gamma}_{ij}; \Psi_i)$ where

[1] In this study, we instantiate f_ϕ with OpenAI's `text-embedding-3-large` model.

[2] Here, [prompt] is the template instructing the encoder to represent the transition from s_{ij}^{t-1} to s_{ij}^t.

[3] K is a hyperparameter which specifies the size of the latent structure and set as 12 in this study.

$\sum_{j\in\mathcal{W}_i} w_{ij} = 1$. The optimization alternates between updating $\mathbf{\Gamma}$ via block-coordinate descent with Sinkhorn iterations and updating Ψ_i yielding the estimated transport plan $\widehat{\mathbf{\Gamma}}_{ij}$ and the latent structure $\widehat{\Psi}_i$ (see Appendix B.2 for details). From $\widehat{\mathbf{\Gamma}}_{ij}$, the coverage score $m_{ij}^t \in [0,1]$ and the alignment distribution $\boldsymbol{\pi}_{ij}^t = (\pi_{ij}^{t1}, \ldots, \pi_{ij}^{tK})$ for each observed step can be computed as:

$$m_{ij}^t = \min\left\{1,\, T_{ij}\sum_{\ell=1}^{K}[\widehat{\mathbf{\Gamma}}_{ij}]_{t\ell}\right\}, \qquad \pi_{ij}^{tk} = \begin{cases} \dfrac{[\widehat{\mathbf{\Gamma}}_{ij}]_{tk}}{\sum_{\ell=1}^{K}[\widehat{\mathbf{\Gamma}}_{ij}]_{t\ell}}, & \sum_{\ell=1}^{K}[\widehat{\mathbf{\Gamma}}_{ij}]_{t\ell} > 0, \\ 1/K, & \text{otherwise.} \end{cases} \tag{4}$$

Let $\mathcal{A}_{ij} = \{(m_{ij}^t, \boldsymbol{\pi}_{ij}^t)\}_{t=1}^{T_{ij}}$ and $\mathcal{A}_i = \{\mathcal{A}_{ij}\}_{j\in\mathcal{W}_i}$. The worker validity distribution in Equation 2 is specified as $\mathbb{P}\Big(\boldsymbol{g}_{ij} \mid \boldsymbol{v}_i, \widehat{\Psi}_i, \boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}, \boldsymbol{\theta}_j\Big) = \mathbb{P}(\boldsymbol{g}_{ij} \mid \boldsymbol{v}_i, \mathcal{A}_{ij}, \boldsymbol{\theta}_j)$. Approximating $\mathbb{P}(\Psi_i \mid \cdot)$ by $\delta_{\widehat{\Psi}_i}$ (see Appendix B.3 for details) and assuming (i) $g_{ij}^t \perp\!\!\!\perp g_{ij}^{0:t-2} \mid g_{ij}^{t-1}, \boldsymbol{v}_i, m_{ij}^t, \boldsymbol{\pi}_{ij}^t, \boldsymbol{\theta}_j$ with $g_{ij}^0 = 1$, and (ii) $o_{ij} \perp\!\!\!\perp g_{ij}^{1:T_{ij}-1} \mid g_{ij}^{T_{ij}}, \boldsymbol{\theta}_j$, Equation 2 can be factorized as:

$$\mathbb{E}_{\boldsymbol{v}_i\sim\mathbb{P}(\cdot\mid\widehat{\Psi}_i)}\Big[\prod_{j\in\mathcal{W}_i}\Big(\prod_{t=1}^{T_{ij}}\mathbb{P}(g_{ij}^t \mid g_{ij}^{t-1}, \boldsymbol{v}_i, m_{ij}^t, \boldsymbol{\pi}_{ij}^t, \boldsymbol{\theta}_j)\Big)\mathbb{P}(o_{ij} \mid g_{ij}^{T_{ij}}, \boldsymbol{\theta}_j)\Big] \tag{5}$$

Specifically, the structural prior is defined as $\mathbb{P}(\boldsymbol{v}_i \mid \widehat{\Psi}_i) \propto \exp\Big(\lambda_v \sum_{k<l}\Big[\widehat{\boldsymbol{\Delta}}_i + \widehat{\boldsymbol{\Delta}}_i^\top\Big]_{kl}\mathbb{I}[v_i^k = v_i^l]\Big)$ which encourages structurally neighboring latent nodes to share the same validity state. For the transition term, $\boldsymbol{\theta}_j$ contains two worker-specific row-stochastic transition matrices $\mathbf{\Theta}_j^{(u)} \in [0,1]^{2\times 2}$, $u \in \{0,1\}$, shared across all tasks annotated by worker j, where $\sum_{b\in\{0,1\}}[\mathbf{\Theta}_j^{(u)}]_{ab} = 1$ for all $a, u \in \{0,1\}$. For $a, b, u \in \{0,1\}$, $[\mathbf{\Theta}_j^{(u)}]_{ab}$ denotes the probability of transitioning from $g_{ij}^{t-1} = a$ to $g_{ij}^t = b$ under latent-node validity state u. We additionally add the constraint $[\mathbf{\Theta}_j^{(1)}]_{a1} \geq [\mathbf{\Theta}_j^{(0)}]_{a1}$ so that $v_i^k = 1$ is at least as favorable toward a valid reasoning state as $v_i^k = 0$. To model the portion of a step not explained by the latent consensus structure, we introduce a tied row-stochastic transition matrix $\mathbf{\Theta}^{(\varnothing)} \in [0,1]^{2\times 2}$, shared across workers, which decides the validity transition for the unmatched portion independently of the latent validity landscape:

$$\mathbb{P}\big(g_{ij}^t = b \mid g_{ij}^{t-1} = a, \boldsymbol{v}_i, m_{ij}^t, \boldsymbol{\pi}_{ij}^t, \boldsymbol{\theta}_j\big) = (1 - m_{ij}^t)[\mathbf{\Theta}^{(\varnothing)}]_{ab} + m_{ij}^t \sum_{k=1}^{K}\pi_{ij}^{tk}[\mathbf{\Theta}_j^{(v_i^k)}]_{ab} \tag{6}$$

Here, $[\mathbf{\Theta}_j^{(v_i^k)}]_{ab} = (1 - v_i^k)[\mathbf{\Theta}_j^{(0)}]_{ab} + v_i^k[\mathbf{\Theta}_j^{(1)}]_{ab}$. Finally, the terminal validity $g_{ij}^{T_{ij}}$ does not strictly determine the verified outcome o_{ij}, since a sound terminal state may still lead to an incorrect committed answer, whereas an unsound one may occasionally lead to a correct answer. The verified outcome is therefore modeled as a Bernoulli variable with worker-specific correctness parameter $\omega_j^{(a)}$ shared across tasks annotated by worker j:

$$\mathbb{P}\Big(o_{ij} \mid g_{ij}^{T_{ij}} = a, \boldsymbol{\theta}_j\Big) = \big(\omega_j^{(a)}\big)^{o_{ij}}\big(1 - \omega_j^{(a)}\big)^{1-o_{ij}}, \qquad 0 < \omega_j^{(0)} \leq \omega_j^{(1)} < 1 \tag{7}$$

Accordingly, $\boldsymbol{\theta}_j = \{\mathbf{\Theta}_j^{(0)}, \mathbf{\Theta}_j^{(1)}, \mathbf{\Theta}^{(\varnothing)}, \omega_j^{(0)}, \omega_j^{(1)}\}$, where $\mathbf{\Theta}^{(\varnothing)}$ is shared across workers

2.4 Variational Inference of Step Validity

Given the estimated structure $\widehat{\Psi}_i = (\widehat{\boldsymbol{C}}_i, \widehat{\boldsymbol{\Delta}}_i)$ and alignment quantities $\mathcal{A}_i$, we infer the posterior distribution over the latent validity landscape $\boldsymbol{v}_i$ and worker validity sequences $\boldsymbol{g}_i$ under the factorization in Equation 5. By Bayes' rule, the posterior is:

$$\mathbb{P}\Big(\boldsymbol{v}_i, \boldsymbol{g}_i \mid \boldsymbol{o}_i, \widehat{\Psi}_i, \mathcal{A}_i, \boldsymbol{\theta}\Big) \propto \mathbb{P}(\boldsymbol{v}_i \mid \widehat{\Psi}_i)\prod_{j\in\mathcal{W}_i}\left[\prod_{t=1}^{T_{ij}}\mathbb{P}\big(g_{ij}^t \mid g_{ij}^{t-1}, \boldsymbol{v}_i, m_{ij}^t, \boldsymbol{\pi}_{ij}^t, \boldsymbol{\theta}_j\big)\right]\mathbb{P}\Big(o_{ij} \mid g_{ij}^{T_{ij}}, \boldsymbol{\theta}_j\Big) \tag{8}$$

Because $\boldsymbol{v}_i$ couples the validity chains of all workers, exact posterior inference scales exponentially with the size of the latent validity landscape. We therefore approximate the posterior with the structured variational distribution $q_i(\boldsymbol{v}_i, \boldsymbol{g}_i) = q_i^v(\boldsymbol{v}_i)\prod_{j\in\mathcal{W}_i} q_{ij}^g(\boldsymbol{g}_{ij})$, where $q_i^v(\boldsymbol{v}_i) = \prod_{k=1}^{K} q_i^k(v_i^k)$, by maximizing the evidence lower bound (ELBO) $\mathcal{L}_i(q_i; \boldsymbol{\theta})$:

$$\mathcal{L}_i(q_i; \boldsymbol{\theta}) = \mathbb{E}_{q_i}\Big[\log\frac{\mathbb{P}(\boldsymbol{v}_i \mid \widehat{\Psi}_i)}{q_i(\boldsymbol{v}_i, \boldsymbol{g}_i)} + \sum_{j\in\mathcal{W}_i}\Big(\sum_{t=1}^{T_{ij}}\log\mathbb{P}(g_{ij}^t \mid g_{ij}^{t-1}, \boldsymbol{v}_i, m_{ij}^t, \boldsymbol{\pi}_{ij}^t, \boldsymbol{\theta}_j) + \log\mathbb{P}(o_{ij} \mid g_{ij}^{T_{ij}}, \boldsymbol{\theta}_j)\Big)\Big] \tag{9}$$

Let $\widehat{\boldsymbol{A}}_i = \widehat{\boldsymbol{\Delta}}_i + \widehat{\boldsymbol{\Delta}}_i^\top$, $\bar{v}_i^k = q_i^k(v_i^k = 1)$, $\gamma_{ij}^t(a) = q_{ij}^g(g_{ij}^t = a)$, and $\xi_{ij}^t(a,b) = q_{ij}^g(g_{ij}^{t-1} = a, g_{ij}^t = b)$. Substituting the three model components into Equation 9 gives:

$$\begin{aligned}\mathcal{L}_i(q_i;\boldsymbol{\theta}) \overset{c}{=} \frac{\lambda_v}{2}\sum_{k<l}[\widehat{\boldsymbol{A}}_i]_{kl}(2\bar{v}_i^k-1)(2\bar{v}_i^l-1) + \sum_{j\in\mathcal{W}_i}\sum_a \gamma_{ij}^{T_{ij}}(a)\big[o_{ij}\log\omega_j^{(a)} + (1-o_{ij})\log(1-\omega_j^{(a)})\big] \\ + \sum_{j\in\mathcal{W}_i}\sum_{t=1}^{T_{ij}}\sum_{a,b}\xi_{ij}^t(a,b)\,\mathbb{E}_{q_i^v}\Big[\log\Big((1-m_{ij}^t)[\boldsymbol{\Theta}^{(\varnothing)}]_{ab} + m_{ij}^t\textstyle\sum_{k=1}^K \pi_{ij}^{tk}[\boldsymbol{\Theta}_j^{(v_i^k)}]_{ab}\Big)\Big] + \mathcal{H}(q_i)\end{aligned} \tag{10}$$

The expected log-transition term $\mathbb{E}_{q_i^v}[\log(\cdot)]$ in Equation 10 remains intractable because its value depends on the joint combinations of all $\{v_i^k\}_{k=1}^K$. Since $(1-m_{ij}^t)+m_{ij}^t\sum_{k=1}^K\pi_{ij}^{tk}=1$, by Jensen's inequality, for any $a,b\in\{0,1\}$, we have:

$$\mathbb{E}_{q_i^v}[\log(\cdot)] \geq (1-m_{ij}^t)\log[\boldsymbol{\Theta}^{(\varnothing)}]_{ab} + m_{ij}^t\sum_{k=1}^K \pi_{ij}^{tk}\Big[(1-\bar{v}_i^k)\log[\boldsymbol{\Theta}_j^{(0)}]_{ab} + \bar{v}_i^k\log[\boldsymbol{\Theta}_j^{(1)}]_{ab}\Big] \tag{11}$$

Let $\bar{\zeta}_{ij}^t = \sum_{l=1}^K \pi_{ij}^{tl}\bar{v}_i^l \in [0,1]$ denote the alignment-weighted latent validity associated with worker step t. The right-hand side of Equation 11 can then be written as $\widetilde{\ell}_{ij}^t(a,b) = (1-m_{ij}^t)\log[\boldsymbol{\Theta}^{(\varnothing)}]_{ab} + m_{ij}^t\big[(1-\bar{\zeta}_{ij}^t)\log[\boldsymbol{\Theta}_j^{(0)}]_{ab} + \bar{\zeta}_{ij}^t\log[\boldsymbol{\Theta}_j^{(1)}]_{ab}\big]$. Substituting $\widetilde{\ell}_{ij}^t$ into Equation 10 yields a tractable surrogate $\widetilde{\mathcal{L}}_i(q_i;\boldsymbol{\theta}) \leq \mathcal{L}_i(q_i;\boldsymbol{\theta})$ that gives the closed-form updates for the variational factors.

Proposition 2.1 (Update of validity-sequence factor $q_{ij}^g(\boldsymbol{g}_{ij})$)**.** *For fixed latent-landscape factor q_i^v and model parameters $\boldsymbol{\theta}$, maximizing $\widetilde{\mathcal{L}}_i$ with respect to $q_{ij}^g(\boldsymbol{g}_{ij})$ gives:*

$$q_{ij}^g(\boldsymbol{g}_{ij}) \propto \phi_{ij}(g_{ij}^{T_{ij}})\prod_{t=1}^{T_{ij}}\psi_{ij}^t(g_{ij}^{t-1}, g_{ij}^t), \quad \psi_{ij}^t(a,b) = [\boldsymbol{\Theta}^{(\varnothing)}]_{ab}^{1-m_{ij}^t}[\boldsymbol{\Theta}_j^{(0)}]_{ab}^{m_{ij}^t(1-\bar{\zeta}_{ij}^t)}[\boldsymbol{\Theta}_j^{(1)}]_{ab}^{m_{ij}^t\bar{\zeta}_{ij}^t} \tag{12}$$

where $\phi_{ij}(a) = (\omega_j^{(a)})^{o_{ij}}(1-\omega_j^{(a)})^{1-o_{ij}}$. With the fixed initial state $g_{ij}^0 = 1$, its marginals $\gamma_{ij}^t(a)$ and $\xi_{ij}^t(a,b)$ can be computed exactly in $O(T_{ij})$ time using forward–backward recursions.

Proof sketch. For fixed q_i^v and $\boldsymbol{\theta}$, the surrogate $\widetilde{\mathcal{L}}_i$ depends on q_{ij}^g only through ξ_{ij}^t, $\gamma_{ij}^{T_{ij}}$ and $\mathcal{H}(q_{ij}^g)$. Maximizing these terms subject to normalization yields a Gibbs distribution with pairwise factors $\exp(\widetilde{\ell}_{ij}^t)$ and terminal factor ϕ_{ij}. Because each pairwise factor involves only (g_{ij}^{t-1}, g_{ij}^t), the resulting distribution is a first-order Markov chain. See Appendix B.4.1 for details.

Proposition 2.2 (Update of validity-node factor $q_i^k(v_i^k)$)**.** *For fixed worker validity-sequence factors $\{q_{ij}^g(\boldsymbol{g}_{ij})\}_{j\in\mathcal{W}_i}$, remaining validity-node factors $\{q_i^l(v_i^l)\}_{l\neq k}$ and model parameters $\boldsymbol{\theta}$, maximizing $\widetilde{\mathcal{L}}_i$ with respect to the validity-node factor $q_i^k(v_i^k)$ gives:*

$$\bar{v}_i^k = q_i^k(v_i^k=1) = \sigma(\eta_i^k), \quad \eta_i^k = \lambda_v\sum_{l\neq k}[\widehat{\boldsymbol{A}}_i]_{kl}(2\bar{v}_i^l-1) + \sum_{j\in\mathcal{W}_i}\sum_{t=1}^{T_{ij}} m_{ij}^t\pi_{ij}^{tk}\sum_{a,b}\xi_{ij}^t(a,b)\log\frac{[\boldsymbol{\Theta}_j^{(1)}]_{ab}}{[\boldsymbol{\Theta}_j^{(0)}]_{ab}} \tag{13}$$

where $\sigma(z) = (1+e^{-z})^{-1}$, and η_i^k can be computed in $O(K + \sum_{j\in\mathcal{W}_i} T_{ij})$.

Proof sketch. For fixed $\{q_{ij}^g\}_{j\in\mathcal{W}_i}$, $\{q_i^l\}_{l\neq k}$ and $\boldsymbol{\theta}$, the terms in $\widetilde{\mathcal{L}}_i$ that depend on $\bar{v}_i^k$ are $\bar{v}_i^k\eta_i^k + \mathcal{H}(q_i^k)$. The structural prior gives the first component of η_i^k, while the transition term gives the second through the linear relation $\bar{\zeta}_{ij}^t = \sum_{l=1}^K \pi_{ij}^{tl}\bar{v}_i^l$. Setting the derivative with respect to $\bar{v}_i^k$ to zero yields $\bar{v}_i^k = \sigma(\eta_i^k)$. See Appendix B.4.2 for details.

For fixed model parameters $\boldsymbol{\theta}$, we alternately update the validity-sequence factors $\{q_{ij}^g(\boldsymbol{g}_{ij})\}_{j\in\mathcal{W}_i}$ using Proposition 2.1 and the validity-node factors $\{q_i^k(v_i^k)\}_{k=1}^K$ using Proposition 2.2. These updates are repeated for all tasks until $\widetilde{\mathcal{L}}_i$ converges. Given the updated variational distribution, we then optimize the model parameters as $\boldsymbol{\theta}^{\text{new}} \in \arg\max_{\boldsymbol{\theta}}\{\sum_{i=1}^N \widetilde{\mathcal{L}}_i(q_i;\boldsymbol{\theta}) + \log\mathbb{P}(\boldsymbol{\theta})\}$. Please refer to Appendix B.4.3 for update details. We alternate the variational-factor and model-parameter updates until convergence. After convergence, $\gamma_{ij}^t(1)$ gives the inferred posterior validity probability of worker step s_{ij}^t. Accordingly, $\mathcal{D}_{\text{proc}}$ can be augmented as $\mathcal{D}_{\text{aug}} = \{(x_i, \{\tau_{ij}, o_{ij}, \{\gamma_{ij}^t(1)\}_{t=1}^{T_{ij}}\}_{j\in\mathcal{W}_i})\}_{i=1}^N$. Finally, Algorithm 2 summarizes how the proposed method infers step-level validity from crowdsourced worker reasoning trajectories.

Table 1: Comparing step-validity recovery performance of the proposed method and each baseline methods across five reasoning domains in terms of **AUROC (%)** and **AUPRC(%)**. Higher values indicate better performance for both metrics. The best result for each metric is highlighted in bold. All results are averaged over 5 runs.

Method	MATH		GSM8K		LogiQA		P-FOLIO		ZebraLogic	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Outcome-Only	45.2	30.2	49.6	38.2	52.7	49.6	47.1	44.9	54.4	25.2
Self-Check	69.1	53.4	72.6	64.1	65.7	63.0	51.6	47.1	70.7	39.1
Dawid–Skene	71.4	45.2	74.8	64.4	64.9	57.2	60.3	53.0	46.1	20.2
AutoPSV	46.2	35.5	58.8	57.4	53.2	48.0	58.1	58.0	64.0	37.3
FreePRM	65.6	49.6	59.9	52.9	60.7	60.9	52.9	51.6	68.5	40.3
Math-Shepherd	69.3	47.4	77.9	67.8	61.7	57.8	58.9	53.3	54.7	24.6
PAV	53.4	37.0	52.3	47.5	50.3	49.8	51.6	49.1	51.0	23.6
Ours	**82.5**	**76.8**	**86.0**	**80.7**	**81.6**	**81.1**	**82.1**	**80.2**	**76.7**	**47.4**

3 EXPERIMENTS

3.1 EVALUATION SETUP

Datasets. Following recent work (Lu et al., 2024; Havrilla et al., 2024; Agarwal et al., 2024; Setlur et al., 2025), we collect worker trajectories on five task sources spanning mathematical reasoning (MATH (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021)), logical reasoning (LogiQA (Liu et al., 2020), P-FOLIO (Han et al., 2024)), and constraint-satisfaction reasoning (ZebraLogic (Lin et al., 2025)), which differ substantially in reasoning complexity and solution-strategy diversity.

LLMs as Crowdworkers and Trajectory Collection. Motivated by the increasing use of synthetic reasoning data (Havrilla et al., 2024) and LLM-assisted crowdwork in data collection (Veselovsky et al., 2023), we first study whether the process supervision can be recovered when reasoning trajectories are collected from a heterogeneous crowd of LLM workers. We first construct an independent heterogeneous pool of LLM workers for each task source from multiple model families. A worker is defined by a fixed prompt configuration that consists of a persistent worker persona with a domain-specific reasoning strategy, such that the same LLM can instantiate multiple workers under different reasoning prompts. For each task source, we sample questions from the available training split for trajectory collection. Each question is assigned to 10 distinct LLM workers, and each worker needs to complete 20 questions within the corresponding task resource. For each question, the LLM worker receives only the question description and independently produces a complete free-form reasoning trajectory together with a final answer. Workers do not observe other workers' trajectories or outcome labels. To evaluate step-validity recovery, for each task source, we further randomly sample 100 questions as the eval set and annotate all 10 trajectories collected for each question. Each trajectory is independently annotated by two graduate students with backgrounds in AI/ML, who assign a binary validity label to each reasoning step. Disagreements are resolved by a third annotator to determine the final labels. Detailed LLM crowd worker configurations, statistics of collected trajectories, and inter-annotator agreement results are provided in Appendix D.1.

Baselines and Operationalizing the Proposed Method. We first compare against baselines that recover reasoning step-level validity using only the collected trajectories and their outcome signals without access to external knowledge. *Outcome-Only* assigns each reasoning step the correctness of the final outcome. *Self-Check* (Manakul et al., 2023; Zhang et al., 2023) estimates step validity from semantic agreement across workers' reasoning steps. *Dawid–Skene (DS)* (Dawid & Skene, 1979) aggregates outcome signals over semantically matched steps. *AutoPSV* (Lu et al., 2024) learns a prefix-level outcome verifier and derives step-validity labels from changes in verifier confidence, and *FreePRM* (Yuan et al., 2025) trains an outcome reward model from response-level labels and derives implicit process rewards without step-level supervision.We further include external process-verification baselines, *Math-Shepherd* (Wang et al., 2024) and *PAV* (Setlur et al., 2025), which query an additional LLM (Qwen2.5-7B-Instruct) to sample continuations from each reasoning prefix and use the resulting continuation outcomes to estimate step quality. For the proposed method, we initialize $\boldsymbol{C}_i$ using K-means clustering over collected reasoning steps for task x_i and initialize $\boldsymbol{\Delta}_i$ to zero. We set $\alpha = 0.5$ and $\rho = 1$ in Equation 3 across all tasks. We optimize the latent consensus structure using alternating FUGW updates and subsequently perform variational posterior inference over step validity until convergence. To ensure robust evaluation, we repeat all experiments five times and report the average performance. AUROC and AUPRC are used to evaluate each method's ability to recover step-level validity on the eval set where invalid reasoning steps are treated as the positive class for AUPRC.

Table 2: Comparing the downstream reasoning performance of Llama-3.1-8B-Base and Mistral-7B-v0.1 after supervised fine-tuning with the proposed method and different baseline strategies across five reasoning benchmarks, in terms of **Pass@1 Accuracy (%)**. The best result for each backbone on each benchmark is highlighted in **bold**. All results are averaged over 3 runs.

Method	MATH		GSM8K		LogiQA		P-FOLIO		ZebraLogic	
	Llama	Mistral	Llama	Mistral	Llama	Mistral	Llama	Mistral	Llama	Mistral
Base	12.7	6.8	25.6	16.3	24.5	26.8	2.1	1.7	1.7	1.1
Raw-All	14.7	10.7	45.7	27.8	38.7	34.7	36.7	34.8	28.0	18.3
Outcome-Filtering	15.7	11.2	43.9	29.6	36.4	36.8	33.1	33.6	24.8	23.8
Long-Is-More	13.2	10.5	43.2	29.2	34.5	34.1	36.2	34.5	26.8	17.5
STM	14.9	11.1	47.3	31.2	39.9	36.3	38.2	37.2	27.8	25.3
Token Cleaning	16.1	11.8	45.9	29.7	38.4	37.5	31.6	36.7	29.8	26.3
Ours	**16.7**	**12.5**	**47.8**	**31.5**	**42.5**	**38.7**	**40.6**	**38.9**	**32.8**	**29.3**

3.2 Evaluation Results

Table 1 compares our method with the baselines across five task sources in terms of AUROC and AUPRC for step-validity recovery. Overall, our method consistently outperforms all baselines across reasoning tasks. Averaged across the five task sources, our method achieves an AUROC of 81.8% and an AUPRC of 73.2%, which improves over the strongest baseline by 15.8% and 19.8% respectively. Below, we summarize several key observations.

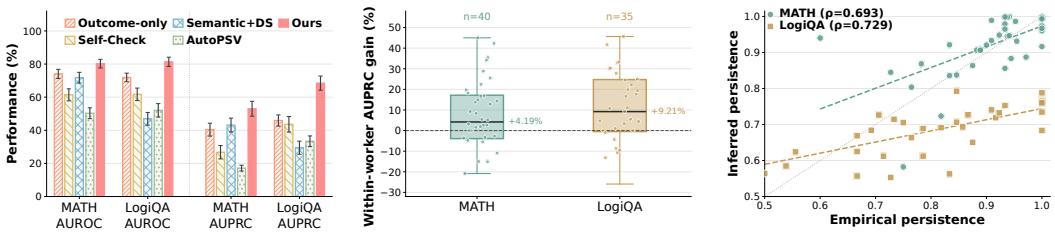
Structured cross-worker inference improves step-validity recovery. Compared with existing baselines, our proposed method generally achieves better recovery performance across different reasoning tasks. For example, on LogiQA, our method achieves 81.6% AUROC and 81.1% AUPRC, compared with 65.7% and 63.0% for the best baseline. Similar improvements are also observed on other reasoning tasks.

Cross-worker information provides useful process signals beyond final outcomes. Among methods that rely only on the collected trajectories and their outcome signals, *Self-Check* and *Dawid–Skene*, which exploit information across multiple worker trajectories, generally outperform methods that primarily derive process supervision from final outcomes. For example, on MATH, *Dawid–Skene* achieves 71.4% AUROC, compared with 45.2% for *Outcome-Only*, 46.2% for *AutoPSV*, and 65.6% for *FreePRM*. Similarly, on GSM8K, *Self-Check* and *Dawid–Skene* achieve AUPRC scores of 64.1% and 64.4%, respectively, compared with 38.2% for *Outcome-Only* and 52.9% for *FreePRM*. These results suggest that information shared across independently collected reasoning trajectories provides additional signals for recovering intermediate reasoning validity.

Cross-worker inference provides strong process supervision without external reasoning evidence. External process-verification baselines provide useful but not strictly information-matched references. *Math-Shepherd* and *PAV* query an additional LLM to sample continuations from each reasoning prefix and infer step validity from the resulting continuation outcomes. Their performance can depend on the rollout model and budget, but these queries also introduce reasoning evidence and generation costs beyond the collected trajectories. Under the fixed evaluated setting with Qwen2.5-7B-Instruct and eight rollouts per prefix, *Math-Shepherd* is competitive on some tasks, while our method achieves stronger overall recovery performance using only the originally collected crowd trajectories and their verified final-answer outcomes. These results suggest that cross-worker trajectory structure provides an effective source of step-level process supervision without requiring per-prefix rollout generation or an external step-level judge. Appendix D.3–D.7 provides additional analyses on sensitivity, worker scaling, and ablations of evidence sources and structured inference.

3.3 Does Inferred Step Validity Improve Reasoning Supervised Finetuning?

We next evaluate whether the recovered step-level validity signals provide useful supervision for downstream reasoning model training. We fine-tune Llama-3.1-8B-Base (Grattafiori et al., 2024) and Mistral-7B-v0.1 (Jiang et al., 2023) on the collected reasoning trajectories using different data-selection and supervision strategies, and evaluate the resulting models on the benchmark evaluation sets using *Pass@1 accuracy*. Full training and evaluation details are provided in Appendix D.8. In addition to training on all collected trajectories (*Raw-All*) and filtering trajectories based on final-answer correctness (*Outcome-Only*), we compare against recent data-selection and token-filtering approaches, including *Long-Is-More* (Zhao et al., 2024), *STM* (Wu et al., 2025), and *Token Cleaning* (Pang et al., 2025). For our method, the inferred posterior step-validity probabilities are used as soft weights on the corresponding reasoning-step tokens during supervised fine-tuning. Table 2



(a) Step-validity recovery (b) Within-worker recovery (c) Persistence fidelity

Figure 1: Comparing step-validity recovery performance across different methods in terms of AUROC and AUPRC, within-worker AUPRC gains of our method over the strongest baseline for each individual worker, and the correspondence between our method's inferred reasoning persistence and empirical reasoning persistence across individual workers on real human-crowd reasoning trajectories collected for the MATH and LogiQA tasks in (a), (b), and (c), respectively. Error bars in (a) represent the standard error of the mean. Each point in (b) and (c) represents an individual worker, and ρ in (c) denotes Spearman's correlation coefficient.

compares the downstream reasoning performance of different backbone models after supervised fine-tuning with our proposed method and different baseline strategies. Overall, our method achieves the highest Pass@1 accuracy across all five benchmarks with both backbone models. Soft supervision with inferred step-validity probabilities consistently improves downstream performance over training on all trajectories or filtering by final-answer correctness. For example, with Llama-3.1-8B-Base, our method improves GSM8K accuracy from 45.7% with *Raw-All* and 43.9% with *Outcome-Only* to 47.8%, while on P-FOLIO the corresponding accuracy improves from 36.7% and 33.1% to 40.6%. Our method also outperforms recent data-selection and token-filtering approaches. Similar trends are observed with Mistral-7B-v0.1. These results show that the inferred step-validity signals provide useful supervision for reasoning SFT on heterogeneous crowdsourced reasoning trajectories beyond trajectory-level filtering and existing token-level data-selection strategies.

3.4 Real-World Study with Human Crowdworkers

Finally, we move on to examine how the proposed method performs on reasoning trajectories collected from real human crowdworkers. We recruited participants through Prolific [4] to solve questions from MATH and LogiQA. Each participant was assigned to one task source and was asked to complete eight problem-solving tasks sampled from that source. For each task, the participant was required to provide a step-by-step reasoning trajectory and a final answer. Overall, we collected 309 trajectories from 40 participants over 30 MATH questions and 265 trajectories from 35 participants over 30 LogiQA questions. [5] To evaluate step-validity recovery, each collected human reasoning trajectory is independently annotated by two graduate students with backgrounds in AI/ML, who assign a binary validity label to each reasoning step. Disagreements are resolved by a third annotator to determine the final labels.

We first examine the overall step-validity recovery performance on the human-collected reasoning trajectories. Specifically, all methods are refit on the human-collected data and evaluated against the expert-annotated step-validity labels. Figure 1(a) compares the proposed method with *Outcome-only*, *Self-Check*, *DS*, and *AutoPSV* over MATH and LogiQA in terms of AUROC and AUPRC, where invalid reasoning steps are treated as the positive class for AUPRC. Overall, our method achieves the best step-validity recovery performance on both tasks. In terms of AUPRC, our method achieves 53.00% on MATH and 68.48% on LogiQA, which outperforms the strongest baselines by 9.83% and 22.57%, respectively. These results show that the advantage of our structured cross-worker inference extends beyond LLM-generated trajectories to reasoning trajectories collected from real human crowdworkers. We next examine whether the overall improvement is driven primarily by differences in average quality across workers, or whether our method can also distinguish valid and invalid reasoning steps within the same worker. Specifically, for each worker, we aggregate all expert-annotated reasoning steps across the questions assigned to that worker and compute the AUPRC of each method on this worker's steps. We then compare our method with the strongest

[4] https://www.prolific.com/

[5] Although each participant was assigned eight tasks, some participants did not complete all assigned tasks.

baseline selected separately for each worker, and report the corresponding AUPRC difference in Figure 1(b). Overall, our method achieves positive median within-worker AUPRC gains of 4.19% on MATH and 9.21% on LogiQA. These results indicate that the improvement of our method is not solely explained by differences in overall reasoning quality across workers, but also reflects its ability to recover step-level validity variations within individual workers. Finally, we explore whether the inferred validity states capture meaningful differences in workers' reasoning dynamics. We focus on reasoning persistence, which characterizes a worker's tendency to remain in a valid reasoning state after a valid step. For each worker, empirical persistence is computed from the expert annotations as the proportion of adjacent step transitions that remain valid among those whose preceding step is valid, and the inferred persistence is computed from our method's posterior joint probabilities over adjacent validity states. Both quantities are computed only over adjacent steps within the same trajectory across different human workers. As shown in Figure 1(c), inferred persistence is strongly positively correlated with empirical persistence on both tasks, with Spearman's correlation coefficients of $\rho = 0.693$ on MATH and $\rho = 0.729$ on LogiQA. This correspondence suggests that the inferred validity states preserve meaningful worker-level reasoning dynamics in the annotation process, rather than merely providing accurate predictions for individual reasoning steps.

4 Related Work

Learning from Noisy Crowdsourced Annotations. Learning from crowds aims to recover latent ground-truth labels from multiple noisy annotations while accounting for heterogeneous worker reliability Zhang et al. (2016); Daniel et al. (2018); Zheng et al. (2017). Classical approaches model each worker using a confusion matrix or related competence parameters and jointly estimate worker quality and task labels through probabilistic inference (Raykar et al., 2010; Hovy et al., 2013; Guo et al., 2023). Subsequent work has considered more flexible forms of worker expertise, task difficulty, worker dependence, and sparse annotation patterns (Karger et al., 2011; Zheng et al., 2017; Nguyen et al., 2024; Zhang et al., 2025). Learning from crowds has also been extended to structured annotations, including sequence tags and coreference chains (Simpson & Gurevych, 2019; Xia et al., 2024b). Prior crowdsourcing research has largely focused on aggregating noisy worker annotations over a shared annotation space for the task. We study a new crowdsourcing setting where workers provide free-form reasoning trajectories and final answers, with intermediate steps varying substantially across workers in content, length, and solution path. We therefore formulate process-supervision recovery from heterogeneous worker trajectories as structured latent-state inference over step-level validity.

Process Supervision for LLM Fine-Tuning Process supervision and data selection provide ways to improve the training signal available to LLMs. Prior work obtains process-level feedback from annotations on reasoning steps (Lightman et al., 2024), continuation-based estimates of eventual success (Wang et al., 2024), progress under a prover policy (Setlur et al., 2025), and outcome-supervised models (Lu et al., 2024; Yuan et al., 2025).Another line of prior work on instruction tuning studies how to allocate training emphasis across data. LIMA and Long-Is-More investigate curated datasets and response-length-based selection (Zhou et al., 2023; Zhao et al., 2024), LESS selects examples using gradient-based influence estimates (Xia et al., 2024a), and ScaleBiO optimizes data-source sampling weights against a validation objective through bilevel optimization (Pan et al., 2025). Token-level methods further filter supervision using perplexity or the influence of model updates (Wu et al., 2025; Pang et al., 2025). Prior work mainly studies how to obtain step-level feedback for a given reasoning trajectory, or how to optimize the use of supervision from a given training corpus during fine-tuning. We instead study how to recover reliable process supervision from heterogeneous workers' free-form reasoning trajectories and final answers, using a crowdsourcing formulation that jointly infers step-level validity and worker reliability.

5 Discussions and Conclusion

This work shows that reliable process supervision can be recovered directly from noisy and heterogeneous crowd reasoning trajectories. Our structured inference framework uses a task-specific latent consensus process to establish cross-trajectory correspondences through soft alignment, enabling structured variational inference of step validity from aligned trajectories and observed final outcomes. Experiments with LLM and human crowdworkers demonstrate improved step-validity recovery, while LLM fine-tuning confirms the value of the recovered signals across backbone models and reasoning benchmarks.

AI USE STATEMENT

For manuscript preparation, we used AI tools for grammar checking and language editing. The authors reviewed the text and take full responsibility for its content. The use of LLMs as experimental crowdworkers is described in Section 3 and Appendix D.1.

ETHICS STATEMENT

Our human-participant study involved minimal risk. Data collection was approved by our institutional review board (IRB). Participants recruited through Prolific completed mathematical and logical reasoning tasks from MATH and LogiQA, as described in Section 3.4.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the proposed method in Section 2 and provide detailed derivations and the inference algorithm in Appendix B and Algorithm 2. The datasets, trajectory collection procedures, baselines, and evaluation protocols are described in Section 3. LLM worker models, prompts, collection statistics, and hyperparameter settings are provided in Appendix D.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei Zhang, et al. Many-shot in-context learning. In *Advances in Neural Information Processing Systems*, 2024.

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

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40, 2018.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. *Advances in Neural Information Processing Systems*, 36:347–386, 2023.

Simeng Han, Aaron Yu, Rui Shen, Zhenting Qi, Martin Riddell, Wenfei Zhou, Yujie Qiao, Yilun Zhao, Semih Yavuz, Ye Liu, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Dragomir Radev, Rex Ying, and Arman Cohan. P-FOLIO: Evaluating and improving logical reasoning with abundant human-written reasoning chains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16553–16565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.966. URL https://aclanthology.org/2024.findings-emnlp.966/.

Alexander Havrilla, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. GLoRe: When, where, and how to improve LLM reasoning via global and local refinements. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 1120–1130, 2013.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4615–4634. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.254. URL https://aclanthology.org/2024.acl-long.254/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. *Advances in neural information processing systems*, 24, 2011.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37889–37905. PMLR, 2025. URL https://proceedings.mlr.press/v267/lin25i.html.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/501. URL https://www.ijcai.org/proceedings/2020/501.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, Yunlong Feng, and Zhijiang Guo. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37:79935–79962, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. *Advances in Neural Information Processing Systems*, 37: 97261–97298, 2024.

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. ScaleBiO: Scalable bilevel optimization for LLM data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31959–31982. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1543. URL https://aclanthology.org/2025.acl-long.1543/.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. Token cleaning: Fine-grained data selection for LLM supervised fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47837–47858. PMLR, 2025. URL https://proceedings.mlr.press/v267/pang25a.html.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11: 1297–1322, 2010.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, 2025.

Edwin Simpson and Iryna Gurevych. A bayesian approach for sequence tagging with crowds. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 1093–1104, 2019.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.06585.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263. Association for Computational Linguistics, 2008.

Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced gromov wasserstein. *Advances in neural information processing systems*, 35:21792–21804, 2022.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Titouan Vayer, Nicolas Courty, Romain Tavenard, Laetitia Chapel, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 2019.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023. doi: 10.48550/arXiv.2306.07899. URL https://arxiv.org/abs/2306.07899.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 70938–70970, 2025.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4):791–813, 2023. doi: 10.1007/s00778-022-00775-9.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, Shao-Hua Sun, and Hung-yi Lee. Mitigating forgetting in llm fine-tuning via low-perplexity token learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0064.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024a. URL https://proceedings.mlr.press/v235/xia24c.html.

Mingxuan Xia, Zenan Huang, Runze Wu, Gengyu Lyu, Junbo Zhao, Gang Chen, and Haobo Wang. Unbiased multi-label learning from crowdsourced annotations. In *Forty-first International Conference on Machine Learning*, 2024b.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27723–27730, 2025. doi: 10.1609/aaai.v39i26.34987. URL https://ojs.aaai.org/index.php/AAAI/article/view/34987.

Zhaohui Yang, Chenghua He, Xiaowen Shi, Shihong Deng, Linjing Li, Qiyue Yin, and Daxin Jiang. Beyond the first error: Process reward models for reflective mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4711–4728. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.253. URL https://aclanthology.org/2025.findings-emnlp.253/.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. Sac3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15445–15458, 2023.

Jing Zhang, Xindong Wu, and Victor S Sheng. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review*, 46(4):543–576, 2016.

Wenjun Zhang, Liangxiao Jiang, and Chaoqun Li. Tllc: transfer learning-based label completion for crowdsourcing. In *Forty-second International Conference on Machine Learning*, 2025.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60674–60703. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24b.html.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endow.*, 10(5):541–552, 2017.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

A NOTATION SUMMARY

Table 3 summarizes the main notation used throughout the proposed framework.

Table 3: Summary of the main notation.

Notation	Description
Tasks and worker trajectories	
N, M	Number of tasks and number of workers.
x_i	Task i.
$\mathcal{W}_i$	Set of workers contributing to task x_i.
s_{ij}^t	The t-th reasoning step provided by worker j for task i.
T_{ij}	Number of reasoning steps in trajectory τ_{ij}.
y_{ij}	Final answer committed by worker j.
τ_{ij}	Reasoning trajectory $(s_{ij}^1, \ldots, s_{ij}^{T_{ij}})$.
o_{ij}	Verified final-answer outcome, $o_{ij} = V_i(y_{ij}) \in \{0, 1\}$.
g_{ij}^t	Latent validity of reasoning step t, with $g_{ij}^t \in \{0, 1\}$ and $g_{ij}^0 = 1$.
$\boldsymbol{g}_{ij}$	Latent validity sequence $(g_{ij}^1, \ldots, g_{ij}^{T_{ij}})$.
$\boldsymbol{g}_i$, $\boldsymbol{o}_i$	Task-level collections $\{\boldsymbol{g}_{ij}\}_{j \in \mathcal{W}_i}$ and $\{o_{ij}\}_{j \in \mathcal{W}_i}$.
$\mathcal{D}_{\text{proc}}$	Collected process-supervised dataset.
$\mathcal{D}_{\text{aug}}$	Process-annotated dataset produced by the framework.
Trajectory representation and latent consensus	
f_ϕ	Frozen pretrained encoder used to embed reasoning steps.
$\boldsymbol{r}_{ij}^t$	Representation of reasoning step s_{ij}^t.
$\boldsymbol{R}_{ij}$	Matrix containing the representations of all steps in τ_{ij}.
$\boldsymbol{\Delta}_{ij}$	Directed transition structure of worker trajectory τ_{ij}.
K	Number of latent consensus nodes.
$\Psi_i = (\boldsymbol{C}_i, \boldsymbol{\Delta}_i)$	Task-specific latent consensus process structure.
$\boldsymbol{c}_i^k$, $\boldsymbol{C}_i$	Representation of latent consensus node k, and their stacked matrix.
$\boldsymbol{\Delta}_i$	Transition strength between latent consensus nodes.
$\widehat{\boldsymbol{A}}_i$	Symmetrized structure $\widehat{\boldsymbol{\Delta}}_i + \widehat{\boldsymbol{\Delta}}_i^\top$.
$\boldsymbol{v}_i$, v_i^k	Latent validity landscape $(v_i^1, \ldots, v_i^K)$ and the state of node k.
λ_v	Strength of the structural prior over $\boldsymbol{v}_i$.
Consensus alignment	
$\boldsymbol{\Gamma}_{ij}$	FUGW transport plan between τ_{ij} and Ψ_i.
α, ρ	Semantic–relational trade-off and marginal relaxation strength.
m_{ij}^t	Coverage score measuring how strongly step t is covered by $\widehat{\Psi}_i$.
$\boldsymbol{\pi}_{ij}^t$	Alignment distribution of step t over the K latent consensus nodes.
$\mathcal{A}_i$	Alignment quantities $\{(m_{ij}^t, \boldsymbol{\pi}_{ij}^t)\}_{j \in \mathcal{W}_i,\, t \le T_{ij}}$.
Validity-transition and outcome model	
$\boldsymbol{\Theta}_j^{(u)}$	Worker-specific validity-transition matrix conditioned on latent-node validity u.
$\boldsymbol{\Theta}^{(\varnothing)}$	Globally tied validity-transition matrix for the unmatched portion of a step.
$\omega_j^{(a)}$	Probability of a correct final answer conditioned on terminal validity state a.
$\boldsymbol{\theta}_j$	Worker-specific parameters $\{\boldsymbol{\Theta}_j^{(0)}, \boldsymbol{\Theta}_j^{(1)}, \omega_j^{(0)}, \omega_j^{(1)}, \boldsymbol{\Theta}^{(\varnothing)}\}$.
$\boldsymbol{\theta}$	All reliability parameters, $\{\boldsymbol{\theta}_j\}_{j=1}^M$.
Variational inference	
$q_i^v(\boldsymbol{v}_i)$	Variational distribution over the latent validity landscape.
$q_{ij}^g(\boldsymbol{g}_{ij})$	Variational distribution over worker j's validity sequence.
$\bar{v}_i^k$	Posterior validity probability $q_i^k(v_i^k = 1)$.
$\gamma_{ij}^t(a)$	Step-validity marginal $q_{ij}^g(g_{ij}^t = a)$.
$\xi_{ij}^t(a, b)$	Pairwise validity marginal $q_{ij}^g(g_{ij}^{t-1} = a, g_{ij}^t = b)$.
$\bar{\zeta}_{ij}^t$	Alignment-weighted latent validity $\sum_{k=1}^K \pi_{ij}^{tk} \bar{v}_i^k$.
$\psi_{ij}^t(a, b)$	Effective validity-transition potential of step t.
η_i^k	The update of q_i^k.
$\mathcal{L}_i$	Evidence lower bound (ELBO) for task i.
$\widetilde{\mathcal{L}}_i$	Tractable surrogate lower bound on $\mathcal{L}_i$.

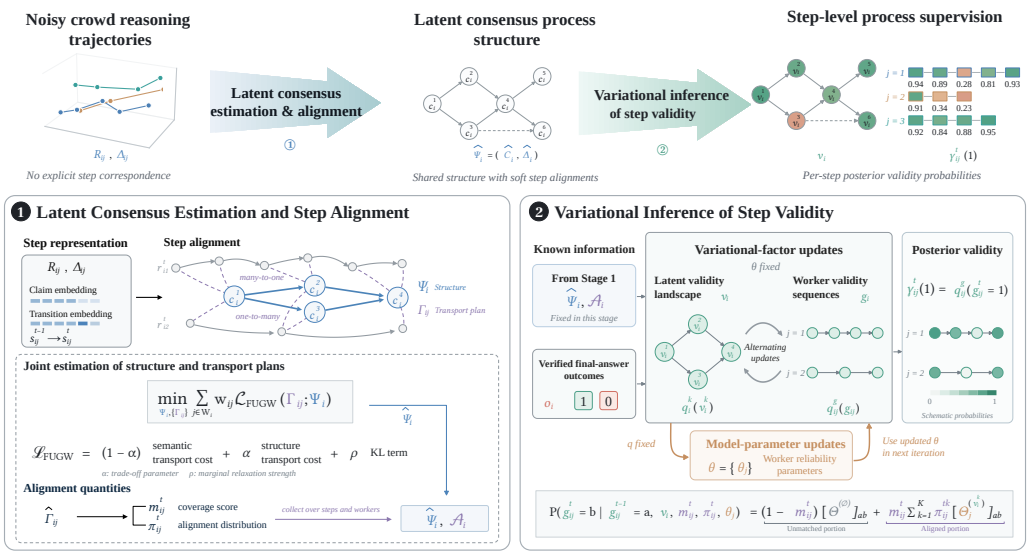


Figure 2: Overview of learning step-level process supervision from heterogeneous crowd reasoning trajectories without step-level expert annotation. (1) Claim and transition embeddings form step representations. FUGW jointly estimates the latent consensus process structure Ψ_i and transport plans Γ_{ij}, yielding coverage scores and alignment distributions collected in $\mathcal{A}_i$. (2) With $\widehat{\Psi}_i$ and $\mathcal{A}_i$ fixed, variational inference uses verified final-answer outcomes o_i to infer the latent validity landscape v_i and worker validity sequences g_i, while learning worker reliability parameters θ. Alternating variational-factor and model-parameter updates yield posterior step-validity probabilities $\gamma_{ij}^t(1) = q_{ij}^g(g_{ij}^t = 1)$ for the original trajectories. Node colors and numerical probabilities are illustrative.

B Additional Details in Section 2

B.1 The illustration of the proposed Framework

B.2 Optimization of the Latent Consensus Structure

For each task x_i, we jointly estimate the latent consensus structure $\Psi_i = (\boldsymbol{C}_i, \boldsymbol{\Delta}_i)$ and the trajectory-specific transport plans by solving

$$\min_{\Psi_i, \{\boldsymbol{\Gamma}_{ij}\}_{j \in \mathcal{W}_i}} \sum_{j \in \mathcal{W}_i} w_{ij}\, \mathcal{L}_{\mathrm{FUGW}}(\boldsymbol{\Gamma}_{ij}; \Psi_i). \tag{14}$$

In our implementation, all trajectories for the same task receive equal weight, $w_{ij} = 1/|\mathcal{W}_i|$. These weights are fixed during consensus estimation and do not depend on the worker reliability parameters $\boldsymbol{\theta}_j$, consistent with the conditional-independence assumption in Section 2.2. Because Equation 14 is jointly nonconvex, we optimize it by alternating among the transport plans, the latent node representations, and the latent transition strengths.

Transport-plan update. For fixed Ψ_i, the objective separates across trajectories, so the plans $\boldsymbol{\Gamma}_{ij}$ can be updated independently. The semantic term in Equation 3 is linear in the transport plan, whereas the relational term is quadratic. At inner iteration q, we linearize the relational term at the current plan and form the local cost

$$\boldsymbol{S}_{ij}^{(q)} = (1-\alpha)\boldsymbol{M}_{ij} + \alpha \nabla_{\boldsymbol{\Gamma}} \mathcal{L}_{\mathrm{rel}}\big(\boldsymbol{\Gamma}_{ij}^{(q)}; \boldsymbol{\Delta}_i\big), \tag{15}$$

where $[\boldsymbol{M}_{ij}]_{tk} = 1 - \langle \boldsymbol{r}_{ij}^t, \boldsymbol{c}_i^k \rangle / (\|\boldsymbol{r}_{ij}^t\|_2 \|\boldsymbol{c}_i^k\|_2)$. This gives an unbalanced optimal-transport subproblem with the two marginal KL penalties in Equation 3. We solve this subproblem using generalized Sinkhorn iterations with a KL-proximal term centered at the current plan. This proximal term is used only inside the numerical update and is not added to the alignment objective. The inner iterations continue until the relative change of the plan is below a prescribed tolerance or a maximum number

Algorithm 1 Latent consensus estimation for task x_i

Require: $\{(\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij})\}_{j\in\mathcal{W}_i}$, latent size K, and parameters α, ρ
Ensure: $\widehat{\Psi}_i = (\widehat{\boldsymbol{C}}_i, \widehat{\boldsymbol{\Delta}}_i)$ and $\{\widehat{\boldsymbol{\Gamma}}_{ij}\}_{j\in\mathcal{W}_i}$
1: Set $w_{ij} = 1/|\mathcal{W}_i|$ and initialize $\boldsymbol{C}_i$
2: Initialize $\{\boldsymbol{\Gamma}_{ij}\}$ and $\boldsymbol{\Delta}_i$
3: **repeat**
4: **for** $j \in \mathcal{W}_i$ **do**
5: Update $\boldsymbol{\Gamma}_{ij}$ by linearized FUGW and generalized Sinkhorn
6: **end for**
7: Update $\boldsymbol{C}_i$ using Equation 16
8: Update $\boldsymbol{\Delta}_i$ using Equation 17
9: **until** the weighted objective converges
10: **return** $\widehat{\Psi}_i$ and $\{\widehat{\boldsymbol{\Gamma}}_{ij}\}_{j\in\mathcal{W}_i}$

of iterations is reached.

Latent-node update. For fixed transport plans and $\boldsymbol{\Delta}_i$, each latent node is updated by the transport-weighted spherical mean of the step representations. Writing $\bar{\boldsymbol{r}}_{ij}^t = \boldsymbol{r}_{ij}^t / \|\boldsymbol{r}_{ij}^t\|_2$, we compute

$$\widetilde{\boldsymbol{c}}_i^k = \sum_{j\in\mathcal{W}_i} w_{ij} \sum_{t=1}^{T_{ij}} [\boldsymbol{\Gamma}_{ij}]_{tk} \bar{\boldsymbol{r}}_{ij}^t, \qquad \boldsymbol{c}_i^k \leftarrow \frac{\widetilde{\boldsymbol{c}}_i^k}{\|\widetilde{\boldsymbol{c}}_i^k\|_2}. \tag{16}$$

If $\widetilde{\boldsymbol{c}}_i^k = \mathbf{0}$, we retain the previous node representation. When $\alpha = 1$, the semantic term is absent and $\boldsymbol{C}_i$ is left unchanged.

Latent-transition update. For fixed transport plans and $\boldsymbol{C}_i$, the transition strength between latent nodes k and ℓ is updated by transporting the observed transition matrices into the latent space and normalizing by the transported mass:

$$[\boldsymbol{\Delta}_i]_{k\ell} \leftarrow \frac{\sum_{j\in\mathcal{W}_i} w_{ij} \sum_{t,t'} [\boldsymbol{\Gamma}_{ij}]_{tk} [\boldsymbol{\Gamma}_{ij}]_{t'\ell} [\boldsymbol{\Delta}_{ij}]_{tt'}}{\sum_{j\in\mathcal{W}_i} w_{ij} \sum_{t,t'} [\boldsymbol{\Gamma}_{ij}]_{tk} [\boldsymbol{\Gamma}_{ij}]_{t'\ell}}. \tag{17}$$

The previous value is retained if the denominator is zero. Since $[\boldsymbol{\Delta}_{ij}]_{tt'} \in \{0, 1\}$, Equation 17 is a transported-mass-normalized average and therefore remains in $[0, 1]$. It measures the average immediate transition evidence between steps softly aligned to latent nodes k and ℓ; no row normalization is imposed because $\boldsymbol{\Delta}_i$ represents transition strength rather than a Markov transition matrix. When $\alpha = 0$, the relational term is absent and $\boldsymbol{\Delta}_i$ is left unchanged. ***Initialization and stopping.*** We initialize $\boldsymbol{C}_i$ by applying spherical K-means to the pooled normalized step representations. We then obtain initial transport plans from semantic matching and initialize $\boldsymbol{\Delta}_i$ to zero. The three blocks are alternated until the relative change of the weighted objective in Equation 14 is below a tolerance or a maximum number of outer iterations is reached. Because the complete problem is nonconvex, the procedure returns an initialization-dependent local solution and does not guarantee the global optimum.

su ***Complexity.*** For a dense trajectory relation matrix, evaluating the linearized relational cost requires $\mathcal{O}(T_{ij}^2 K + T_{ij} K^2)$ time per inner iteration. Since $\boldsymbol{\Delta}_{ij}$ has only $T_{ij} - 1$ nonzero consecutive-transition entries in our setting, multiplications involving $\boldsymbol{\Delta}_{ij}$ are sparse, reducing the cost to $\mathcal{O}(T_{ij} K^2)$. The values of K, α, ρ, and all optimization tolerances are reported together with the experimental configuration.

B.3 Point-Mass Approximation of the Latent Process Structure

We provide additional details for the point-mass approximation $\mathbb{P}(\Psi_i \mid \cdot) \approx \delta_{\widehat{\Psi}_i}$ used in Section 2.4. Recall that the latent process structure Ψ_i and the trajectory-to-structure couplings $\{\boldsymbol{\Gamma}_{ij}\}_{j\in\mathcal{W}_i}$ are estimated jointly through the alignment objective in Section 2.3. Let $\boldsymbol{\Gamma}_i = \{\boldsymbol{\Gamma}_{ij}\}_{j\in\mathcal{W}_i}$ and denote the joint alignment objective as:

$$\mathcal{J}_i(\Psi_i, \boldsymbol{\Gamma}_i) = \sum_{j\in\mathcal{W}_i} w_{ij} \mathcal{L}_{\mathrm{FUGW}}(\boldsymbol{\Gamma}_{ij}; \Psi_i) \tag{18}$$

The corresponding solution can be obtained by minimizing the above objective:

$$(\Psi_i^\star, \mathbf{\Gamma}_i^\star) \in \arg\min_{\Psi_i, \mathbf{\Gamma}_i} \mathcal{J}_i(\Psi_i, \mathbf{\Gamma}_i). \tag{19}$$

To instantiate the conditional distribution over Ψ_i appearing in Equation 2, we construct a joint Gibbs distribution over the process structure and alignment couplings from the alignment objective:

$$\mathbb{P}(\Psi_i, \mathbf{\Gamma}_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j\in\mathcal{W}_i}) = \frac{1}{Z_i} \exp[-\kappa \mathcal{J}_i(\Psi_i, \mathbf{\Gamma}_i)] \tag{20}$$

Here, $\kappa > 0$ controls the concentration of the distribution and Z_i denotes the corresponding normalizing constant. This construction depends only on the task and its trajectory representations and does not involve the worker reliability parameters $\boldsymbol{\theta}$ or the verified outcomes $\boldsymbol{o}_i$, consistent with the structural conditional distribution in Equation 2. Since the exponential in Equation 20 is strictly decreasing in $\mathcal{J}_i$, maximizing the joint Gibbs distribution is equivalent to minimizing the alignment objective for any $\kappa > 0$:

$$(\Psi_i^\star, \mathbf{\Gamma}_i^\star) \in \arg\max_{\Psi_i, \mathbf{\Gamma}_i} \mathbb{P}(\Psi_i, \mathbf{\Gamma}_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j\in\mathcal{W}_i}) \tag{21}$$

Rather than integrating over the uncertainty in $(\Psi_i, \mathbf{\Gamma}_i)$, we approximate their joint conditional distribution by a point mass at the estimated local solution:

$$\mathbb{P}(\Psi_i, \mathbf{\Gamma}_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j\in\mathcal{W}_i}) \approx \delta_{(\widehat{\Psi}_i, \widehat{\mathbf{\Gamma}}_i)}(\Psi_i, \mathbf{\Gamma}_i) \tag{22}$$

Marginalizing Equation 22 over $\mathbf{\Gamma}_i$ gives the point-mass approximation of the structural distribution appearing in Equation 2:

$$\mathbb{P}(\Psi_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j\in\mathcal{W}_i}) \approx \delta_{\widehat{\Psi}_i}(\Psi_i) \tag{23}$$

Therefore, for any integrable function $F_i(\Psi_i)$:

$$\int F_i(\Psi_i)\mathbb{P}(\Psi_i \mid x_i, \{\boldsymbol{R}_{ij}, \boldsymbol{\Delta}_{ij}\}_{j\in\mathcal{W}_i})\, d\Psi_i \approx \int F_i(\Psi_i)\delta_{\widehat{\Psi}_i}(\Psi_i)\, d\Psi_i = F_i(\widehat{\Psi}_i) \tag{24}$$

Applying Equation 24 to Equation 2 yields the plug-in factorization used in Section 2.4. The estimated couplings $\widehat{\mathbf{\Gamma}}_i$ are subsequently used to construct the alignment summaries $\mathcal{A}_i$ used in the validity model.

B.4 Variational Inference and Parameter Estimation

We follow the notation of Section 2.4 and hold the estimated consensus structures $\widehat{\Psi}_i$ and alignment quantities $\mathcal{A}_i$ fixed throughout.

B.4.1 Derivation of the Validity-Sequence Update

Fix a task i, a worker $j \in \mathcal{W}_i$, the latent-landscape factor q_i^v, and the model parameters $\boldsymbol{\theta}$. By Equation 7, the terminal observation likelihood conditioned on $g_{ij}^{T_{ij}} = a$ is:

$$\phi_{ij}(a) = \mathbb{P}\big(o_{ij} \mid g_{ij}^{T_{ij}}{=}a, \boldsymbol{\theta}_j\big) = \big(\omega_j^{(a)}\big)^{o_{ij}} \big(1-\omega_j^{(a)}\big)^{1-o_{ij}} \tag{25}$$

In Equation 10, the structural term does not involve q_{ij}^g, whereas the transition and terminal terms associated with worker j do, and the entropy contributes $\mathcal{H}(q_{ij}^g)$. After replacing each expected log-transition by the bound in Equation 11, the part of the surrogate that depends on q_{ij}^g is therefore:

$$\widetilde{\mathcal{L}}_{ij}^g\big(q_{ij}^g\big) \stackrel{c}{=} \sum_{t=1}^{T_{ij}} \sum_{a,b} \xi_{ij}^t(a,b)\, \widetilde{\ell}_{ij}^t(a,b) + \sum_a \gamma_{ij}^{T_{ij}}(a)\, \log \phi_{ij}(a) + \mathcal{H}\big(q_{ij}^g\big) \tag{26}$$

Because ξ_{ij}^t and $\gamma_{ij}^{T_{ij}}$ are marginals of q_{ij}^g, both sums in Equation 26 are linear functionals of the joint distribution. Hence, Equation 26 can be rewritten as a single expectation under q_{ij}^g:

$$\widetilde{\mathcal{L}}_{ij}^g\big(q_{ij}^g\big) \stackrel{c}{=} \mathbb{E}_{q_{ij}^g}\Big[\sum_{t=1}^{T_{ij}} \widetilde{\ell}_{ij}^t\big(g_{ij}^{t-1}, g_{ij}^t\big) + \log \phi_{ij}\big(g_{ij}^{T_{ij}}\big)\Big] + \mathcal{H}\big(q_{ij}^g\big) \tag{27}$$

We then maximize Equation 27 over distributions q_{ij}^g supported on the validity sequences $\boldsymbol{g}_{ij} \in \{0,1\}^{T_{ij}}$ with the fixed initial state $g_{ij}^0 = 1$, subject to normalization. Introducing a multiplier λ for the normalization constraint gives the Lagrangian:

$$\mathcal{J}(q_{ij}^g) = \mathbb{E}_{q_{ij}^g}\Big[\sum_{t=1}^{T_{ij}} \widetilde{\ell}_{ij}^t(g_{ij}^{t-1}, g_{ij}^t) + \log \phi_{ij}(g_{ij}^{T_{ij}})\Big] + \mathcal{H}(q_{ij}^g) + \lambda\Big(\sum_{\boldsymbol{g}_{ij}} q_{ij}^g(\boldsymbol{g}_{ij}) - 1\Big) \tag{28}$$

Treating $\{q_{ij}^g(\boldsymbol{g}_{ij})\}$ as free variables and using $\partial\mathcal{H}/\partial q_{ij}^g(\boldsymbol{g}_{ij}) = -\log q_{ij}^g(\boldsymbol{g}_{ij}) - 1$, the stationary condition is:

$$\frac{\partial \mathcal{J}}{\partial q_{ij}^g(\boldsymbol{g}_{ij})} = \sum_{t=1}^{T_{ij}} \widetilde{\ell}_{ij}^t(g_{ij}^{t-1}, g_{ij}^t) + \log \phi_{ij}(g_{ij}^{T_{ij}}) - \log q_{ij}^g(\boldsymbol{g}_{ij}) - 1 + \lambda = 0 \tag{29}$$

Solving Equation 29 for $\log q_{ij}^g(\boldsymbol{g}_{ij})$ gives:

$$\log q_{ij}^g(\boldsymbol{g}_{ij}) = \sum_{t=1}^{T_{ij}} \widetilde{\ell}_{ij}^t(g_{ij}^{t-1}, g_{ij}^t) + \log \phi_{ij}(g_{ij}^{T_{ij}}) + \text{const} \tag{30}$$

Exponentiating Equation 30 yields the Gibbs form:

$$q_{ij}^g(\boldsymbol{g}_{ij}) \propto \phi_{ij}(g_{ij}^{T_{ij}}) \prod_{t=1}^{T_{ij}} \exp\Big\{\widetilde{\ell}_{ij}^t(g_{ij}^{t-1}, g_{ij}^t)\Big\} \tag{31}$$

By Equation 11, $\widetilde{\ell}_{ij}^t(a,b)$ is a weighted sum of the logarithms of the corresponding entries of $\boldsymbol{\Theta}^{(\varnothing)}$, $\boldsymbol{\Theta}_j^{(0)}$, and $\boldsymbol{\Theta}_j^{(1)}$, with weights $1 - m_{ij}^t$, $m_{ij}^t(1 - \bar{\zeta}_{ij}^t)$, and $m_{ij}^t \bar{\zeta}_{ij}^t$, respectively. Therefore:

$$\exp\Big\{\widetilde{\ell}_{ij}^t(a,b)\Big\} = \big[\boldsymbol{\Theta}^{(\varnothing)}\big]_{ab}^{1-m_{ij}^t} \big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab}^{m_{ij}^t(1-\bar{\zeta}_{ij}^t)} \times \big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}^{m_{ij}^t \bar{\zeta}_{ij}^t} \tag{32}$$

Substituting Equation 32 into Equation 31 gives:

$$q_{ij}^g(\boldsymbol{g}_{ij}) \propto \phi_{ij}(g_{ij}^{T_{ij}}) \prod_{t=1}^{T_{ij}} \psi_{ij}^t(g_{ij}^{t-1}, g_{ij}^t), \quad \psi_{ij}^t(a,b) = \big[\boldsymbol{\Theta}^{(\varnothing)}\big]_{ab}^{1-m_{ij}^t} \big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab}^{m_{ij}^t(1-\bar{\zeta}_{ij}^t)} \big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}^{m_{ij}^t \bar{\zeta}_{ij}^t} \tag{33}$$

Equation 33 is exactly the update in Equation 12. Assuming strictly positive transition probabilities, $\mathcal{H}(q_{ij}^g)$ is strictly concave and the remaining terms of Equation 27 are linear in q_{ij}^g. Hence the objective is strictly concave and Equation 33 is its unique maximizer. Every pairwise potential in Equation 33 involves only the adjacent states (g_{ij}^{t-1}, g_{ij}^t), while the verified outcome enters only through the terminal potential $\phi_{ij}(g_{ij}^{T_{ij}})$. Hence q_{ij}^g has a first-order Markov structure over the binary validity states, with ψ_{ij}^t serving as unnormalized pairwise potentials. Its marginals can therefore be computed exactly by forward–backward recursion. Define the forward messages as:

$$F_{ij}^0(a) = \mathbb{I}[a = 1], \qquad F_{ij}^t(b) = \sum_{a \in \{0,1\}} F_{ij}^{t-1}(a)\, \psi_{ij}^t(a,b), \qquad t = 1, \ldots, T_{ij} \tag{34}$$

Here F_{ij}^0 encodes the fixed initial state $g_{ij}^0 = 1$. Define the backward messages as:

$$B_{ij}^{T_{ij}}(a) = \phi_{ij}(a), \qquad B_{ij}^{t-1}(a) = \sum_{b \in \{0,1\}} \psi_{ij}^t(a,b)\, B_{ij}^t(b), \qquad t = T_{ij}, \ldots, 1 \tag{35}$$

By construction, $F_{ij}^t(a)$ accumulates all pairwise potentials up to step t conditioned on $g_{ij}^t = a$, whereas $B_{ij}^t(a)$ accumulates all subsequent pairwise potentials together with the terminal potential conditioned on $g_{ij}^t = a$. Therefore, the normalization constant of Equation 33 is:

$$Z_{ij} = \sum_{\boldsymbol{g}_{ij}} \phi_{ij}(g_{ij}^{T_{ij}}) \prod_{t=1}^{T_{ij}} \psi_{ij}^t(g_{ij}^{t-1}, g_{ij}^t) = \sum_{a \in \{0,1\}} F_{ij}^{T_{ij}}(a)\, \phi_{ij}(a) \tag{36}$$

Summing Equation 33 over all configurations satisfying $g_{ij}^t = a$ gives the single-state marginals:

$$\gamma_{ij}^t(a) = q_{ij}^g\big(g_{ij}^t = a\big) = \frac{F_{ij}^t(a)\, B_{ij}^t(a)}{Z_{ij}} \tag{37}$$

Similarly, summing over all configurations satisfying $(g_{ij}^{t-1}, g_{ij}^t) = (a, b)$ gives the adjacent-pair marginals:

$$\xi_{ij}^t(a,b) = q_{ij}^g\big(g_{ij}^{t-1} = a, g_{ij}^t = b\big) = \frac{F_{ij}^{t-1}(a)\, \psi_{ij}^t(a,b)\, B_{ij}^t(b)}{Z_{ij}} \tag{38}$$

The two sets of marginals are mutually consistent. Using the backward recursion in Equation 35 gives:

$$\sum_{b\in\{0,1\}} \xi_{ij}^t(a,b) = \frac{F_{ij}^{t-1}(a) B_{ij}^{t-1}(a)}{Z_{ij}} = \gamma_{ij}^{t-1}(a) \tag{39}$$

Similarly, using the forward recursion in Equation 34 gives:

$$\sum_{a\in\{0,1\}} \xi_{ij}^t(a,b) = \frac{F_{ij}^t(b) B_{ij}^t(b)}{Z_{ij}} = \gamma_{ij}^t(b) \tag{40}$$

Finally, since each validity variable takes only two possible values, every message update in Equations 34–35 requires a constant number of operations. Therefore, all $\{\gamma_{ij}^t, \xi_{ij}^t\}_{t=1}^{T_{ij}}$ can be computed exactly in $O(T_{ij})$ time.

B.4.2 Derivation of the Validity-Node Update

Fix a task i and a latent node k, and hold the worker validity-sequence factors $\{q_{ij}^g\}_{j\in\mathcal{W}_i}$, the remaining validity-node factors $\{q_i^l\}_{l\neq k}$, and the model parameters $\boldsymbol{\theta}$ fixed. Since q_i^k is a Bernoulli factor, it is fully determined by $\bar{v}_i^k = q_i^k(v_i^k = 1)$. We therefore isolate the terms of $\widetilde{\mathcal{L}}_i$ that depend on $\bar{v}_i^k$. Only two components contribute: the structural prior term and the bounded transition terms.

Consider first the structural contribution in Equation 10. To distinguish the summation indices from the fixed node k, write this term as:

$$\frac{\lambda_v}{2} \sum_{k'<l'} [\widehat{\boldsymbol{A}}_i]_{k'l'} \big(2\bar{v}_i^{k'} - 1\big)\big(2\bar{v}_i^{l'} - 1\big) \tag{41}$$

Because $\widehat{\boldsymbol{A}}_i$ is symmetric and the summation in Equation 41 runs over unordered pairs, node k participates in exactly one term for each $l \neq k$. Therefore, the part that depends on $\bar{v}_i^k$ is:

$$\frac{\lambda_v}{2}\big(2\bar{v}_i^k - 1\big) \sum_{l\neq k} [\widehat{\boldsymbol{A}}_i]_{kl}\big(2\bar{v}_i^l - 1\big) \tag{42}$$

Expanding the factor $(2\bar{v}_i^k - 1)$ shows that Equation 42 is affine in $\bar{v}_i^k$. Up to terms independent of $\bar{v}_i^k$, it becomes:

$$\frac{\lambda_v}{2}\big(2\bar{v}_i^k - 1\big) \sum_{l\neq k} [\widehat{\boldsymbol{A}}_i]_{kl}\big(2\bar{v}_i^l - 1\big) \overset{c}{=} \bar{v}_i^k \lambda_v \sum_{l\neq k} [\widehat{\boldsymbol{A}}_i]_{kl}\big(2\bar{v}_i^l - 1\big) \tag{43}$$

We next consider the transition contribution. Expanding $\bar{\zeta}_{ij}^t = \sum_{l=1}^K \pi_{ij}^{tl} \bar{v}_i^l$ in the Jensen surrogate writes the transition term as a sum over latent nodes:

$$\widetilde{\ell}_{ij}^t(a,b) = \big(1 - m_{ij}^t\big) \log\big[\boldsymbol{\Theta}^{(\varnothing)}\big]_{ab} + m_{ij}^t \sum_{l=1}^K \pi_{ij}^{tl} \Big[\big(1 - \bar{v}_i^l\big) \log\big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab} + \bar{v}_i^l \log\big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}\Big] \tag{44}$$

For the coordinate update of node k, every summand indexed by $l \neq k$ in Equation 44 is constant with respect to $\bar{v}_i^k$. The summand corresponding to $l = k$ is:

$$m_{ij}^t \pi_{ij}^{tk} \Big[\big(1 - \bar{v}_i^k\big) \log \big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab} + \bar{v}_i^k \log \big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab} \Big] \overset{c}{=} \bar{v}_i^k m_{ij}^t \pi_{ij}^{tk} \log \frac{\big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}}{\big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab}} \tag{45}$$

In the surrogate objective, each $\widetilde{\ell}_{ij}^t(a,b)$ is weighted by the pairwise marginal $\xi_{ij}^t(a,b)$ and then aggregated over validity transitions, steps, and workers. Applying these weights to Equation 45, the part of the transition contribution that depends on $\bar{v}_i^k$ is:

$$\bar{v}_i^k \sum_{j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} m_{ij}^t \pi_{ij}^{tk} \sum_{a,b \in \{0,1\}} \xi_{ij}^t(a,b) \log \frac{\big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}}{\big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab}} \tag{46}$$

Both Equation 43 and Equation 46 are linear in $\bar{v}_i^k$. Let η_i^k denote the sum of their coefficients. The only remaining term involving q_i^k is its entropy. Since q_i^k is Bernoulli, its entropy is:

$$\mathcal{H}\big(q_i^k\big) = -\bar{v}_i^k \log \bar{v}_i^k - \big(1 - \bar{v}_i^k\big) \log\big(1 - \bar{v}_i^k\big) \tag{47}$$

Therefore, up to terms independent of $\bar{v}_i^k$, the coordinate objective for latent node k is:

$$\widetilde{\mathcal{L}}_{ik}^v\big(\bar{v}_i^k\big) \overset{c}{=} \bar{v}_i^k \eta_i^k - \bar{v}_i^k \log \bar{v}_i^k - \big(1 - \bar{v}_i^k\big) \log\big(1 - \bar{v}_i^k\big) \tag{48}$$

Differentiating Equation 48 with respect to $\bar{v}_i^k$ gives:

$$\frac{\partial \widetilde{\mathcal{L}}_{ik}^v}{\partial \bar{v}_i^k} = \eta_i^k + \log\big(1 - \bar{v}_i^k\big) - \log \bar{v}_i^k \tag{49}$$

Setting Equation 49 to zero gives the coordinate log-odds condition:

$$\log \frac{\bar{v}_i^k}{1 - \bar{v}_i^k} = \eta_i^k \tag{50}$$

Solving Equation 50 and substituting the structural and transition coefficients gives:

$$\begin{aligned} \bar{v}_i^k &= q_i^k\big(v_i^k = 1\big) = \sigma\big(\eta_i^k\big) \\ \eta_i^k &= \lambda_v \sum_{l \neq k} [\widehat{\boldsymbol{A}}_i]_{kl} \big(2\bar{v}_i^l - 1\big) + \sum_{j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} m_{ij}^t \pi_{ij}^{tk} \sum_{a,b \in \{0,1\}} \xi_{ij}^t(a,b) \log \frac{\big[\boldsymbol{\Theta}_j^{(1)}\big]_{ab}}{\big[\boldsymbol{\Theta}_j^{(0)}\big]_{ab}} \end{aligned} \tag{51}$$

Equation 51 is exactly the update in Equation 13. Assuming strictly positive transition probabilities, η_i^k is finite. The Bernoulli entropy in Equation 48 is strictly concave on $(0,1)$, while the remaining term is linear in $\bar{v}_i^k$. Hence the coordinate objective is strictly concave and Equation 51 is its unique maximizer. For a fixed node k, evaluating the structural contribution in Equation 51 requires $O(K)$ operations. Evaluating the transition contribution requires $O(\sum_{j \in \mathcal{W}_i} T_{ij})$ operations because the inner summation over $(a,b) \in \{0,1\}^2$ has constant size. Therefore, each validity-node coordinate update requires $O\left(K + \sum_{j \in \mathcal{W}_i} T_{ij}\right)$. Finally, the structural contribution for node k depends on the remaining posterior means $\{\bar{v}_i^l\}_{l \neq k}$. Accordingly, the derivation above is a coordinate update in which all remaining validity-node factors are held fixed. We update the K factors sequentially, using the most recently updated values of the other coordinates. Since each coordinate update exactly maximizes $\widetilde{\mathcal{L}}_i$ with respect to one q_i^k while holding the remaining factors fixed, each update is nondecreasing in the surrogate objective.

B.4.3 Derivation of the Model-Parameter Updates

With all variational distributions held fixed, the model parameters are updated by maximizing the surrogate objective together with the parameter prior:

$$\boldsymbol{\theta}^{\text{new}} \in \arg\max_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{N} \widetilde{\mathcal{L}}_i(q_i; \boldsymbol{\theta}) + \log \mathbb{P}(\boldsymbol{\theta}) \right\} \tag{52}$$

To make Equation 52 a well-defined MAP objective, we specify independent conjugate priors for the model parameters. For each worker j, each row of the worker-specific transition matrix $\mathbf{\Theta}_j^{(u)}$, $u \in \{0,1\}$, is assigned a Dirichlet prior:

$$[\mathbf{\Theta}_j^{(u)}]_{a\cdot} \sim \mathrm{Dir}\Big(\beta_{j,a0}^{(u)}, \beta_{j,a1}^{(u)}\Big), \qquad u \in \{0,1\} \tag{53}$$

Each row of the globally tied unmatched-transition matrix is assigned an independent Dirichlet prior:

$$[\mathbf{\Theta}^{(\varnothing)}]_{a\cdot} \sim \mathrm{Dir}\Big(\beta_{a0}^{(\varnothing)}, \beta_{a1}^{(\varnothing)}\Big) \tag{54}$$

Each terminal-observation probability is assigned an independent Beta prior:

$$\omega_j^{(a)} \sim \mathrm{Beta}\big(\beta_{j,a}^{+}, \beta_{j,a}^{-}\big) \tag{55}$$

We assume all prior hyperparameters are at least one. Although $\mathbf{\Theta}^{(\varnothing)}$ is included in every worker-associated parameter block $\boldsymbol{\theta}_j$ for notational convenience, it represents a single globally tied parameter. It therefore contributes only one prior factor through Equation 54 and is updated once using statistics pooled across all workers.

We first derive the transition-matrix updates. Collecting the bounded transition terms of the surrogate objective over all tasks and workers gives:

$$\begin{aligned}\widetilde{\mathcal{L}}_{\mathrm{trans}} = \sum_{i=1}^{N} \sum_{j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} \sum_{a,b} \xi_{ij}^t(a,b) \Big[& (1 - m_{ij}^t) \log[\mathbf{\Theta}^{(\varnothing)}]_{ab} \\ & + m_{ij}^t (1 - \bar{\zeta}_{ij}^t) \log[\mathbf{\Theta}_j^{(0)}]_{ab} + m_{ij}^t \bar{\zeta}_{ij}^t \log[\mathbf{\Theta}_j^{(1)}]_{ab} \Big]\end{aligned} \tag{56}$$

The three coefficients $1 - m_{ij}^t$, $m_{ij}^t(1 - \bar{\zeta}_{ij}^t)$, and $m_{ij}^t \bar{\zeta}_{ij}^t$ are nonnegative and sum to one. They coincide with the exponents of the three transition components in Equation 12. Reading off the coefficients of the corresponding log-transition parameters defines the posterior-weighted counts for $\mathbf{\Theta}_j^{(0)}$:

$$C_{j,ab}^{(0)} = \sum_{i:\, j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} \xi_{ij}^t(a,b) m_{ij}^t (1 - \bar{\zeta}_{ij}^t) \tag{57}$$

Similarly, the posterior-weighted counts for $\mathbf{\Theta}_j^{(1)}$ are:

$$C_{j,ab}^{(1)} = \sum_{i:\, j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} \xi_{ij}^t(a,b) m_{ij}^t \bar{\zeta}_{ij}^t \tag{58}$$

Because the unmatched-transition matrix is globally tied, its posterior-weighted counts are pooled over all tasks and workers:

$$C_{ab}^{(\varnothing)} = \sum_{i=1}^{N} \sum_{j \in \mathcal{W}_i} \sum_{t=1}^{T_{ij}} \xi_{ij}^t(a,b)(1 - m_{ij}^t) \tag{59}$$

Thus, $C_{j,ab}^{(0)}$ and $C_{j,ab}^{(1)}$ allocate the posterior transition mass explained by the latent consensus structure according to the inferred validity of the aligned latent nodes, whereas $C_{ab}^{(\varnothing)}$ collects the transition mass assigned to the unmatched component. Consider one row a of the worker-specific transition matrices. Let

$$p_{j,a}^{(u)} = [\mathbf{\Theta}_j^{(u)}]_{a1}, \qquad [\mathbf{\Theta}_j^{(u)}]_{a0} = 1 - p_{j,a}^{(u)}, \qquad u \in \{0,1\} \tag{60}$$

The monotonicity constraint in the transition model requires $p_{j,a}^{(0)} \le p_{j,a}^{(1)}$. Combining the posterior-weighted counts with the Dirichlet priors, the part of the MAP objective involving this row is:

$$\begin{aligned}\sum_{u \in \{0,1\}} \Big[& \Big(C_{j,a1}^{(u)} + \beta_{j,a1}^{(u)} - 1\Big) \log p_{j,a}^{(u)} \\ & + \Big(C_{j,a0}^{(u)} + \beta_{j,a0}^{(u)} - 1\Big) \log(1 - p_{j,a}^{(u)}) \Big],\end{aligned} \qquad \text{subject to} \quad p_{j,a}^{(0)} \le p_{j,a}^{(1)} \tag{61}$$

Ignoring the ordering constraint temporarily, the unconstrained MAP estimates are given as:

$$\widetilde{p}_{j,a}^{(u)} = \frac{C_{j,a1}^{(u)} + \beta_{j,a1}^{(u)} - 1}{\sum_{b\in\{0,1\}} \left(C_{j,ab}^{(u)} + \beta_{j,ab}^{(u)} - 1\right)}, \qquad u \in \{0,1\} \tag{62}$$

If $\widetilde{p}_{j,a}^{(0)} \leq \widetilde{p}_{j,a}^{(1)}$, the unconstrained estimates also solve the constrained problem. Otherwise, the ordering constraint is active at the optimum. Because the row-wise objective is concave in $(p_{j,a}^{(0)}, p_{j,a}^{(1)})$, the constrained optimum lies on the boundary $p_{j,a}^{(0)} = p_{j,a}^{(1)}$. Maximizing the pooled objective over this common value gives:

$$\bar{p}_{j,a} = \frac{\sum_{u\in\{0,1\}} \left(C_{j,a1}^{(u)} + \beta_{j,a1}^{(u)} - 1\right)}{\sum_{u\in\{0,1\}} \sum_{b\in\{0,1\}} \left(C_{j,ab}^{(u)} + \beta_{j,ab}^{(u)} - 1\right)} \tag{63}$$

The constrained update is therefore defined as:

$$p_{j,a}^{(u)} = \begin{cases} \widetilde{p}_{j,a}^{(u)}, & \widetilde{p}_{j,a}^{(0)} \leq \widetilde{p}_{j,a}^{(1)}, \\ \bar{p}_{j,a}, & \widetilde{p}_{j,a}^{(0)} > \widetilde{p}_{j,a}^{(1)}, \end{cases} \qquad u \in \{0,1\}, \tag{64}$$

and the corresponding transition probabilities are:

$$[\mathbf{\Theta}_j^{(u)}]_{a1} = p_{j,a}^{(u)}, \qquad [\mathbf{\Theta}_j^{(u)}]_{a0} = 1 - p_{j,a}^{(u)} \tag{65}$$

Applying the same derivation to the globally tied unmatched-transition matrix gives:

$$[\mathbf{\Theta}^{(\varnothing)}]_{ab} = \frac{C_{ab}^{(\varnothing)} + \beta_{ab}^{(\varnothing)} - 1}{\sum_{b'\in\{0,1\}} \left(C_{ab'}^{(\varnothing)} + \beta_{ab'}^{(\varnothing)} - 1\right)} \tag{66}$$

When the effective counts in a row are nonzero, these updates maximize the corresponding row-wise MAP objective. If a row receives zero posterior-weighted transition mass and both corresponding Dirichlet hyperparameters equal one, the objective is constant over that row and the maximizer is not unique. In this degenerate case, we retain the previous row. Choosing hyperparameters strictly greater than one instead provides positive pseudocounts and keeps the transition probabilities in the interior of the probability simplex. We next derive the terminal-observation updates. For worker j and terminal validity state a, define the posterior-weighted correct and incorrect outcome counts as:

$$R_{j,a}^{+} = \sum_{i:\, j\in\mathcal{W}_i} \gamma_{ij}^{T_{ij}}(a) o_{ij}, \qquad R_{j,a}^{-} = \sum_{i:\, j\in\mathcal{W}_i} \gamma_{ij}^{T_{ij}}(a)(1 - o_{ij}) \tag{67}$$

Combining the terminal terms of the surrogate objective with the Beta prior, the part of the MAP objective involving $\omega_j^{(a)}$ is:

$$\left(R_{j,a}^{+} + \beta_{j,a}^{+} - 1\right) \log \omega_j^{(a)} + \left(R_{j,a}^{-} + \beta_{j,a}^{-} - 1\right) \log(1 - \omega_j^{(a)}) \tag{68}$$

Ignoring the ordering constraint between the two terminal states temporarily, setting the derivative of Equation 68 to zero gives:

$$\frac{R_{j,a}^{+} + \beta_{j,a}^{+} - 1}{\omega_j^{(a)}} - \frac{R_{j,a}^{-} + \beta_{j,a}^{-} - 1}{1 - \omega_j^{(a)}} = 0 \tag{69}$$

Solving Equation 69 gives the unconstrained MAP estimate:

$$\widetilde{\omega}_j^{(a)} = \frac{R_{j,a}^{+} + \beta_{j,a}^{+} - 1}{R_{j,a}^{+} + R_{j,a}^{-} + \beta_{j,a}^{+} + \beta_{j,a}^{-} - 2} \tag{70}$$

The terminal model additionally imposes the ordering constraint $\omega_j^{(0)} \leq \omega_j^{(1)}$. If the two unconstrained estimates in Equation 70 satisfy this ordering, they also solve the constrained problem.

Otherwise, the ordering constraint is active at the optimum. Because the objective is concave in $(\omega_j^{(0)}, \omega_j^{(1)})$, the constrained optimum then lies on the boundary $\omega_j^{(0)} = \omega_j^{(1)}$. Maximizing the pooled objective over this commonvalue gives:

$$\omega_j^{(0)} = \omega_j^{(1)} = \frac{\sum_{a\in\{0,1\}} \left(R_{j,a}^{+} + \beta_{j,a}^{+} - 1\right)}{\sum_{a\in\{0,1\}} \left(R_{j,a}^{+} + R_{j,a}^{-} + \beta_{j,a}^{+} + \beta_{j,a}^{-} - 2\right)} \tag{71}$$

C ALGORITHM

Algorithm 2 summarizes how the proposed framework infers step-level validity from crowdsourced worker reasoning trajectories.

Algorithm 2 Inference of Step-Level Validity from Crowdsourced Reasoning Trajectories

Require: Process-supervised dataset $\mathcal{D}_{\mathrm{proc}}$, frozen encoder f_ϕ, latent size K, transport parameters α, ρ, prior strength λ_v, and parameter-prior hyperparameters
Ensure: Process-annotated dataset $\mathcal{D}_{\mathrm{aug}}$ and worker reliability parameters $\widehat{\boldsymbol{\theta}}$

for $i = 1$ to N **do**
  for $j \in \mathcal{W}_i$ **do**
    Encode τ_{ij} into $\boldsymbol{R}_{ij}$ and $\boldsymbol{\Delta}_{ij}$ via f_ϕ
  end for
  Estimate $\widehat{\Psi}_i$ and $\{\widehat{\boldsymbol{\Gamma}}_{ij}\}_{j\in\mathcal{W}_i}$ by Algorithm 1
  Compute $\mathcal{A}_i = \{(m_{ij}^t, \boldsymbol{\pi}_{ij}^t)\}$ via equation 4
end for
Initialize $\boldsymbol{\theta}$ and $\{\bar{v}_i^k\}$
repeat
  for $i = 1$ to N **do**
    repeat
      for $j \in \mathcal{W}_i$ **do**
        Update q_{ij}^g via equation 12 and obtain $\gamma_{ij}^t, \xi_{ij}^t$ by forward–backward recursions
      end for
      for $k = 1$ to K **do**
        Update $\bar{v}_i^k$ via equation 13
      end for
    until $\widetilde{\mathcal{L}}_i$ converges
  end for
  for $j = 1$ to M **do**
    Update $\{\boldsymbol{\Theta}_j^{(u)}\}_{u\in\{0,1\}}$ via equation 64
    Update $\{\omega_j^{(a)}\}_{a\in\{0,1\}}$ via equation 70, using equation 71 when the ordering constraint is active
  end for
  Update the globally tied $\boldsymbol{\Theta}^{(\varnothing)}$ via equation 66
until $\sum_{i=1}^N \widetilde{\mathcal{L}}_i(q_i;\boldsymbol{\theta}) + \log \mathbb{P}(\boldsymbol{\theta})$ converges
$\mathcal{D}_{\mathrm{aug}} \leftarrow \{(x_i, \{\tau_{ij}, o_{ij}, \{\gamma_{ij}^t(1)\}_{t=1}^{T_{ij}}\}_{j\in\mathcal{W}_i})\}_{i=1}^N$
return $\mathcal{D}_{\mathrm{aug}}$ and $\widehat{\boldsymbol{\theta}}$

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 LLM AS CROWDWORKERS AND TRAJECTORY COLLECTION (ADDITIONAL DETAILS)

D.1.1 LLM WORKER MODELS

We instantiate LLM workers from a diverse pool of models spanning multiple providers and model families, which are summarized in Table 4.

Table 4: LLM models used to instantiate crowdworkers. Multiple workers may be instantiated from the same model using different prompt configurations.

Provider / Family	Model
OpenAI	`gpt-5.6-luna`
	`gpt-5-mini`
	`gpt-5-nano`
DeepSeek	`deepseek-v4-flash`
	`deepseek-v4-pro`
Moonshot / Kimi	`kimi-k2.6`
	`kimi-k3`
Google Gemini	`gemini-2.5-flash-lite`
	`gemini-2.5-flash`
	`gemini-3.5-flash-lite`
Anthropic Claude	`claude-sonnet-4-6`
	`claude-haiku-4-5-20251001`
Meta Llama	`Llama-3.1-8B-Instruct`
	`Llama-4-Maverick-17B-128E`
	`Llama-4-Scout-17B-16E-Instruct`

Table 5: Prompt components used to instantiate LLM crowdworkers.

Component	**Purpose**	**Prompt Content**
Shared prompt structure	Specify the common output structure and independent reasoning requirement.	Solve the problem independently and provide the reasoning as clearly separated numbered steps, with one coherent reasoning move per step. Use the designated solution route and maintain the fixed worker style across tasks. Finish with `Final Answer: <answer>`.
Persistent persona	Induce stable worker-specific reasoning habits across assigned tasks.	Maintain this fixed worker style across tasks: {`PERSISTENT_PERSONA`}. We use ten personas that vary verification intensity, representation preference, search behavior, reasoning granularity, and other persistent reasoning characteristics.
Domain-specific strategy	Vary the solution route within each reasoning domain.	Use the following solution strategy for this task: {`DOMAIN_STRATEGY`}. For each task source, we define ten strategies covering different solution routes such as equation-first, backward reasoning, constraint propagation, countermodel construction, and verification-oriented reasoning.

D.1.2 Worker Prompt Construction

Each LLM worker is instantiated using a prompt composed of a persistent worker persona, and a domain-specific reasoning strategy. For each worker, the persona and strategy are independently sampled when the worker is instantiated and remain fixed across all questions assigned to that worker within the corresponding task source. Table 5 summarizes the prompt components.

D.1.3 Worker Assignment and Collection Statistics

Table 6 summarizes the detailed statistics of questions used for trajectory collection and the resulting worker trajectories and reasoning steps. Each question is assigned to 10 distinct workers. To evaluate step-validity recovery performance, we randomly sample 100 collected questions from each task source and annotate all 10 trajectories associated with each sampled question. These annotated trajectories are used exclusively as the evaluation set. Table 7 reports Krippendorff's α and raw agreement between annotators on reasoning step-validity labels.

D.2 Additional Details for Baselines and Evaluation setup

For *Self-Check*, we estimate the validity of each reasoning step from its semantic agreement with the other worker trajectories for the same question. Specifically, for each target step, we identify the

Table 6: Detailed statistics of trajectory collection and step-validity recovery evaluation across the five task sources.

Dataset	Collection Questions	Collected Trajectories	Eval Questions	Eval Trajectories
MATH	450	4,500	100	1,000
GSM8K	300	3,000	100	1,000
LogiQA	600	6,000	100	1,000
P-FOLIO	600	6,000	100	1,000
ZebraLogic	200	2,000	100	1,000
Total	**2,150**	**21,500**	**500**	**5,000**

Table 7: Inter-annotator agreement on reasoning step-validity annotations across the five task sources.

Metric	MATH	GSM8K	LogiQA	P-FOLIO	ZebraLogic
Krippendorff's α	0.929	0.942	0.798	0.829	0.761
Raw agreement (%)	97.0	97.2	89.9	91.5	91.0

most semantically similar step from each other worker trajectory using cosine similarity between step embeddings and average the resulting similarity scores across workers. For *Dawid–Skene (DS)*, we first group semantically similar reasoning steps within each question using agglomerative clustering over the step embeddings. Each semantic cluster is treated as a latent binary item, and the verified final outcome of a worker trajectory is used as its noisy observation for a cluster when the trajectory contains at least one step assigned to that cluster. Each trajectory contributes at most one observation to each cluster. Dawid–Skene is then used to estimate worker reliability and the posterior validity of each semantic cluster, which is assigned back to the corresponding reasoning steps. For step-validity evaluation, we randomly sample 100 questions from each task source and obtain expert step-validity annotations for all collected trajectories associated with these questions. We further split these 100 questions into 10 validation questions and 90 test questions. The validation set is used for hyperparameter selection for all compared methods, while the test set is held out for final evaluation. The same validation and test splits are used across all methods. Expert step-validity labels from the test set are used only to compute AUROC and AUPRC and are never used for model fitting and hyperparameter selection. Consistent with our transductive recovery setting, the observed reasoning trajectories and their verified final-answer outcomes on the test questions remain available to the inference methods.

D.3 Hyperparameter Sensitivity

We vary the number of latent consensus nodes K, the semantic–structural trade-off α, the marginal-relaxation strength ρ, and the structural validity regularizer λ_v one at a time around the reference configuration. Unless it is being varied, the FUGW configuration is K=12, α=0.5, ρ=1, and λ_v=0.15 for all five datasets. Each entry in Table 8 is the mean AUPRC-Error over five optimization seeds.

Our method is generally robust to moderate hyperparameter changes. The reference configuration remains competitive across all five datasets, although the best setting slightly varies by task. Performance is most sensitive to K, the number of nodes used to represent the latent consensus process, and λ_v, which controls how strongly connected latent nodes are encouraged to have similar validity.

Too small a value of K forces the model to merge distinct reasoning states into the same consensus node, making it harder to represent the full reasoning process. An overly large λ_v makes neighboring consensus nodes too similar, so changes in validity become harder to detect. For example, reducing K from 12 to 4 lowers AUPRC from 0.807 to 0.618 on GSM8K. The best value of λ_v varies across datasets, but the shared reference value of 0.15 remains within 0.01 AUPRC of the best result on every dataset, while larger values generally perform worse. By comparison, the reference values of α and ρ remain competitive across datasets, suggesting less sensitivity to moderate changes in the alignment trade-off and marginal relaxation.

Table 8: Hyperparameter sensitivity across five reasoning domains in terms of **AUPRC (%)**. Higher values indicate better performance. Each block varies one hyperparameter while holding the remaining parameters fixed at the reference configuration used in the main experiments. The reference value of each hyperparameter is highlighted in bold. All results are averaged over five runs.

Axis	Value	MATH	GSM8K	LogiQA	P-FOLIO	ZebraLogic
K	4	69.1	61.8	67.3	69.1	42.6
	8	76.1	70.5	81.6	78.9	44.7
	12	**76.8**	**80.7**	**81.1**	**80.2**	**47.4**
	16	76.4	79.2	80.9	78.9	46.6
	24	81.1	75.7	80.4	72.9	43.5
α	0	74.1	76.4	80.2	78.0	44.8
	0.25	73.0	80.7	81.1	79.5	46.9
	0.5	**76.8**	**80.7**	**81.1**	**80.2**	**47.4**
	0.75	83.5	84.0	79.9	78.2	46.9
	1	74.6	77.9	77.2	74.2	42.7
ρ	0.1	73.0	83.5	69.6	75.8	45.4
	0.3	77.6	83.3	76.6	76.9	46.5
	1	**76.8**	**80.7**	**81.1**	**80.2**	**47.4**
	3	65.6	80.8	80.8	79.4	46.2
	10	62.5	81.0	80.5	79.4	45.7
λ_v	0	77.0	78.3	81.8	79.5	45.2
	0.1	77.6	79.6	81.7	79.6	45.3
	0.15	**76.8**	**80.7**	**81.1**	**80.2**	**47.4**
	0.5	75.1	75.0	79.4	77.3	45.5
	1	67.9	58.7	76.8	76.9	43.6

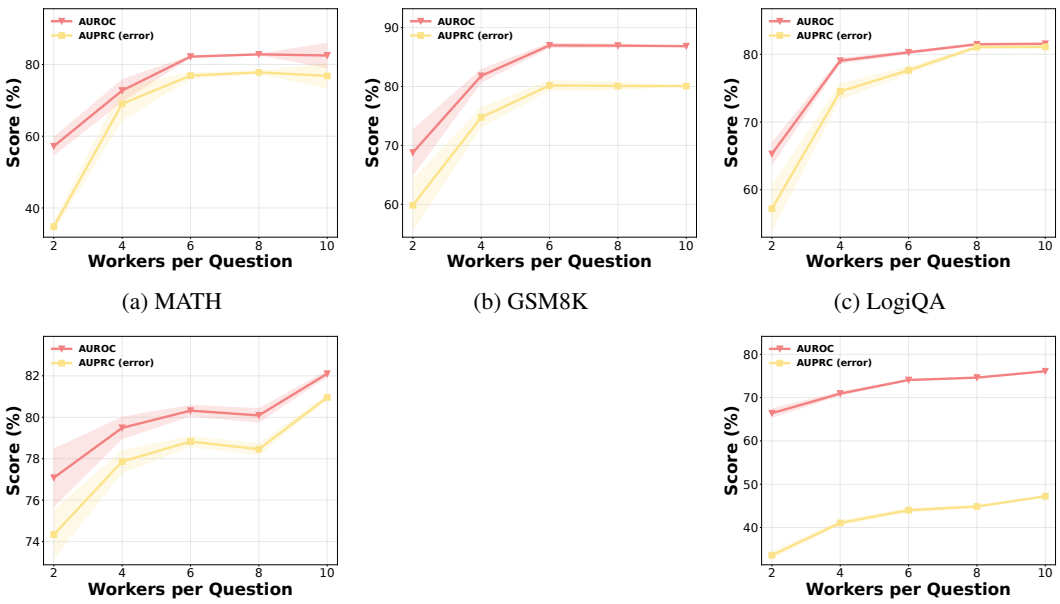
Table 9: Comparing step-validity recovery performance of the proposed method with different embedding encoders across five reasoning domains in terms of **AUPRC (%)**. Higher values indicate better performance. The encoder used to construct the claim and transition embeddings is varied while all remaining settings are held fixed at those used in the main experiments. Parentheses report the change in percentage points relative to the OpenAI reference encoder. All results are averaged over 5 runs.

Dataset	OpenAI (1024-d)	BGE (1024-d)	E5 (1024-d)	MPNet (768-d)
MATH	76.8	75.7 (−1.1)	77.9 (+1.1)	73.3 (−3.5)
GSM8K	80.7	80.3 (−0.4)	80.8 (+0.1)	79.7 (−1.0)
LogiQA	81.1	81.6 (+0.5)	82.1 (+1.0)	81.8 (+0.7)
P-FOLIO	80.2	80.3 (+0.1)	79.8 (−0.4)	79.9 (−0.3)
ZebraLogic	47.4	51.1 (+3.7)	51.1 (+3.7)	51.7 (+4.3)

D.4 Embedding-Encoder Ablation

To examine whether our method depends on the text encoder used to construct the claim and transition embeddings, we replace f_ϕ while holding all other model and inference settings fixed. In Table 9, OpenAI denotes `text-embedding-3-large`, BGE denotes `bge-large-en-v1.5`, E5 denotes `e5-large-v2`, and MPNet denotes `all-mpnet-base-v2`. We use each encoder at its native dimensionality and do not retune the other hyperparameters. OpenAI, BGE, and E5 produce 1024-dimensional embeddings, while MPNet produces 768-dimensional embeddings. All encoders produce scores for the same set of labeled steps.

Our method performs consistently well across embedding encoders. On GSM8K, LogiQA, and P-FOLIO, changing the encoder changes AUPRC by at most 0.010. The variation is larger on MATH and ZebraLogic, but even the weakest encoder variant outperforms the strongest baseline in the



(a) MATH (b) GSM8K (c) LogiQA

(d) P-FOLIO (e) ZebraLogic

Figure 3: Effect of the number of workers per question on step-validity recovery. We report AUROC and AUPRC when varying the number of available worker trajectories from 2 to 10. Results are averaged over five runs, and shaded regions indicate the standard error of the deviation.

main evaluation by at least 0.071 on every dataset. No single encoder performs best across all five datasets: for example, MPNet performs best on ZebraLogic but worst on MATH. These results show that different encoders have task-specific trade-offs and that the gains of our approach are not tied to a particular embedding encoder.

D.5 Effect of the Number of Workers

We further examine how step-validity recovery changes with the number of worker trajectories available for each question. We vary the number of workers per question from 2 to 10 and evaluate the proposed method on the eval set from each task source. For each setting, we repeat the experiment 5 times and report the mean AUROC and AUPRC, with the shaded regions indicating the standard error of deviation. To simulate smaller worker pools, for each question we uniformly subsample 2, 4, 6, or 8 workers from the original 10-worker pool, while the 10-worker setting uses the full set of collected trajectories. Figure 3 shows that recovery performance generally improves as more worker trajectories become available. The largest gains typically occur when increasing the worker pool from 2 to 4 or 6 workers. For example, on MATH, increasing the number of workers from 2 to 6 improves AUROC from 57.22% to 82.17% and AUPRC from 34.79% to 76.89%. Similarly, on GSM8K, AUROC increases from 68.78% to 86.97% and AUPRC from 59.85% to 80.18%.

D.6 Ablation Analysis of Target Outcomes and Peer Evidence

We examine how a trajectory's final-answer correctness and cross-trajectory evidence from other workers on the same question affect step-validity recovery. Specifically, *w/o both* removes both information sources. *w/o peer evidence* removes cross-trajectory alignment and latent consensus while retaining the target trajectory's final-answer correctness. *w/o target outcome* excludes the target trajectory's final-answer correctness from inference and parameter estimation, while keeping the other workers' trajectories and outcomes. *Full* uses both information sources.

Table 10 shows that removing cross-trajectory evidence reduces both AUROC and AUPRC across all five task sources. For example, on GSM8K, AUROC decreases from 86.0% to 61.9%, and AUPRC decreases from 80.7% to 43.1%. We observed a mixed effect when removing the target outcome

Table 10: Comparing step-validity recovery performance of the proposed method under different ablations of the target trajectory's own outcome and same-question peer evidence across five reasoning domains in terms of **AUROC (%)** and **AUPRC (%)**. Higher values indicate better performance for both metrics. The best result for each metric is highlighted in bold. All results are averaged over 5 runs.

Variant	MATH		GSM8K		LogiQA		P-FOLIO		ZebraLogic	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
w/o both	50.2	30.1	49.9	35.2	55.5	44.9	54.8	49.9	66.6	30.9
w/o peer evidence	60.3	37.2	61.9	43.1	69.1	62.2	64.6	61.5	69.8	33.5
w/o target outcome	75.5	42.7	**87.9**	**86.3**	79.6	75.6	78.3	**81.9**	75.7	**50.7**
Full	**82.5**	**76.8**	86.0	80.7	**81.6**	**81.1**	**82.1**	80.2	**76.7**	47.4

Table 11: Comparing step-validity recovery performance of the proposed method and simplified static aggregation controls using the same FUGW-derived cross-trajectory correspondence across five reasoning domains in terms of **AUROC (%)** and **AUPRC (%)**. Higher values indicate better performance for both metrics. The best result for each metric is highlighted in bold. All results are averaged over 5 runs.

Variant	MATH		GSM8K		LogiQA		P-FOLIO		ZebraLogic	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
FUGW+Vote	58.1	40.0	63.3	40.7	58.4	53.4	55.7	50.8	58.6	26.6
FUGW+DS	73.9	52.9	74.2	65.4	66.1	55.8	61.9	59.6	61.7	29.1
Full	**82.5**	**76.8**	**86.0**	**80.7**	**81.6**	**81.1**	**82.1**	**80.2**	**76.7**	**47.4**

from the inference and parameter estimation of the proposed method. Both metrics decreased on MATH and LogiQA but increased on GSM8K. These results highlight the contribution of cross-trajectory evidence beyond a trajectory's final-answer correctness, whose additional benefit varies across task sources and metrics.

D.7 Ablation Analysis of Structured Latent-State Inference

The ablation in Appendix D.6 examines which evidence sources contribute to step-validity recovery. We further ask a complementary question that whether the gains of the full method can be explained primarily by establishing cross-trajectory correspondence, or whether the proposed structured process inference that models sequential evolution of latent step validity provides additional benefit once such correspondence is available. To isolate these effects, we construct two simplified inference methods that reuse the same FUGW-based cross-trajectory alignment as the full method, but replace the proposed latent-state inference with static crowd aggregation. *FUGW+Vote* propagates each worker's verified final outcome correctness to the latent consensus nodes according to its FUGW alignment weights, and *FUGW+DS* uses the same FUGW-aligned evidence, but replaces uniform aggregation with Dawid–Skene aggregation to account for heterogeneous worker reliability. As shown in Table 11, *FUGW+DS* consistently outperforms *FUGW+Vote* across all five reasoning domains, indicating that accounting for heterogeneous worker reliability remains beneficial after cross-trajectory correspondence has been established. However, both static aggregation methods remain substantially below the full method. For example, on LogiQA, AUPRC increases from 55.8% with *FUGW+DS* to 81.10% with the full method. These results show that the improvement of the proposed method cannot be explained by cross-trajectory alignment or reliability-aware static aggregation alone. While establishing correspondence across heterogeneous reasoning trajectories makes conventional crowd-aggregation signals applicable, explicitly modeling latent step validity and its sequential evolution provides substantial additional benefit for recovering process supervision.

D.8 Supervised Fine-Tuning and Evaluation Details

We perform supervised fine-tuning for both Llama-3.1-8B-Base and Mistral-7B-v0.1 using AdamW with a learning rate of 1×10^{-5}. We train for 3 epochs with a batch size of 32 on 8 H100 GPUs, using bfloat16 precision and a maximum sequence length of 8192. The same training configuration is used across both backbone models and all compared fine-tuning methods. For the proposed method, the inferred posterior validity probability $\gamma_{ij}^{t}(1)$ is used as the supervision weight for the reasoning tokens associated with step t of trajectory τ_{ij}. All compared methods use the same underlying training configuration, differing only in their data-selection or supervision weighting strategy. For

evaluation, we use greedy decoding with temperature 0 and generate one response per prompt, with a maximum generation length of 4096 tokens. *Pass@1 accuracy* is computed by extracting the final answer from each generated response and evaluating it using the corresponding dataset-specific answer verifier. All downstream results are averaged over 3 runs.