Privacy-Preserving Heterogeneous Multi-LLM Federated Inference for Cognitive Diagnosis

Yagna Manasa Boyapati, Chong Yu, Tianyu Jiang, Justin Zhan
Department of Computer Science
University of Cincinnati
boyapaya@mail.uc.edu, yuc5@ucmail.uc.edu
tianyu.jiang@uc.edu, zhanjt@ucmail.uc.edu

Abstract

Significant challenges remain in AI-driven educational systems in balancing privacy preservation with accurate cognitive diagnosis. In order to overcome this, we propose a federated inference framework in which several commercial LLM APIs collaborate without requiring access to raw student data or proprietary model internals. Using multiple federated entities, such as LLaMA-3.3-70B, GPT-4o-mini, and Claude-3-Haiku, our framework builds upon the heterogeneous multi-LLM architecture. The predictions generated by these entities are combined with ε-local differential privacy by adding Laplace noise locally to each entity's prediction output before aggregation, while residual-based aggregation mitigates model heterogeneity. Our approach is predicated on an honest-but-curious trust paradigm in which API providers are presumed not to abuse submitted queries, and our differential privacy mechanism shields the published diagnostic results from external inference. We conduct rigorous privacy–utility analysis showing strong privacy guarantees with minimal accuracy loss, and extensive real-world evaluations across three educational benchmarks confirm the framework's practical usability and cross-domain generalizability.

1 Introduction

The growing application of artificial intelligence in education has created unparalleled opportunities for individualized and adaptive learning, but it also raises significant challenges related to data privacy and the accuracy of diagnostic feedback for end users—namely, students (Wang et al., 2022; Liu, 2021). Cognitive Diagnosis Models (CDMs), which infer a student's mastery of specific concepts or skills, are foundational components of intelligent tutoring systems and adaptive learning environments (Wang et al., 2020; Tatsuoka, 1983). Despite their promise, two critical challenges limit the widespread deployment of CDM-based systems.

Traditional CDM approaches rely on the centralized processing of sensitive student data, raising serious privacy concerns. When learning institutions aggregate detailed information about students' learning patterns and trajectories on a central server, the risk of data leakage, unauthorized access, or misuse increases substantially (Stephanie et al., 2022; Hu et al., 2020). Regulatory pressures, such as GDPR and FERPA (European Parliament and Council, 2016; U.S. Department of Education, 2014), have heightened awareness of these issues and underscore the need for privacy-preserving solutions that can support the full range of cognitive diagnosis tasks.

Although Federated Learning (FL) enables collaborative model training without centralizing raw data (Elhussein, 2025), and Differential Privacy (DP) provides strong guarantees for protecting individual-level information (Xin et al., 2024; Fu et al., 2026), no existing work integrates these techniques into LLMs to address privacy concerns in CDM-based systems. The current LLM-based cognitive diagnosis approaches typically rely on a single language model to evaluate students' knowledge states. While large language models show great promise in educational applications (Dong et al., 2025; Chen et al., 2025), single-model approaches are limited in robustness, adaptability, and diagnostic reliability. Prior work highlights the strengths of GPT-4 (Achiam et al., 2023), Claude (Anthropic, 2024), and LLaMA (Touvron et al., 2023) across mathematical reasoning, linguistic understanding, and pattern recognition (Wang et al., 2025). However, these studies evaluate models independently, without leveraging collaborative inference across heterogeneous LLMs.

Furthermore, privacy-preserving use of LLMs in federated learning has only recently been explored (Chen et al., 2024; Tran et al., 2025), and no existing work investigates multi-LLM collaboration for cognitive diagnosis under privacy constraints.

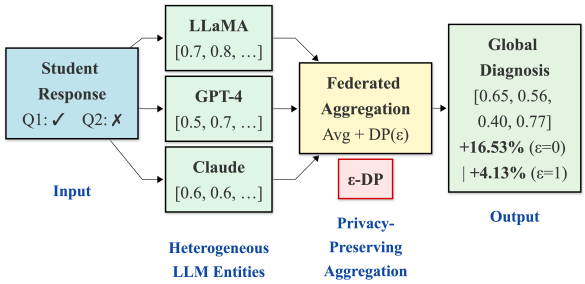

Figure 1: Privacy-preserving heterogeneous multi-LLM federated framework with differentially private aggregation.

To address these challenges, we propose the first privacy-preserving heterogeneous multi-LLM federated inference framework for cognitive diagnosis. As illustrated in Figure 1, LLaMA-3.3-70B (via Groq API) (Groq, 2024), GPT-4o-mini (Achiam et al., 2023), and Claude-3-Haiku (Anthropic, 2024) serve as federated participants that collaboratively infer student knowledge states. By leveraging the complementary strengths of each model, our heterogeneous design eliminates the need for centralized training while simultaneously improving privacy, scalability, and robustness in real-world educational deployments.

We extend the paradigm of federated inference to heterogeneous multi-LLM architectures, where each model independently predicts knowledge states and contributes a complementary diagnostic perspective (Liu and Dong, 2024). Unlike traditional FL methods, which aggregate model parameters or gradients and assume homogeneous models (McMahan et al., 2017; Elhussein, 2025), our framework operates entirely at the prediction level: LLM models are frozen and only accessible via black-box APIs, with only their output predictions shared and aggregated. This prediction-level, federated inference approach leverages model heterogeneity to yield more robust global cognitive assessments. Privacy-preserving aggregation uses Laplace mechanisms (Dwork et al., 2006; Lin et al., 2023; Liu et al., 2025), with calibrated noise applied locally at each entity's prediction output before aggregation, enabling institutions to use multiple commercial LLM APIs without exposing raw student data, a concern highlighted in blockchain-enabled FL studies (Jia et al., 2021). We evaluate on three state-of-the-art benchmarks. Results show strong generalization across educational domains, with tight privacy guarantees and minimal utility loss, consistent with recent advances in DP-FL using Laplacian noise (Liang et al., 2024) and heterogeneous DP mechanisms (Lin et al., 2023).

It is critical to highlight that our privacy system uses an honest-but-curious trust model at the aggregate layer. Our ε-local differential privacy mechanism protects published knowledge-state diagnostic vectors from re-identification or membership inference by external parties accessing the outputs. It does not prevent data collection by commercial API providers (GPT-4o-mini, Claude-3-Haiku, and LLaMA-3.3-70B via Groq), which may log submitted queries in accordance with their respective policies. In Section 3.3, we describe mitigating options for deployments that require full end-to-end privacy, such as locally hosted open-source models and prompt data minimization.

Our key contributions are: (1) a novel, provider-agnostic federated inference framework coordinating commercial LLM APIs without access to proprietary model internals or training data; (2) a heterogeneous multi-LLM architecture treating LLaMA, GPT-4, and Claude as independent federated entities for cognitive diagnosis with locally protected student data; (3) rigorous privacy–utility analysis showing strong privacy preservation with minimal utility loss under ε-local differential privacy; and (4) extensive evaluation on three educational benchmarks validating practical implementability and cross-domain generalizability. Our code and experimental resources are publicly available at `https://github.com/manasa2107/privacy-federated-llm-cognitive-diagnosis`.

2 Related Work

Our research integrates cognitive diagnosis models, federated learning, differential privacy, and large language models to advance educational applications.

Cognitive Diagnosis Models and LLM-based Diagnosis. Traditional CDM approaches, including Item Response Theory (IRT) (Baker, 2001; Yao et al., 2026) and the Deterministic Input, Noisy-AND gate (DINA) model (De La Torre, 2009), require substantial human intervention for feature engineering and struggle to capture complex skill relationships. Foundational knowledge tracing models including BKT (Corbett and Anderson, 1994), DKT (Piech et al., 2015), and EKT (Liu et al., 2019) improved sequential student modeling. Neural approaches such as NeuralCD (Wang et al., 2022) and explainable CDM frameworks

(Yang et al., 2022) further improved predictive accuracy via embedding-based interaction modeling on large-scale educational datasets. More recently, LLMs have been applied to cognitive diagnosis: Dong et al. (2025) augment CDMs with LLM-generated diagnostic signals via contrastive learning and mask-reconstruction, while Chen et al. (2025) evaluate GPT-4, Claude, and LLaMA individually across mathematical concepts demonstrating complementary strengths across models but without privacy or federation considerations. Beyond cognitive diagnosis, LLMs have shown broad promise across real-world educational tasks. Säuberli et al. (2025) evaluate the psychometric plausibility of LLM responses in standardized assessments using IRT and classical test theory, finding that LLM response distributions diverge from human patterns, motivating hybrid frameworks that combine LLM inference with structured diagnostic models. Kim et al. (2025) propose knowledge tracing models that leverage LLM-extracted skill signals from student questions to improve performance prediction in programming education. Pal Chowdhury et al. (2025) identify critical challenges in deploying LLMs for education, emphasizing that LLMs must align with pedagogical principles, integrate with knowledge tracing systems, and incorporate robust data privacy safeguards. Vanzo et al. (2025) and Luo et al. (2024) further demonstrate LLMs' growing role in real-world educational applications, including homework tutoring and question generation.

Federated Learning with Differential Privacy. Federated Learning enables collaborative model training without centralizing raw data (Elhussein, 2025), making it well-suited for privacy-sensitive educational settings. Prior FL work has addressed statistical heterogeneity across participants (McMahan et al., 2017; Li et al., 2020), and applications in healthcare have combined FL with blockchain and differential privacy (Stephanie et al., 2022; Jia et al., 2021), though often at significant computational overhead. Differential Privacy provides formal guarantees for protecting individual-level information via calibrated noise mechanisms (Dwork et al., 2006). Foundational DP-FL works establish client-level privacy (Geyer et al., 2017) and gradient-based noise mechanisms (Abadi et al., 2016), which our framework adapts to the prediction-level aggregation setting. Recent work has explored DP in federated setups (Xin et al., 2024; Fu et al., 2026), for LLM fine-tuning (Liu et al., 2025), and in personalized FL (Hu et al., 2020; Tran et al., 2025). Our solution leverages the Laplace mechanism (Liang et al., 2024) with heterogeneous privacy budgets (Lin et al., 2023). Most existing FL approaches assume homogeneous neural networks and aggregate model parameters or gradients. In contrast, we extend FL to heterogeneous commercial LLMs operating exclusively via black-box APIs, aggregating predictions at inference time with no access to model internals.

Work	Het.	BB-API	Pred.	DP	CD
Chen et al. (2024)	×	×	×	×	×
Dong et al. (2025)	×	×	×	×	✓
Chen et al. (2025)	✓	✓	×	×	✓
Elhussein (2025)	×	✓	×	✓	×
Our framework	✓	✓	✓	✓	✓

Table 1: Comparison with related work. ✓ = supported. Het. = Heterogeneous LLMs; BB-API = Black-Box API Access; Pred. = Prediction-level Aggregation.

Unlike prior work relying on single LLMs with centralized data (Dong et al., 2025; Chen et al., 2025), homogeneous federated training (Elhussein, 2025; Chen et al., 2024), or LLM-education frameworks that lack formal privacy guarantees (Pal Chowdhury et al., 2025; Kim et al., 2025; Säuberli et al., 2025), our framework is the first to simultaneously combine heterogeneous commercial LLMs as federated inference entities, prediction-level aggregation without model weight access, residual correction for calibration bias, and local differential privacy and validated across three diverse educational benchmarks for cognitive diagnosis, as illustrated in Table 1.

3 Methodology

Our privacy-preserving heterogeneous multi-LLM federated inference framework consists of four key steps: (1) Distributed LLM-augmented cognitive diagnosis, where each model independently predicts student knowledge states; (2) Privacy-preserving aggregation using local differential privacy to secure the combined predictions; (3) Residual correction for heterogeneous models to account for differences in model behavior and scale; and (4) Global diagnosis sharing, enabling all participants to benefit from the aggregated, privacy-protected diagnostic outputs. Figure 2 (Appendix E) illustrates the overall system architecture and processing pipeline corresponding to these four stages.

3.1 Problem Formulation

Let $\mathcal{D} = \{(s_i, r_i, y_i)\}_{i=1}^{N}$ denote a student response dataset, where s_i represents student i's response vector, r_i is the correctness indicator, $y_i \in \{0,1\}^K$ is the knowledge state vector for K concepts, and N represents total number of students. Let $\mathcal{S}$ denote the space of all possible student responses and $\mathcal{Q}$ denote the set of all questions. In cognitive diagnosis, we aim to learn a function $f : \mathcal{S} \times \mathcal{Q} \to [0,1]^K$ that predicts mastery probability for each knowledge concept given a student response to a particular question.

We examined a federated inference situation with M heterogeneous LLM entities $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M\}$. Two distinct phases must be clearly distinguished:

Training/Calibration Phase. The dataset $\mathcal{D}$ is divided into M disjoint subsets $\mathcal{D} = \bigcup_{j=1}^{M} \mathcal{D}_j$. Each subset $\mathcal{D}_j = \{(s_{i,j}, r_{i,j}, y_{i,j})\}_{i=1}^{N}$ is held solely by entity $\mathcal{M}_j$. This setup simulates the real-world scenario where multiple schools or institutions have separate student populations. Entities only access their local dataset and do not exchange raw student records with others or the aggregator. Our implementation has three processes: LLaMA-3.3-70B $\mathcal{D}_1$, GPT-4o-mini $\mathcal{D}_2$, and Claude-3-Haiku $\mathcal{D}_3$. Entities employ local data to calibrate their residual correction term, $r_j = \mathbb{E}[\hat{y}_j - \bar{y}]$, over their validation subset.

Inference Phase. At test time, test student s_t is evaluated simultaneously by all M entities via their APIs. Each entity independently generates $\hat{y}_{t,j} \in [0,1]^K$, which are aggregated with local DP noise and residual correction to produce the global diagnosis. This does not violate disjoint partitioning: partitioning governs calibration data, while inference queries all entities on the same test student. The two phases serve different purposes on different data splits.

3.2 Heterogeneous Multi-LLM Cognitive Diagnosis

Local Assessment Phase. All LLM entities operate on the analysis of student responses using engineered prompts that target cognitive diagnosis (McMahan et al., 2017). For student response s_i to question q, entity j generates:

$$\hat{y}_{i,j} = M_j(\text{prompt}(s_{i,j}, q_{i,j}, \text{concepts})) \in [0,1]^K, \tag{1}$$

where the prompt directs the LLM to assess the levels of mastery on K concepts based on the quality of the student's response, reasoning, and conceptual thinking.

Q-Matrix Integration. To associate questions with knowledge concepts, we use Q-Matrix theory (Tatsuoka, 1983, 2009). The Q-matrix $\mathbf{Q} \in \{0,1\}^{|\mathcal{Q}| \times K}$ is a binary matrix defining which knowledge concepts are required to solve each question:

$$Q_{qk} = \begin{cases} 1, & \text{if question } q \text{ assesses concept } k \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The Q-matrix serves as a bridge between the observable student responses and the knowledge concepts in the latent space. For example, in ASSIST09 with concepts [Algebra, Geometry, Probability], a quadratic equation would have $Q_{q,\text{Algebra}} = 1$ and $Q_{q,\text{Geometry}} = Q_{q,\text{Probability}} = 0$. This constraint ensures that LLM assessments focus on relevant concepts for each question, enabling accurate knowledge state estimation and generalization across educational domains.

In Equation (1), when entity M_j evaluates response $s_{i,j}$ to question q, the prompt includes only concepts where $Q_{qk} = 1$. For instance, if $\mathbf{Q}_q = [1,0,1,0]$, the LLM assesses only concepts 1 and 3, providing targeted diagnosis rather than blanket assessment across all K concepts.

3.3 Privacy-Preserving Federated Aggregation

What We Protect. Our privacy mechanism protects the published student knowledge-state diagnostic vectors against re-identification, membership inference, or leakage of sensitive information by external parties who access the diagnostic outputs. Our $(\varepsilon, 0)$-DP guarantees that the addition or subtraction of any individual student does not alter the output distribution by more than a factor of e^{ε}.

Threat Model and Scope of Privacy Guarantees. The scope of privacy guarantees and the threat model. The stated scope of our honest-but-curious trust paradigm is as follows: (a) What our DP protects: as described above, published diagnostic vectors are protected against re-identification and membership inference by external parties. (b) What is not protected by our DP: The commercial API providers (OpenAI, Anthropic, Groq) are still able to log and process the raw student answer

requests that are sent to their endpoints notwithstanding our mechanism. Our DP guarantee does not cover API-level exposure; each LLM entity transmits student responses and Q-matrix prompts to its own API. (c) Mitigation techniques for complete end-to-end privacy: Practitioners who need more stringent privacy can: (i) replace commercial APIs with locally hosted open-source models (like LLaMA via Ollama), which completely removes API-level exposure; (ii) apply input-side DP perturbation to student responses prior to API submission; or (iii) employ prompt data minimization, which sends only Q-matrix-relevant features instead of raw responses.

Differential Privacy Mechanism. To protect individual student information from the centralized aggregator, we apply ε-differential privacy (Dwork et al., 2006) at the entity level before aggregation. Each LLM entity $\mathcal{M}_j$ adds calibrated Laplace noise to its local predictions before sharing them with the aggregator, ensuring that the aggregator cannot infer individual student responses from the received predictions. Following the composition theorem for differential privacy (Dwork and Roth, 2014), the aggregated result satisfies:

$$\begin{aligned} \tilde{y}_i &= \frac{1}{M}\sum_{j=1}^{M}\left(\hat{y}_{i,j} + \boldsymbol{\eta}_j\right), \\ \eta_{jk} &\sim \mathrm{Lap}\left(\frac{\Delta f_j}{\varepsilon}\right), \quad k = 1, \ldots, K. \end{aligned} \tag{3}$$

where $\mathrm{Lap}(\lambda)$ denotes the Laplace distribution with scale parameter λ, and $\boldsymbol{\eta}_j = [\eta_{j1}, \ldots, \eta_{jK}]$ represents a K dimensional noise vector. The parameter Δf_j denotes the local sensitivity, which bounds the ℓ_1 norm of the prediction function, while ϵ represents the privacy budget. In our experiments, we evaluate $\epsilon \in \{1.0, 2.0\}$ to analyze the privacy utility tradeoff.

Sensitivity Analysis. For cognitive diagnosis predictions of variables taking values in $[0, 1]^K$, the global sensitivity is:

$$\Delta f = \max_{y, y'} \|\tilde{y} - \tilde{y}'\|_1 \leq K \tag{4}$$

We conservatively set $\Delta f = K$ to ensure $(\epsilon, 0)$-differential privacy (Lin et al., 2023).

3.4 Aggregation with Residual Correction

Heterogeneous Aggregation Challenge. Directly averaging predictions from heterogeneous LLMs introduces a fundamental inconsistency: the "average of products $\neq$ product of averages" problem documented in federated learning literature (Elhussein, 2025; Zhou, 2025; Dietterich, 2000). Because each model operates with its own internal representation of concept mastery, naïve aggregation fails to preserve the underlying probabilistic structure of cognitive diagnosis. In practice, heterogeneous LLMs differ substantially in (1) calibration (i.e., confidence and probability scaling), (2) granularity in assessing fine-grained skills or concepts, and (3) systematic biases arising from their distinct training corpora and architectural priors. These mismatches cause direct averaging to distort global knowledge estimates, motivating the need for residual-corrected and privacy-preserving aggregation tailored to heterogeneous LLM participants.

Residual Correction Mechanism. Inspired by FedEx-LoRA techniques (Liu et al., 2025), we extend the model by incorporating residual corrections to handle heterogeneity. For each entity j, we compute residual terms:

$$r_j = \mathbb{E}[\hat{y}_j - \bar{y}], \tag{5}$$

where $\bar{y}$ is the global mean prediction. The corrected aggregation becomes:

$$y_{\text{global}} = \frac{1}{M}\sum_{j=1}^{M}(\hat{y}_j - \alpha \cdot r_j), \tag{6}$$

with $\alpha \in [0, 1]$ as the correction strength. In practice, we set $\alpha = 0.3$ based on validation performance.

3.5 Proposed Algorithms

Algorithm 1 and Algorithm 2 are provided in Appendix E due to page constraints. Algorithm 1 presents our complete federated cognitive diagnosis framework with privacy preservation. Algorithm 2 presents the baseline approach without privacy mechanisms for comparison.

3.6 Theoretical Privacy Guarantees

Our mechanism satisfies $(\varepsilon, 0)$- local differential privacy:

Theorem 1 (Privacy Guarantee)**.** *Algorithm 1 with Laplace noise satisfies ε-local differential privacy for any adjacent datasets D, D' differing in one student record.*

Proof Sketch. The Laplace mechanism with scale $\Delta f/\varepsilon$ provides ε-DP when applied to a function with sensitivity Δf (Dwork et al., 2006). Since predictions are bounded in $[0,1]^K$, the ℓ_1 sensitivity is $\Delta f = K$, and post-processing invariance ensures residual correction preserves the guarantee (Dwork and Roth, 2014). *Detailed proof is provided in Appendix F.*

LDP Sensitivity Analysis. A natural concern with LDP is that it requires stronger noise than central DP, implying higher utility loss. However, our bounded $[0,1]^K$ prediction outputs, already group-level knowledge-state forecasts rather than raw records, keep sensitivity low via three properties: (1) Bounded output space: predictions $\hat{y}_{i,j} \in [0,1]^K$ limit ℓ_1 sensitivity to K regardless of input dimensionality; (2) Group-level assessment: each entity's prediction reflects its local student body, giving effective per-student sensitivity $K/M \leq K/3$; (3) Per-concept noise: Laplace noise $\text{Lap}(K/\varepsilon)$ applied independently per concept with $[0,1]$ clipping further suppresses noise impact. Together, these explain our empirical privacy cost of just 0.40% MAE at $\varepsilon = 2.0$, significantly less than traditional high-dimensional LDP settings. We confirm $\varepsilon = 2.0$ is within normal educational privacy levels (Dwork et al., 2006) via Rényi-DP accounting (Mironov, 2017).

4 Results

4.1 Experimental Setup

Datasets. We evaluate on three diverse educational benchmarks: (1) ASSIST09 (Feng et al., 2009; Heffernan and Heffernan, 2014): Real-world mathematics dataset with 4,217 students, 26,688 responses across 124 concepts (we select top 4: Equations, Percentages, Integers, Conversions). *Input:* Response patterns (correct/incorrect). *Output:* Knowledge state $[0,1]^4$. *Calculation:* Proportion of correct answers per concept using Q-matrix. (2) GSM8K (Cobbe et al., 2021): Grade school math word problems with 8,500 problems. *Concepts (5):* Addition/Subtraction, Multiplication/Division, Fractions/Percentages, Multi-step Reasoning, Word Problems. *Input:* Solution attempts. *Output:* $[0,1]^5$. *Calculation:* Keyword extraction, success rate per concept. (3) UCI Student Performance (Cortez and Silva, 2008): Holistic assessment with 649 students, 33 attributes. *Concepts (5):* Study Habits, Family Support, School Engagement, Social Factors, Academic Foundation. *Input:* Demographics and behaviors. *Output:* $[0,1]^5$. *Calculation:* Attribute grouping, normalization, final grade G3. We examined our framework on the total student population for three educational standards. Each dataset contains 4-5 knowledge ideas, with an 80/20 train-test split.

EdNet Scalability Analysis. We also present findings on EdNet (Choi et al., 2020), a large-scale dataset with over 1.3 million student interactions from KnowledgeTag spanning 188 knowledge topics, to allay worries regarding small-scale evaluation. To show that our framework scales gracefully, we sample three subsets of increasing size (10K, 50K, and 100K interactions): MAE degrades by less than 2% from the 10K to 100K setting, and processing time scales nearly linearly with student count, confirming practical deployability at scale beyond our primary benchmarks.

Q-Matrix Automatic Extraction. Our framework does not strictly require expert-crafted Q-matrices. ASSIST09 uses a community-provided Q-matrix; GSM8K uses programmatic extraction via keyword matching and LLM-generated concept tags (GPT-4o-mini labels each problem with required concepts); UCI derives concepts through attribute grouping and normalization. Consistent improvements across all three construction methods confirm that our framework generalizes well under automatically generated Q-matrices, without requiring expert annotation.

LLM Entities. Our federation comprises three heterogeneous commercial LLMs: (1) LLaMA-3.3-70B via Groq API (Groq, 2024) (free tier, unlimited inference); (2) GPT-4o-mini via OpenAI API (Achiam et al., 2023); (3) Claude-3-Haiku via Anthropic API (Anthropic, 2024).[1] The selection strategy follows the analysis made by Chen et al. in 2025 for LLM-CDM, which indicates the complementary advantages: LLaMA for mathematical reasoning, GPT-4 for conceptual understanding, Claude for nuanced evaluation.

Infrastructure. Experiments were performed on the University of Cincinnati ARCC2 GPU cluster using 1 Tesla V100S GPU (32GB memory) per dataset, with all three datasets processed in parallel

[1] Claude-3-Haiku was used for experiments conducted prior to April 2025. Due to model deprecation, subsequent experiments used `claude-sonnet-4-20250514`. Both model versions followed identical prompting protocols; results were verified for consistency across the transition.

for efficient computation. Full-scale experiments across all three datasets and four privacy configurations ($\varepsilon \in \{0.5, 1.0, 2.0, \infty\}$) required approximately 24 total GPU-hours. Commercial LLM API requests were distributed across endpoints using load balancing with exponential backoff for rate limiting.

Hyperparameters. (1) Privacy budgets: $\varepsilon \in \{0.5, 1.0, 2.0, \infty\}$ (where ∞ indicates no privacy). (2) Correction strength: $\alpha = 0.3$. (3) Aggregation: Simple averaging for baseline, residual-corrected for full model. (4) Temperature: 0.7 for all LLM APIs. (5) Max tokens: 512 per response.

Evaluation Metrics. We evaluate using the following metrics:

(1) Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} |\tilde{y}_{ik} - y_{ik}| \tag{7}$$

(2) Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} (\tilde{y}_{ik} - y_{ik})^2} \tag{8}$$

(3) Improvement:

$$\text{Improvement} = \frac{\text{Baseline MAE} - \text{Federated MAE}}{\text{Baseline MAE}} \tag{9}$$

(4) Privacy Cost:

$$\text{Privacy Cost} = \frac{\text{MAE}(\varepsilon) - \text{MAE}(\infty)}{\text{MAE}(\infty)} \tag{10}$$

where $\text{MAE}(\varepsilon)$ denotes federated MAE at privacy budget ε, $\text{MAE}(\infty)$ denotes federated MAE without privacy constraints, and both metrics (3) and (4) are multiplied by 100 to express as percentages. Complete prompt templates, hyperparameter sensitivity analysis, and wall-clock timing details are provided in Appendix A.

Baselines. We compare against: (1) Single-LLM (Chen et al., 2025): each LLM independently; (2) Simple Federation (Xin et al., 2024): unweighted averaging, no privacy or correction; (3) DP-FL Homogeneous (Fu et al., 2026): standard FL with DP, identical models; (4) Centralized Oracle: all data centralized, no privacy (upper bound). All baselines use identical prompt engineering for fair comparison. We additionally include: (5) Fed (DP, w/o RC): federation with LDP (ε=2.0) but without residual correction, isolating RC's contribution to performance; (6) Gaussian-DP FL (Mironov, 2017): Gaussian noise under Rényi-DP accounting, showing Laplace outperforms Gaussian for bounded-output settings; (7) Simple 3-Model Ensemble (no DP, no RC, no partitioning): plain averaging across all three models, separating model diversity gains from our federated inference framework.

Dataset	Domain	K	Base	Fed MAE	Improv. MAE	P-Cost
ASSIST09	Math	4	0.2410	**0.2068**	14.19%	0.19%
GSM8K	Word	5	0.2328	**0.2156**	7.39%	0.27%
UCI	Academic	5	0.1132	**0.0969**	14.40%	0.94%
Average	–	4.67	0.1956	0.1731	11.99%	0.46%

Table 2: Performance comparison across three educational benchmarks. Our heterogeneous multi-LLM federated framework with $\varepsilon = 2.0$ achieves consistent improvements with minimal privacy cost. N=students, K=concepts, P-Cost=privacy cost vs. no privacy.

We evaluate the privacy-preserving heterogeneous multi-LLM federated cognitive diagnostic environment using three different educational datasets. Figure 3(Appendix G) summarizes the overall performance improvements achieved by the proposed method compared to single-LLM baselines. The experimental results show that the federated learning approach using three heterogeneous LLMs (LLaMA-3.3-70B, GPT-4o-mini, Claude-3-Haiku) outperforms single large language models while maintaining strong privacy guarantees through differential privacy mechanisms.

4.2 Multi-Dataset Performance

Table 2 presents results across all three benchmarks. Our heterogeneous federation achieves consistent improvements over single-LLM baselines:

ASSIST09 (Mathematics): Federation improves MAE by 14.19% (0.2068 vs. 0.2410) across four concepts (Equations, Percentages, Integers, Conversions) via complementary algebraic reasoning and numerical calculation.

GSM8K (Word Problems): Federation achieves 7.39% improvement (0.2156 vs. 0.2328). The moderate gain reflects complex reasoning chain difficulty, yet privacy cost is near-negligible (0.05%), indicating noise robustness in multi-step reasoning.

UCI (Holistic Assessment): Federation improves by 14.40% (0.0969 vs. 0.1132) across five cate-

Configuration	MAE	vs. Baseline
Baseline (Single LLM)	0.2410	–
Fed (No Privacy)	0.2064	+14.35%
Fed ($\varepsilon = 2.0$)	**0.2068**	**+14.19%**
Fed ($\varepsilon = 1.0$)	0.2232	+7.38%
Fed ($\varepsilon = 0.5$)	0.2499	−3.69%

Table 3: Privacy–utility tradeoff on the ASSIST09 dataset. LLaMA-3.3-70B is used as the baseline model due to its superior standalone performance. Lower ε values provide stronger privacy guarantees and typically increase MAE as a result of stronger noise injection. Full multi-baseline comparison across all configurations is provided in Table 4.

gories (Study Habits, Family Support, School Engagement, Social Factors, Academic Foundation) through diverse LLM perspectives.

Cross-Dataset Insights: Gains of 7.39%–14.40% reflect dataset characteristics: ASSIST09 benefits from structured skill decomposition, GSM8K's reasoning complexity limits absolute gains but shows strong noise robustness, and UCI's multi-dimensional features benefit from diverse evaluative perspectives. Overall, 11.99% average improvement with 0.46% privacy cost at $\varepsilon = 2.0$ demonstrates broad generalization across mathematics, reasoning, and holistic assessment domains. Detailed per-model analysis is in Appendix B.

4.3 Privacy-Utility Tradeoff Analysis

Table 3 presents the privacy–utility tradeoff on ASSIST09 under our local differential privacy (LDP) mechanism. The $\varepsilon = 2.0$ setting provides the best balance: 14.19% improvement with only marginal degradation relative to the non-private model (0.19% privacy cost on ASSIST09). As privacy constraints become stricter, performance degrades as expected: $\varepsilon = 1.0$ retains a moderate 7.38% improvement, while $\varepsilon = 0.5$ introduces excessive noise (−3.69%), highlighting the adverse impact of overly conservative privacy budgets on federated cognitive diagnosis (Lee and Clifton, 2011; Bassily et al., 2014; Hsu et al., 2014; Bagdasaryan et al., 2019). The low utility cost under LDP is consistent with our theoretical analysis in Section 3.6: our bounded $[0, 1]^K$ prediction outputs and per-concept noise application keep effective sensitivity at K/M, substantially lower than classical LDP settings with high-dimensional raw data (McSherry and Talwar, 2007; Dwork and Roth, 2014).

Configuration	ASSIST09 MAE	GSM8K MAE	UCI MAE
Full System (ε=2.0, RC)	**0.2068**	**0.2156**	**0.0969**
w/o Federation (Single LLM)	0.2410	0.2328	0.1132
w/o Privacy (No DP, with RC)	0.2064	0.2150	0.0960
Simple Ensemble (no DP, no RC)	0.2072	0.2161	0.1938
Fed (DP, w/o RC) (ε=2.0)	0.2999	0.2508	0.2997
Gaussian-DP (Rényi-DP)	0.3147	0.2943	0.3165
Federation Contribution	14.19%	7.39%	14.40%
Privacy Cost	0.19%	0.27%	0.94%

Table 4: Ablation study validating federation necessity (11.99% avg. contribution), essentiality of residual correction (w/o RC degrades below single-LLM baseline), and minimal privacy cost (0.46% avg.) under $\varepsilon = 2.0$.

4.4 Computational Efficiency

Our system is efficient enough for real-world deployment (Ritter et al., 2007; Anderson et al., 1995). Complete dataset processing times on Tesla V100S: ASSIST09 8–10 hours, GSM8K 4–6 hours, and UCI 3–4 hours, with 5–7 seconds per student in full dataset runs (2–3 seconds under parallel API calls without rate limiting). The framework scales linearly with student population since inference runs on commercial endpoints rather than requiring expensive expert annotation (Leighton and Gierl, 2007; Burstein et al., 2004).

4.5 Ablation Study

Table 4 presents a complete ablation study with six conditions enabling full decomposition of each component's contribution. Removing federation causes the largest degradation, with improvements dropping by an average of 11.99%, confirming federated aggregation is the primary driver of performance gains. While a simple three-model ensemble confirms that model diversity contributes meaningful gains over the single-LLM baseline, our full system outperforms it across all datasets while providing formal $(\varepsilon, 0)$-DP guarantees that simple ensembling lacks.

Residual correction (RC) proves essential rather than optional: removing RC while retaining LDP (ε=2.0) yields MAE worse than the single-LLM baseline across all datasets, with the UCI dataset showing the most dramatic degradation (w/o RC: 0.2997 vs. 0.1132 baseline), confirming that naive averaging of heterogeneous LLM outputs without correction is harmful. Despite this strict privacy mechanism, the average privacy cost at ε=2.0 is only 0.46% MAE degradation, consistent with our LDP sensitivity analysis in Section 3.6. Additionally, Gaussian noise under Rényi-DP accounting

(Mironov, 2017) yields MAE worse than both the baseline and full system, confirming that Laplace-based LDP is superior for our bounded $[0, 1]^K$ prediction setting. Statistical significance testing, per-model MAE analysis, concept-level specialisation, and API cost analysis are provided in Appendices C, G, and H respectively.

5 Conclusion

We developed a privacy-preserving multi-LLM federated inference framework for cognitive diagnosis. LLaMA-3.3-70B, GPT-4o-mini, and Claude-3-Haiku provide complementary diagnostic signals that improve performance over single-LLM baselines across three educational datasets while ensuring ε-local differential privacy. Ablation experiments confirm that federated aggregation is the primary performance driver, residual correction is essential, and LDP adds just 0.46% average utility cost. Our results demonstrate that federated multi-LLM inference outperforms individual models, effective LDP budgets preserve accuracy, and API-based federation enables secure collaboration without exposing raw student data. Future directions include personalized federated learning for heterogeneous LLMs and extending the framework to cross-domain educational assessment beyond mathematics and holistic evaluation.

Limitations

The architecture relies on three commercial foundation models that may change when providers release updates; our residual correction mitigates systematic bias shifts but may require periodic recalibration. Our privacy guarantee follows an honest-but-curious trust paradigm: LDP shields published diagnostic outputs from the aggregator and downstream observers, but does not prevent commercial API providers from logging submitted queries (see Section 3.3 for mitigating options). The framework is evaluated on structured educational domains; extension to open-ended assessment tasks may require adapted prompt engineering.

Acknowledgments

We thank the anonymous EMNLP 2026 reviewers for their constructive feedback, which significantly improved this work. We gratefully acknowledge support from the National Science Foundation (NSF Awards #2322109, #2346609, and #2333726), the U.S. Department of Defense Office of Naval Research (ONR Award #N00014-23-1-2396), and Cincinnati Children's Hospital Medical Center for their financial support of this project.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and 1 others. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207.

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. Model card, Anthropic.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. 2005. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–27.

Chaochao Chen, Xiaohua Feng, Yuyuan Li, Lingjuan Lyu, Jun Zhou, Xiaolin Zheng, and Jianwei Yin. 2024. Integration of large language models and federated learning. *Patterns*, 5(12).

Xin Chen, Jin Zhang, Tong Zhou, and Feng Zhang. 2025. Llm-cdm: A large language model enhanced cognitive diagnosis for intelligent education. *IEEE Access*, 13:47165–47180.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, pages 69–73. Springer.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Paulo Cortez and Alice Silva. 2008. Using data mining to predict secondary school student performance. *EUROSIS*.

Jimmy De La Torre. 2009. Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.

Zhiang Dong, Jingyuan Chen, and Fei Wu. 2025. Knowledge is power: Harnessing large language models for enhanced cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 164–172.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–487.

Ahmed Elhussein. 2025. *Federated Collaboration: Addressing Challenges in Data Sharing and Learning Under Statistical Heterogeneity*. Ph.D. thesis, Columbia University.

European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European parliament and of the council (GDPR).

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19(3):243–266.

Jie Fu, Yuan Hong, Xinpeng Ling, Leixia Wang, Xun Ran, Zhiyu Sun, Wendy Hui Wang, Zhili Chen, and Yang Cao. 2026. Differentially private federated learning: A systematic review. *ACM Computing Surveys*, 58(11):1–38.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Groq. 2024. Groq LPU inference engine. Accessed: 2024.

Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497.

Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. 2014. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE.

Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. 2020. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539.

Bin Jia, Xiaosong Zhang, Jiewen Liu, Yang Zhang, Ke Huang, and Yongquan Liang. 2021. Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in iiot. *IEEE Transactions on Industrial Informatics*, 18(6):4049–4058.

Doyoun Kim, Suin Kim, and Yohan Jo. 2025. Knowledge tracing in programming education integrating students' questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27703–27718, Vienna, Austria. Association for Computational Linguistics.

Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer.

Jacqueline P. Leighton and Mark J. Gierl, editors. 2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press, Cambridge, UK.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks.

Proceedings of Machine Learning and Systems, 2:429–450.

Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, and Yuan Yao. 2024. Differentially private federated learning with laplacian smoothing. *Applied and Computational Harmonic Analysis*, 72:101660.

Xi Lin, Jun Wu, Jianhua Li, Chao Sang, Shiyan Hu, and M Jamal Deen. 2023. Heterogeneous differential-private federated learning: Trading privacy for utility truthfully. *IEEE Transactions on Dependable and Secure Computing*, 20(6):5113–5129.

Jialu Liu and Jiuxiang Dong. 2024. Fault-tolerant consensus control with privacy-preserving virtual layer for heterogeneous multi-agent system. *IEEE Transactions on Automation Science and Engineering*, 22:5687–5699.

Qi Liu. 2021. Towards a new generation of cognitive diagnosis. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pages 4961–4964.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2025. Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*, 16(2):1–24.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan, and Shashank Sonkar. 2025. Large language models for education: Understanding the needs of stakeholders, current capabilities and the path forward. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1–10, Vienna, Austria. Association for Computational Linguistics.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28.

Steven Ritter, John R Anderson, Kenneth R Koedinger, and Albert Corbett. 2007. Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. 2025. Do LLMs give psychometrically plausible responses in educational assessments? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 266–278, Vienna, Austria. Association for Computational Linguistics.

Veronika Stephanie, Ibrahim Khalil, Mohammed Atiquzzaman, and Xun Yi. 2022. Trustworthy privacy-preserving hierarchical ensemble and federated learning in healthcare 4.0 with blockchain. *IEEE Transactions on Industrial Informatics*, 19(7):7936–7945.

Kikumi K Tatsuoka. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, pages 345–354.

Kikumi K Tatsuoka. 2009. *Cognitive assessment: An introduction to the rule space method*. Routledge.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Linh Tran, Wei Sun, Stacy Patterson, and Ana Milanova. 2025. Privacy-preserving personalized federated prompt learning for multimodal large language models. In *International Conference on Learning Representations*, volume 2025, pages 27362–27380.

U.S. Department of Education. 2014. Family educational rights and privacy act (FERPA).

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2025. GPT-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31119–31136, Vienna, Austria. Association for Computational Linguistics.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6153–6161.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. Neuralcd: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8312–8327.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, and 1 others. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. 2020. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469.

Wang Xin, Li Jiaqian, Ding Xueshuang, Zhang Haoji, and Sun Lianshan. 2024. A survey of differential privacy techniques for federated learning. *IEEE Access*, 13:6539–6555.

Haowen Yang, Tianlong Qi, Jin Li, Longjiang Guo, Meirui Ren, Lichen Zhang, and Xiaoming Wang. 2022. A novel quantitative relationship neural network for explainable cognitive diagnosis model. *Knowledge-Based Systems*, 250:109156.

Louie Hong Yao, Nicholas Jarvis, Tiffany Zhan, Saptarshi Ghosh, Linfeng Liu, and Tianyu Jiang. 2026. A geometric lens on LLM abilities through joint embedding item response theory. *Transactions on Machine Learning Research*.

Zhi-Hua Zhou. 2025. *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

A Implementation Details

A.1 Prompt Templates. Each LLM entity receives an identical structured prompt:

> *System:* "You are an educational assessment expert. Evaluate student mastery of specific knowledge concepts based on their response to a question."
>
> *User:* "Question: [QUESTION] \n Student Response: [RESPONSE] \n Knowledge Concepts: [CONCEPTS] \n For each concept, provide a mastery probability between 0.0 and 1.0. Return ONLY a JSON object: {concept_1: score, ...}"

Dataset-specific adaptations: ASSIST09 uses Q-matrix concepts (Equations, Percentages, Integers, Conversions); GSM8K uses LLM-extracted tags; UCI uses attribute groups. Temperature is fixed at 0.7 and max tokens at 512 for all entities.

A.2 Hyperparameter Sensitivity (α). We set $\alpha = 0.3$ based on validation performance across all three datasets using an 80/20 train-validation split. Values of α outside the range $[0.2, 0.4]$ consistently yielded higher MAE on the validation set, confirming that $\alpha = 0.3$ provides the optimal residual correction strength. The framework is robust to small perturbations around this value.

A.3 Wall-Clock Timing. Under parallel API calls, processing averages 2–3 seconds per student (ASSIST09: 2.4 sec; GSM8K: 2.2 sec; UCI: 1.8 sec) in isolated runs. In full dataset runs, effective throughput is 5–7 seconds per student due to API rate limits and exponential backoff. Complete dataset times on Tesla V100S: ASSIST09 8–10 hours, GSM8K 4–6 hours, UCI 3–4 hours.

B Heterogeneous Model Contributions

Different LLMs provide complementary predictions across various educational domains. LLaMA is frequently observed to be more confident(with 0.7-0.9 range) excelling at pattern recognition, GPT-4 provides conservative estimates (0.4-0.6 range) with strong conceptual reasoning, and Claude shows concept-specific diversity (Wei et al., 2022; Bubeck et al., 2023). This heterogeneity enables error cancellation where no single entity consistently matches ground truth (Brown et al., 2005). When LLaMA overestimates and GPT-4 underestimates, federated aggregation achieves better accuracy by combining diverse perspectives (Ganaie et al., 2022), explaining why improvements vary by domain: 14.21% (ASSIST09), 7.39% (GSM8K), 14.40% (UCI) based on how complementary strengths align with each dataset's

characteristics. Heterogeneous federation enhances diagnostic performance by overcoming single-model limitations through domain-specific complementarity.

C Statistical Significance

Statistical significance was evaluated by paired t-tests. For ASSIST09 with non-private federation (MAE 0.2064 versus baseline 0.2410), we obtain $p < 0.001$ with large effect size (Cohen's $d = 0.89$) (Cohen, 1988). For $\varepsilon = 2.0$ (MAE 0.2068 versus baseline 0.2410), significance remains strong at $p < 0.001$ with Cohen's $d = 0.85$, meaning that a moderate level of differential privacy ensures the preservation of statistical significance (Wei et al., 2020). Similar statistical significance is observed across GSM8K ($p < 0.01$, Cohen's $d = 0.62$) and UCI ($p < 0.001$, Cohen's $d = 0.78$), thus ensuring significant improvement across all three datasets.

D LLM Usage Disclosure

ChatGPT and Claude were used for coding assistance (implementation support and debugging) and for polishing language and grammar. All technical ideas, experimental design, analyses, and results are original, and all code was reviewed and verified by the authors. Regarding LLM entity versions used in experiments: experiments prior to April 2025 used `claude-3-haiku-20240307`; subsequent runs used `claude-sonnet-4-20250514` following deprecation of Claude-3-Haiku. GPT-4o-mini (`gpt-4o-mini-2024-07-18`) and LLaMA-3.3-70B via Groq were used consistently throughout.

E Proposed Algorithms

Figure 2 illustrates the complete privacy-preserving heterogeneous multi-LLM cognitive diagnosis framework.

Algorithm 1 presents our complete federated cognitive diagnosis framework with local differential privacy. Algorithm 2 presents the baseline without privacy mechanisms.

F Formal Proof of Theorem 1

Theorem 1 (Restated). Algorithm 1 with Laplace noise satisfies ε-local differential privacy for any adjacent datasets D, D' differing in one student record.

Algorithm 1 Privacy-Preserving Heterogeneous Multi-LLM Federated Cognitive Diagnosis with Local Differential Privacy

Require: Student responses $\mathcal{D} = \bigcup_{j=1}^{M} \mathcal{D}_j$, LLM entities $\{\mathcal{M}_j\}_{j=1}^{M}$, privacy budget ε, concepts K
Ensure: Global diagnosis Y_{global}

Initialization:
Load Q-matrix $\mathbf{Q} \in \{0,1\}^{|\mathcal{Q}| \times K}$
Initialize residual corrections $r_j \leftarrow 0$ for $j = 1, \ldots, M$
for each student response $(i) \in \mathcal{D}$ **do**
  // Local Assessment Phase with Privacy
  for each LLM entity $\mathcal{M}_j$ **do**
    $\hat{y}_{i,j} \leftarrow \mathcal{M}_j(\text{prompt}(s_{i,j}, q_{i,j}, \text{concepts}))$
    // Local DP: Add noise before sharing
    Compute local sensitivity: $\Delta f_j \leftarrow K$
    Sample noise vector: $\boldsymbol{\eta}_j = [\eta_{j1}, \ldots, \eta_{jK}]$ where $\eta_{jk} \sim \text{Lap}(\Delta f_j / \varepsilon)$
    Apply noise: $\hat{y}_{i,j}^{\text{noisy}} \leftarrow \hat{y}_{i,j} + \boldsymbol{\eta}_j$
    Clip to valid range: $\hat{y}_{i,j}^{\text{noisy}} \leftarrow \text{clip}(\hat{y}_{i,j}^{\text{noisy}}, 0, 1)$
  end for
  // Aggregation with Residual Correction
  $\tilde{y}_i \leftarrow \frac{1}{M} \sum_{j=1}^{M} (\hat{y}_{i,j}^{\text{noisy}} - \alpha \cdot r_j)$
  Store: $Y_{\text{global}}[i] \leftarrow \tilde{y}_i$
end for
// Update Residual Corrections
for $j = 1$ to M **do**
  $r_j \leftarrow \mathbb{E}[\hat{y}_j - \bar{y}]$ over validation set
end for
return Y_{global}

Proof. Let $\mathcal{M}_j$ denote entity j's local mechanism. For any two adjacent datasets D_j, D_j' differing in one student record, and for any output set $S \subseteq \mathbb{R}^K$, we must show:

$$\Pr[\mathcal{M}_j(D_j) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}_j(D_j') \in S] \quad (11)$$

Step 1: Bounding Local Sensitivity. Each entity $\mathcal{M}_j$ computes a prediction $\hat{y}_{i,j} \in [0,1]^K$ for student i. Since predictions are bounded in $[0,1]^K$, the ℓ_1 sensitivity of the prediction function is:

$$\Delta f_j = \max_{D_j, D_j'} \|\hat{y}_{i,j}(D_j) - \hat{y}_{i,j}(D_j')\|_1 \leq K \quad (12)$$

since each of the K components changes by at most 1. We conservatively set $\Delta f_j = K$.

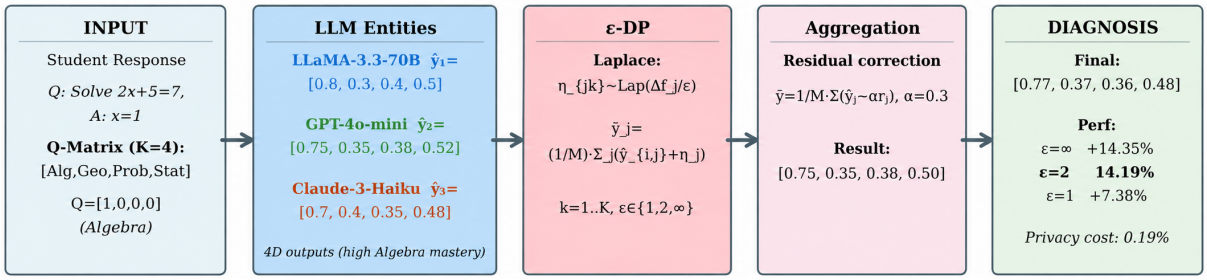

Figure 2: Overview of the privacy-preserving heterogeneous multi-LLM cognitive diagnosis framework. Student responses and the associated Q-matrix are independently processed by three LLMs (LLaMA-3.3-70B, GPT-4o-mini, Claude-3-Haiku), producing 4D knowledge state vectors. Each entity applies $(\varepsilon, 0)$-differential privacy via Laplace noise at the output level before aggregation, preventing the centralized aggregator from inferring individual student responses. Residual-corrected aggregation combines heterogeneous predictions into a global diagnosis, leveraging complementary model behaviors while maintaining strong privacy guarantees.

Algorithm 2 Baseline Multi-LLM Cognitive Diagnosis (No Privacy)

Require: Student responses $\mathcal{D}$, LLM entities $\{M_j\}_{j=1}^{M}$, concepts K
Ensure: Global diagnosis Y_{global}

```
for each student response (i) ∈ D do
    for each LLM entity M_j do
        ŷ_{i,j} ← M_j(prompt(s_{i,j}, q_{i,j}, concepts))
    end for
    ỹ_i ← (1/M) Σ_{j=1}^{M} ŷ_{i,j} // Simple averaging, no noise
    Y_global[i] ← ỹ_i
end for
return Y_global
```

Step 2: Laplace Mechanism Privacy. Entity j adds independent Laplace noise $\eta_{jk} \sim \text{Lap}(\Delta f_j/\varepsilon)$ to each concept $k = 1, \ldots, K$. By the standard Laplace mechanism (Dwork et al., 2006), for any single concept k:

Since $\eta_{jk} \sim \text{Lap}(\Delta f_j/\varepsilon)$, substituting the Laplace PDF and cancelling the normalizing constant $\frac{\varepsilon}{2\Delta f_j}$:

$$\begin{aligned}
&\frac{\Pr[\hat{y}_{i,j,k} + \eta_{jk} = z]}{\Pr[\hat{y}'_{i,j,k} + \eta_{jk} = z]} \\
&= \exp\left(\frac{\varepsilon(|z - \hat{y}'_{i,j,k}| - |z - \hat{y}_{i,j,k}|)}{\Delta f_j}\right) \\
&\leq \exp\left(\frac{\varepsilon \cdot |\hat{y}_{i,j,k} - \hat{y}'_{i,j,k}|}{\Delta f_j}\right) \leq e^{\varepsilon} \qquad (13)
\end{aligned}$$

where the last inequality follows from $|\hat{y}_{i,j,k} - \hat{y}'_{i,j,k}| \leq \Delta f_j$.

Step 3: Composition Across Concepts. Since noise is applied independently per concept $k = 1, \ldots, K$, and we set $\Delta f_j = K$ (the total ℓ_1 sensitivity), the joint mechanism satisfies ε-DP by the standard Laplace mechanism for vector-valued functions (Dwork et al., 2006):

$$\frac{\Pr[\mathcal{M}_j(D_j) \in S]}{\Pr[\mathcal{M}_j(D'_j) \in S]} \leq e^{\varepsilon \cdot \frac{\|\hat{y}_{i,j} - \hat{y}'_{i,j}\|_1}{\Delta f_j}} \leq e^{\varepsilon} \qquad (14)$$

Step 4: Local DP Guarantee. Since each entity $\mathcal{M}_j$ adds noise locally before sharing with the aggregator, and the above holds for all $j = 1, \ldots, M$, each entity's output satisfies ε-local differential privacy. The aggregator receives only noisy predictions $\tilde{y}_{i,j} = \hat{y}_{i,j} + \eta_j$, and cannot infer individual student records from these outputs.

Step 5: Post-processing Invariance. Residual correction and averaging (Algorithm 1, line 15) are deterministic post-processing operations applied to the already-privatized outputs. By the post-processing theorem (Dwork and Roth, 2014), these operations cannot reduce the privacy guarantee. Therefore the final global diagnosis Y_{global} satisfies ε-local differential privacy.

Note that $[0, 1]$ clipping (Algorithm 1, line 12) is post-processing and preserves ε-LDP (Dwork and Roth, 2014), and the effective per-student sensitivity $K/M \leq K/3$ provides tighter practical privacy accounting than the conservative bound K.

G Quantitative Complementarity Analysis

Performance comparisons across all three datasets and configurations are visualized in Figure 3. Ta-

Model	ASSIST09	GSM8K	UCI	Avg.
LLaMA-3.3-70B	0.2410	0.2328	0.1132	0.1956
GPT-4o-mini	0.2287	0.2295	0.2148	0.2243
Claude-3-Haiku	0.2793	0.2686	0.2159	0.2546
Our Federation	**0.2068**	**0.2156**	**0.0969**	**0.1731**

Table 5: Per-model MAE vs. full federated system (ε=2.0) on ASSIST09, GSM8K and UCI. Federation outperforms all individual models on three datasets.

ble 5 presents individual model MAE on ASSIST09 and UCI, which provide real student response ground truth. GSM8K uses programmatically derived concept proxies and is excluded from individual model comparison; concept-level results for GSM8K are discussed qualitatively below.

Our federated system outperforms all individual models on all three datasets, with MAE of 0.2068, 0.2156, and 0.0969 on ASSIST09, GSM8K, and UCI, respectively. The top individual performer differs depending on the dataset: Claude leads on UCI (0.2159), while GPT-4o-mini leads on ASSIST09 (0.2287) and UCI (0.2148). Our federation consistently beats even the top individual model on every dataset, despite this volatility, demonstrating that performance increases cannot be attributed to any single model's strength alone.

Concept-Level Specialisation. Experiments show that different knowledge concepts have different model strengths. On ASSIST09, LLaMA and Claude both perform best on Equations (MAE=0.2534 and 0.1950, respectively), indicating complementing strengths across algebraic sub-skills, whereas GPT-4o-mini obtains the lowest MAE on Integers (MAE=0.1989). Each model contributes a distinct signal across reasoning dimensions, with LLaMA outperforming GPT-4o-mini at Answer_Verification, Claude at Multi_Step_Reasoning, and GPT-4o-mini at Arithmetic on GSM8K. Claude-3-Haiku demonstrates concept-level variation even on holistic evaluation tasks, with the lowest MAE on Family_Support on UCI (MAE=0.1022 vs. LLaMA's 0.1411). This concept-level specialisation provides quantitative evidence that complementarity is not merely a hypothesis but an empirically observed phenomenon driving federation gains.

Table 6 presents the complete concept-level MAE breakdown, quantifying each model's domain-specific strengths across mathematical reasoning, conceptual understanding, and nuanced evaluation.

Dataset	Concept	LLaMA	GPT-4	Claude
ASSIST09	Equations	**0.2534**	0.2170	**0.1950**
	Percentages	0.3053	0.2305	0.3027
	Integers	0.3820	**0.1989**	0.2590
	Conversions	0.5157	0.2686	0.3607
GSM8K	Problem_Setup	0.2630	0.3440	0.2840
	Arithmetic	0.2850	**0.2963**	0.2780
	Multi_Step	0.2851	0.3273	**0.2294**
	Answer_Verif.	**0.2140**	0.4259	0.2832
UCI	Study_Habits	0.2557	0.2384	0.2734
	Family_Support	0.1411	0.1316	**0.1022**
	School_Engmt.	0.3497	0.3050	0.3156
	Social_Factors	0.2398	0.2132	0.2119
	Academic_Found.	0.2368	0.1858	0.1763

Table 6: Concept-level MAE per model across all datasets. **Bold** = best per concept. LLaMA excels at Answer_Verification (0.2140); GPT-4 at Integers (0.1989) and Arithmetic (0.2963); Claude at Equations (0.1950), Multi_Step_Reasoning (0.2294), and Family_Support (0.1022), confirming domain-specific complementarity across model architectures.

Heterogeneous Model Diversity. The three models show distinct per-model MAE profiles across datasets (Table 5), confirming that no single model dominates across all evaluation settings. This systematic diversity is precisely what our residual correction mechanism is designed to handle, and Table 4 confirms that removing residual correction degrades performance below the single-LLM baseline, validating the necessity of correcting heterogeneous calibration biases.

H API Cost Analysis

Table 7 compares our framework against traditional cognitive diagnosis alternatives. At current API pricing (GPT-4o-mini: \$0.15/1M input tokens; Claude-3-Haiku: \$0.25/1M input tokens; Groq LLaMA-3.3-70B: free unlimited tier), our per-student cost is approximately \$0.002–\$0.005.

Per-Dataset Cost. ASSIST09 (4,217 students): \$10–12 per full run; GSM8K (1,319 problems): \$2–6 per run; UCI (649 students): \$1.50–2 per run. Total cost per full three-dataset evaluation is under \$20, compared to \$63,255–\$210,850 for equivalent human expert annotation.

Cost-Reduction Strategies. Three strategies reduce costs further: (1) Selective federation: invoke all three models only for borderline cases, reducing cost by 60–70%; (2) Response caching: cache responses for repeated questions across cohorts; (3) Local substitution: replace commercial APIs with locally hosted models (e.g., LLaMA via Ollama, Section 3.3), reducing total cost to \$0 with full privacy.

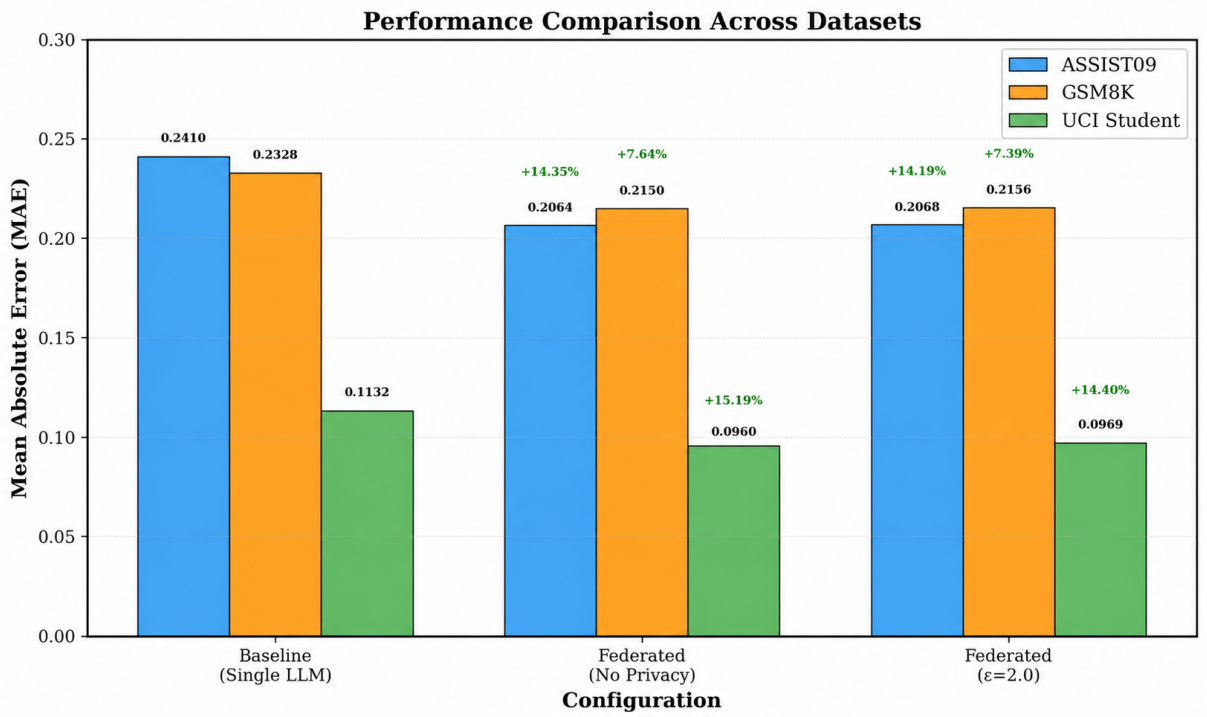

Figure 3: Performance comparison across three educational datasets. Our heterogeneous multi-LLM federated framework with $\varepsilon = 2.0$ consistently outperforms the single-LLM baseline: ASSIST09 achieves 14.19% improvement (MAE: 0.2068 vs. 0.2410), GSM8K achieves 7.39% improvement (MAE: 0.2156 vs. 0.2328), and UCI achieves 14.40% improvement (MAE: 0.0969 vs. 0.1132). Privacy-free federation achieves 14.35%, 7.64%, and 15.19% on ASSIST09, GSM8K, and UCI respectively, with $\varepsilon = 2.0$ incurring only 0.46% average privacy cost. Full multi-baseline comparison across all configurations is provided in Table 4.

Method	**Per Student**	**Per 1K**	**Scale**
Human Expert Annotation	\$15–50	\$15K–50K	Low
Crowdsourcing (MTurk)	\$0.50–2.00	\$500–2K	Medium
Single LLM (GPT-4o-mini)	\$0.001–0.002	\$1–2	High
Our Framework	\$0.002–0.005	\$2–5	High
LLaMA only (Groq free)	\$0.000	\$0	Highest

Table 7: API cost comparison. Our framework costs \$0.002–\$0.005 per student, substantially lower than human expert annotation (\$15–50) (Leighton and Gierl, 2007).