From Diagnosis to Generalization: A Cognitive Approach to Data Selection for Educational LLMs

Yuxiang Guo[1],Yan Zhuang[1],Qi Liu[1],Zhenya Huang[1],Xianquan Wang[1],Liyang He[1], Jiatong Li[1],Rui Li[1],Shijin Wang[2*]

[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]IFLYTEK Research, Hefei, China
{guoyx18,zykb,wxqcn,heliyang,cslijt,ruili2000}@mail.ustc.edu.cn,{qiliuql,huangzhy}@ustc.edu.cn
sjwang3@iflytek.com

Abstract

Specializing Large Language Models for educational domains is a key frontier in creating personalized learning tools. The central challenge is not data scarcity but its abundance: efficiently selecting a curated data subset from vast corpora to enhance specialized skills and foster generalization, without degrading existing abilities. Existing data selection paradigms, relying on superficial semantic similarity or model training dynamics, often lack a principled framework to identify data that promotes true cognitive growth. Our work proposes a paradigm shift from leveraging indirect proxies of learning value, such as semantic similarity and training dynamics, towards a framework that performs a direct, cognitive-level modeling of the learner's state. We introduce CASS, a novel framework that implements this cognitive approach through a clear pipeline, moving from an initial Diagnosis to the ultimate goal of expanding the model's cognitive frontier. First, CASS diagnoses the LLM's cognitive frontier using Multidimensional Item Response Theory. Leveraging this diagnosis, it then employs Fisher Information to select a data subset situated at LLM's cognitive frontier that offers maximum informational gain. Finally, the model is fine-tuned on this curated data using a structured, easy-to-hard curriculum to ensure effective learning. Experiments on our new multi-subject dataset show that models trained with CASS not only achieve superior accuracy in the target domain but also exhibit enhanced generalization. CASS provides a more efficient, effective, and theoretically-grounded paradigm for building expert educational LLMs.

Code — https://github.com/Guoyxustc/CASS

Data — https://huggingface.co/datasets/guoyx18/EduData

Introduction

The development of Educational Large Language Models has opened new frontiers in personalized learning (Maghsudi et al. 2021; Liu et al. 2024). Currently, a common approach is to fine-tune these models on isolated subject domains, enabling high proficiency in specific areas (Wang et al. 2024). However, to meet the evolving demands of intelligent education in a new era, there is a growing need for models that can move beyond single-domain expertise (Pratama, Sampelolo, and Lura 2023). For instance, an LLM fine-tuned specifically for mathematics may struggle with a physics problem that requires synthesizing conceptual physics with the procedural application of trigonometry. Fostering the ability to bridge these domains is essential, as it aligns with a core principle of human learning. Foundational cognitive science holds that the ability to transfer knowledge across domains is the very defining characteristic of deep understanding, distinguishing genuine comprehension from the superficial retrieval of facts (Bransford and Schwartz 1999; Bransford et al. 2000). Therefore, equipping educational LLMs with cross-scenario cognitive generalization is the key pathway to developing systems that exhibit higher-order thinking and fulfill the true promise of next-generation intelligent education.

Achieving such generalization is challenging. Naively fine-tuning on a mixture of the combined data from all domains is not only computationally prohibitive but can also degrade performance due to task interference (Wang et al. 2023). A more feasible and effective approach is to strategically select a high-impact data subset from a vast auxiliary pool. Current data selection techniques predominantly fall into two categories: Representation-based methods, which assess data relevance through semantic similarity in an embedding space (Ivison et al. 2023; Xie et al. 2023), and Training dynamics-based methods, which leverage model feedback like gradients or loss values to assess sample importance (Xia et al. 2024; Killamsetty et al. 2021; Swayamdipta et al. 2020).

However, both mainstream paradigms exhibit significant limitations when the goal is to cultivate true cognitive generalization. To illustrate, consider the goal of teaching a specialized Math LLM to solve physics problems. Representation-based methods, which rely on semantic similarity, might select physics problems laden with equations simply because their textual features resemble complex math problems. This approach often fails because such superficial similarity overlooks the need for foundational physics concepts (e.g., force, mass, acceleration), which are the true building blocks for new understanding. Similarly, training dynamics-based methods, which are computationally expensive, gradient-dependent, and often mistake training noise for learning value, might identify a problem as

*Corresponding author

hard merely due to its intricate calculations, not its core physical principles. Such noisy signals fail to reflect the intrinsic learning value of data, conflating computational difficulty with the conceptual novelty needed for cognitive growth. Consequently, existing paradigms, whether chasing surface features or noisy training signals, fail to address the core challenge: selecting data that truly expands a model's cognitive boundaries. They cannot distinguish between a physics problem that merely practices existing math skills and one that teaches the foundational knowledge required to form a new cognitive frontier.

To overcome these limitations, we argue for a paradigm shift: instead of relying on such indirect proxies, we must directly model the LLM's cognitive state. We introduce CASS (Cognitive Adaptive Selection Strategy), a novel framework that implements this vision. CASS systematically tackles two fundamental challenges: (1) How to accurately diagnose the latent cognitive state of an LLM, and (2) How to quantify a data point's informational value from a purely cognitive perspective. Our solution is a three-stage pipeline. First, CASS employs a cognitive diagnosis by combining Multidimensional Item Response Theory (MIRT) (Reckase 2006), a sophisticated model from human psychometrics that assesses both student abilities and question characteristics, with Behavioral Probing (Ribeiro et al. 2020; Li et al. 2024), a technique that ensures the model's responses reflect true understanding rather than superficial shortcuts. This provides a robust, quantifiable assessment of the LLM's cognitive frontiers. Second, guided by this diagnosis, CASS employs an adaptive selection strategy based on Fisher Information (Ly et al. 2017). This information-theoretic principle allows it to precisely select a data subset that offers maximum cognitive information gain—specifically, by prioritizing data points where the model is most uncertain, which lie at the very edge of its current capabilities. Finally, the model is fine-tuned on this curated data using a structured, easy-to-hard curriculum (Bengio et al. 2009) to ensure effective assimilation. Overall, the main contributions of our work can be summarized as follows:

- **A New Paradigm for Data Selection.** We introduce and formalize a new data selection paradigm rooted in direct cognitive diagnosis, breaking from the reliance on indirect proxies. Our approach is implemented in a gradient-free framework that, by leveraging MIRT, enables robust cognitive profiling even for black-box LLMs.
- **An Information-Theoretic Principle for Generalization.** We introduce a principled method to expand an LLM's cognitive boundaries by maximizing information gain. Concretely, we are the first to apply Fisher Information in this context to quantify the cognitive value of data, providing a theoretically-grounded alternative to heuristic-based selection.
- **A New Benchmark and State-of-the-Art Results.** To facilitate research on this important problem, we construct and release EduData, a large-scale, multi-subject dataset. On this benchmark, our CASS framework establishes a new state-of-the-art, demonstrating significant gains over strong baselines.

Related Works

Data Selection for Targeted Skill Enhancement

In the current data-rich landscape, directly training models with massive, mixed datasets not only incurs high computational costs but may also impair the model's specialized capabilities due to inter-task interference (Wang et al. 2023). Consequently, the problem of data subset selection has garnered extensive research attention (Albalak et al. 2024), as it offers a path to efficiently enhance domain-specific abilities while preserving generalization. A key strategy in this area involves identifying the most beneficial supplementary auxiliary data from vast corpora for targeted fine-tuning.

Existing data selection techniques can be roughly categorized into two main streams. The first, representation-based methods, evaluates data by analyzing text embeddings to measure relevance or diversity (Ivison et al. 2023; Xie et al. 2023). Their core limitation is the assumption that semantic similarity is a reliable proxy for learning utility, which often fails in complex educational domains. The second, more sophisticated stream uses training dynamics, leveraging model feedback like loss or gradient variations to assess sample importance (Xia et al. 2024; Chen et al. 2023; Killamsetty et al. 2021). A key inspiration for our work is the Dataset Cartography framework (Swayamdipta et al. 2020), which also seeks to identify data that is truly valuable for generalization. By tracking training dynamics, it finds that ambiguous samples—not necessarily the hardest ones—are most crucial for OOD performance. However, this approach is limited because its reliance on cognitively-agnostic signals often conflates items that are difficult due to superficial reasons (such as noise or poor formatting) with those that are genuinely valuable for expanding the model's reasoning capabilities. In contrast, our CASS framework diverges from both paradigms. We directly evaluate the cognitive information gain of data, providing a more direct solution to finding items that enhance generalization.

A Cognitive Science Perspective on LLM Evaluation and Training

LLMs are trained on vast amounts of human language and accumulated knowledge, which inevitably leads them to internalize cognitive patterns similar to humans. This theoretical inference is increasingly supported by empirical evidence (Niu et al. 2024). Research shows that LLMs not only exhibit neural representations and predictive behaviors similar to the human brain in language processing tasks (Mischler et al. 2024; Tuckute, Kanwisher, and Fedorenko 2024) but also replicate human-specific cognitive effects in higher-order tasks such as logical reasoning and even cross-modal sensory judgments (Shaki, Kraus, and Wooldridge 2023; Marjieh et al. 2024). This profound similarity, which spans the behavioral to neural levels, provides a solid theoretical foundation for researchers to understand, evaluate, and even optimize LLMs from the perspective of cognitive science.

Researchers have recently demonstrated the effectiveness of migrating cognitive science tools originally designed for human assessment into the LLM domain (Dong et al. 2025), demonstrating their effectiveness. Among these, psychome-

tric models (Liu et al. 2019; Wang et al. 2022; Zhang et al. 2024) represented by Item Response Theory (Hambleton and Swaminathan 2013) have been particularly prominent. This technique has been successfully applied to build benchmarks with higher information efficiency (Polo et al. 2024; Zhuang et al. 2025) and to significantly reduce the computational cost of complex tasks like model merging (Mencattini et al. 2025). These cutting-edge works demonstrate that IRT is not only an effective tool for evaluating LLMs but also a reliable means for passively evaluating and precisely diagnosing their capability boundaries. However, we believe their true potential lies not in passive evaluation, but in actively shaping the learning process. Our work takes this crucial next step, establishing a new paradigm for cognitively-guided LLM training. We re-envision Item Response Theory, transforming it from a post-training analysis tool into a proactive mechanism for training data selection. We leverage its diagnostic power to identify and select data situated at the model's cognitive frontier—data points that offer the greatest learning value—providing a more theoretically grounded solution to the core challenge of cross-domain generalization.

Proposed Method

A central challenge in specializing Educational LLMs is selecting a curated data subset that efficiently expands their cognitive abilities. We address this by proposing CASS (Cognitive Adaptive Selection Strategy), a framework that reframes data selection through the cognitive science perspective. Our proposed CASS framework follows a three-stage pipeline encompassing cognitive diagnosis, data selection, and curriculum-based fine-tuning, as visualized in Figure 1.

Problem Definition

Let $\mathcal{M}_{spec}$ be a Large Language Model that is already specialized in a target knowledge domain, denoted as $\mathcal{K}_T$. We are given a large, auxiliary data pool $\mathcal{D}_{aux} = \{(x_i, y_i)\}_{i=1}^{N}$, which consists of N question-answer pairs from various knowledge domains that are distinct from $\mathcal{K}_T$. We are also given a selection budget k, representing the desired size of our training subset, where $k \ll N$.

Our primary task is to formulate an optimal data selection strategy, denoted by a function π, which selects a subset $\mathcal{D}_{select} \subset \mathcal{D}_{aux}$ of size k.

$$\mathcal{D}_{select} = \pi(\mathcal{M}_{spec}, \mathcal{D}_{aux}, k) \tag{1}$$

The final model, $\mathcal{M}_{final}$, is obtained by further fine-tuning $\mathcal{M}_{spec}$ on the selected subset $\mathcal{D}_{select}$. The objective is to find a selection strategy π^* that maximizes the performance of $\mathcal{M}_{final}$. Specifically, we aim to enhance both its proficiency within the target domain (in-domain performance) and, more critically, its ability to generalize to different knowledge domains (out-of-domain generalization).

Cognitive Diagnosis

Behavioral Probing To reliably diagnose the cognitive state of an educational LLM, we must look beyond traditional evaluation metrics. Methods that solely rely on scoring accuracy are vulnerable to the model exploiting superficial shortcuts, such as positional bias, which masks its true reasoning capabilities. To address this, we adopt a Behavioral Probing process, inspired by the principles of behavioral testing in NLP (Ribeiro et al. 2020; Li et al. 2024). This process assesses the model's robustness and depth of understanding by systematically perturbing non-core features of each test item. Our implementation of this probing involves three specific perturbations: (1) Option Content Permutation, shuffling the text of multiple-choice options; (2) Option ID-Character Shift, applying a Caesar-like shift to option identifiers (e.g., A, B $\rightarrow$ U, V); and (3) Question-Options Position Swap, presenting the options before the question stem. To estimate the cognitive properties of the items within our entire training dataset, we first conduct the Behavioral Probing process across this dataset. The resulting response scores form our input for the subsequent parameter estimation step, where each score y_{ui} is the average of the binary outcomes from four trials: one on the original item and three on its perturbed variants.

Cognitive Parameter Estimation with MIRT To accurately model the cognitive state of the educational LLM and the cognitive characteristics of items, we employ the Multidimensional Item Response Theory (MIRT) model. MIRT is a sophisticated psychometric framework that extends classical Item Response Theory (IRT) by positing that performance on a task is governed not by a single latent ability, but by a vector of multiple cognitive skills. This higher-dimensional representation gives MIRT a significantly greater modeling capacity, making it exceptionally well-suited for capturing the complex, multifaceted cognitive states of educational LLMs. In our work, we utilize the 2-Parameter Logistic variant of the MIRT model.

The 2PL-MIRT model formalizes the probability of a model u correctly answering an item i as follows:

$$P(y_{ui} = 1|\theta_u, a_i, b_i) = \frac{1}{1 + \exp(-(\langle a_i, \theta_u \rangle - b_i))} \tag{2}$$

The key cognitive parameters within this model are:

- **Cognitive Status** (θ_u): A latent vector in $\mathbb{R}^d$ that represents the cognitive profile of model u across d different latent dimensions or skills.
- **Discrimination** (a_i): A vector in $\mathbb{R}^d$ that indicates how effectively item i can differentiate between models.
- **Difficulty** (b_i): A scalar value in $\mathbb{R}$ that represents the overall difficulty of item i. A higher value indicates a more challenging item.

With the model defined, we then estimate the cognitive parameters $\mathbf{\Theta} = \{\theta_u, a_i, b_i\}$ using the response scores y_{ui} from the previous step. The estimation is achieved by optimizing the Mean Squared Error (MSE) loss on the training data $\mathcal{D}_{train}$, which seeks to find the parameters $\hat{\mathbf{\Theta}}$ that minimize the difference between the observed scores and the model's predictions. This optimization problem is formally defined as:

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \sum_{(u,i) \in \mathcal{D}_{train}} \left(y_{ui} - \hat{p}_{ui}(\mathbf{\Theta})\right)^2 \tag{3}$$

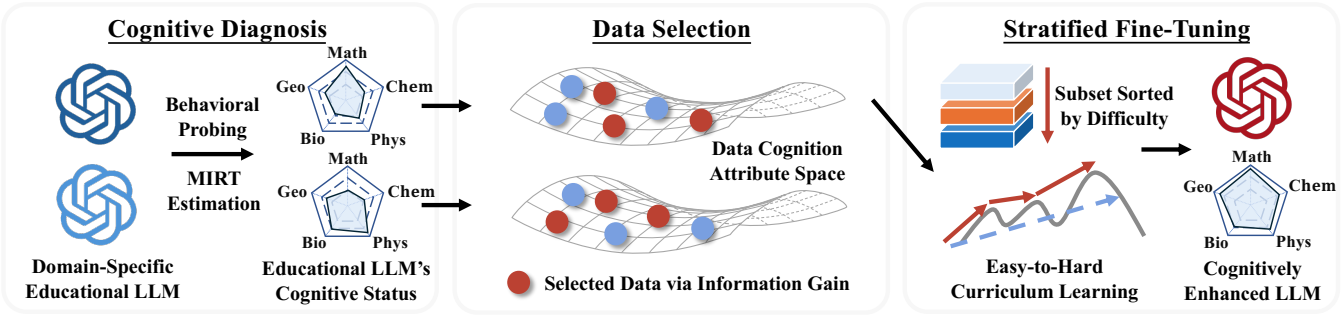


Figure 1: An overview of our proposed CASS framework. CASS first performs a cognitive diagnosis by applying Multidimensional Item Response Theory to an Educational LLM's responses from a behavioral probing task. This step simultaneously estimates both the LLM's latent cognitive status (θ_u) and the key attributes of each data item, such as its difficulty ($\hat{b}_i$) and discrimination ($\hat{a}_i$). Guided by this comprehensive diagnosis, the framework then uses Fisher Information to select a data subset that offers maximum information gain. Finally, the model is fine-tuned on this information-rich subset using an easy-to-hard curriculum, resulting in a cognitively enhanced LLM.

where $\hat{p}_{ui}(\mathbf{\Theta})$ represents the probability predicted by the 2PL-MIRT model for model u on item i using the parameter set $\mathbf{\Theta}$. These resulting parameters—the estimated model proficiency $\hat{\theta}_u$ and the calibrated item characteristics $(\hat{a}_i, \hat{b}_i)$—form the quantitative foundation for our subsequent information-theoretic data selection.

Data Selection via Fisher Information

Our next goal is to select a data subset that is maximally informative for refining the model's cognitive state, θ_u. To achieve this, we employ Fisher Information to quantify the informational value of each item.

Theoretical Justification for Fisher Information Our primary goal is to select data points that offer maximum informational gain relative to the model's current cognitive state. The Cramér-Rao Lower Bound (CRLB), a fundamental result in estimation theory, provides the basis we need. It connects the precision of any unbiased estimator, $\hat{\theta}_u$, to the Fisher Information Matrix, $\mathcal{I}(\theta_u)$, thereby allowing us to quantify informational gain:

$$\text{Cov}(\hat{\theta}_u) \geq \mathcal{I}(\theta_u)^{-1} \tag{4}$$

This inequality reveals that the Fisher Information Matrix, $\mathcal{I}(\theta_u)$, sets the theoretical limit on the best possible precision for our cognitive diagnosis—a smaller estimation variance is only possible with larger Fisher Information. Therefore, maximizing Fisher Information is the most principled way to select data, as it simultaneously maximizes the quantifiable information gain and, as a direct result, yields the most precise cognitive diagnosis.

Derivation for the 2PL-MIRT Model We derive the Fisher Information for item i with respect to the model's cognitive state θ_u, denoted as $\mathcal{I}_i(\theta_u)$. The derivation begins with the log-likelihood function, $L(\theta_u)$, for a single response:

$$L(\theta_u) = y_{ui} \log P_{ui} + (1 - y_{ui}) \log(1 - P_{ui}) \tag{5}$$

where P_{ui} is the probability of a correct response from Equation (2).

The Fisher Information is the negative expectation of the Hessian matrix $\mathbf{H}(\theta_u)$ of this log-likelihood function.

$$\mathbf{H}(\theta_u) = \frac{\partial^2 L}{\partial \theta_u \partial \theta_u^T} = -P_{ui}(1 - P_{ui}) a_i a_i^T \tag{6}$$

Since the Hessian does not contain the random variable y_{ui}, the expectation is trivial. This yields the Fisher Information for item i:

$$\mathcal{I}_i(\theta_u) = -\mathbb{E}[\mathbf{H}(\theta_u)] = P_{ui}(1 - P_{ui}) a_i a_i^T \tag{7}$$

The term $P_{ui}(1 - P_{ui})$ is the response variance, which is maximized when $P_{ui} \approx 0.5$. This confirms that the greatest information gain comes from items where the model is most uncertain—precisely at its cognitive frontier.

Application to Data Selection The result of our derivation, the Fisher Information $\mathcal{I}_i(\theta_u)$, is a $d \times d$ matrix for each item. To create a scalar value for ranking, we employ the D-optimality criterion from optimal design theory (Atkinson 2011). We define the information score for each item $i \in \mathcal{D}_{\text{aux}}$ as the determinant of its Fisher Information matrix, computed using the parameters $(\hat{\theta}_u, \hat{a}_i, \hat{b}_i)$ from our diagnosis step:

$$\text{score}(i) = \det(\mathcal{I}_i(\hat{\theta}_u)) \tag{8}$$

Maximizing this determinant is geometrically equivalent to minimizing the volume of the confidence ellipsoid for the parameter estimate $\hat{\theta}_u$, thus ensuring a precise diagnosis.

We then rank all items by this score and select the top-k items to form our curated subset, $\mathcal{D}_{\text{select}}$. This selection process is formally expressed as:

$$\mathcal{D}_{\text{select}} = \text{Top-k}_{i \in \mathcal{D}_{\text{aux}}}(\text{score}(i)) \tag{9}$$

where the Top-k operator returns the set of k items from the auxiliary pool that have the highest information scores.

Stratified Fine-Tuning with Adaptive Curriculum

Having curated an information-rich data subset, $\mathcal{D}_{select}$, simply fine-tuning in a random order would neglect the crucial cognitive insights from our diagnosis, particularly the estimated item difficulty b_i. A random mix of difficulties can lead to unstable training and hinder the development of robust knowledge. Therefore, inspired by Curriculum Learning (Bengio et al. 2009), we introduce a stratified fine-tuning strategy. Presenting data in a structured, easy-to-hard curriculum based on cognitive difficulty not only emulates effective learning but also allows for a steady and stable progression of the model's capabilities.

This adaptive curriculum is implemented in two phases. First, we group the data by partitioning $\mathcal{D}_{select}$ into K distinct groups based on the difficulty parameter b_i. For instance, with $K = 3$, we create three groups: Foundational (easiest items), Transitional (items of moderate difficulty), and Advanced (most challenging items).

Second, we perform staged fine-tuning, training the LLM sequentially through these ordered groups. The process begins with the Foundational data, continues with the Transitional data using the previously trained adapters, and concludes with the Advanced data.

This staged, curriculum-based approach is superior to naive fine-tuning because it emulates a more natural learning trajectory. By mastering simpler concepts before tackling complex ones, the model can build more stable and generalizable internal representations, ensuring the cognitively-selected data is assimilated in the most effective manner.

Experiments

In this section, we conduct experiments to address the following four research questions:

RQ1: How does CASS compare to existing task specific state-of-the-art data selection methods in simultaneously enhancing an LLM's in-domain expertise and its cross-domain generalization capabilities?

RQ2: What is the impact of each key component within CASS?

RQ3: How does the performance of CASS vary with different data selection ratios, and is its effectiveness consistent across different academic subjects?

Experimental Setup

Datasets Existing educational datasets are often monolithic, lacking the high-quality, multi-subject data necessary for developing and evaluating Large Language Models in complex, practical teaching scenarios. To bridge this gap, we introduce EduData, a new high-quality, multi-subject dataset designed for academic research. EduData consists of 98,000 challenging high-school-level problems curated from mock and college entrance examinations across seven distinct academic subjects (e.g., Mathematics, Physics, History). After a rigorous filtering process to remove items with images or incomplete solutions, each subject contains 14,000 single-choice questions. For each subject, the data is split into a training set of 10,000 and a test set of 4,000 questions, respectively.

Evaluation Protocol and Metrics The effectiveness of our data selection strategy is evaluated based on the accuracy of the fine-tuned models on held-out test datasets, which is the most direct metric of performance in the educational domain. To provide a comprehensive assessment, our evaluation is structured around two key settings designed to measure both specialized proficiency and generalization: In-domain Performance: For a given target subject (e.g., Mathematics), this is the model's accuracy on the test set of that same subject. This metric assesses the model's mastery of the specialized domain. Out-of-domain Generalization: This is the model's average accuracy across the test sets of all other subjects. This metric directly quantifies the model's ability to transfer learned skills to new, unseen domains. This dual-evaluation setup allows us to directly test the model's cross-scenario generalization capabilities.

Backbone models and baselines We evaluated a variety of subset selection methods on five representative backbone models for task-specific fine-tuning. The methods are described as follows:

LESS (Xia et al. 2024): This method utilizes a gradient-based approach to select informative samples for fine-tuning. It identifies data points that offer the highest information gain, ensuring an effective selection process.

DSIR (Xie et al. 2023): DSIR is a lightweight method that assigns importance weights to source instances based on their n-gram overlap with the target task, and then constructs the training subset via importance resampling.

DEFT-Few (Ivison et al. 2023): This technique involves using a universal encoder to generate embeddings for all datasets. A K-Nearest Neighbors (KNN) algorithm is then employed to sample the k most similar examples from a general pool to construct an auxiliary dataset tailored for each target task.

BM25 (Robertson, Zaragoza et al. 2009): We also compared our method against BM25, a classical information retrieval algorithm that uses the target task's examples as queries to rank source instances by lexical relevance and select the top-k matches.

Random Selection: As the most straightforward baseline, this method involves randomly sampling a subset of data from the general training pool for instruction fine-tuning.

Implementation Details Our experimental procedure begins by establishing a specialized baseline model for each subject via initial fine-tuning. Specifically, we fine-tune the base LLM solely on the training data of a single target subject. This crucial first step creates a controlled and fair starting point for all methods, allowing us to isolate and accurately measure the subsequent improvements in cross-domain generalization. The training sets of all other six subjects then serve as the auxiliary data pool for CASS. The CASS pipeline is then applied to this pool to select a cognitively optimal subset for enhancing generalization, proceeding as follows:

To diagnose the cognitive state of the specialized model and the attributes of the data in the training set (e.g., difficulty, discrimination), we first obtain response data from the cohort of subject-specific LLMs established in our initial

Base Model	Datasets	# Samples	Methods						
			Base	Random	BM25	DEFT	DSIR	LESS	CASS
LLaMA3-8B	In-domain	4,000	0.6476	0.6518	0.6534	0.6494	0.6572	<u>0.6651</u>	**0.6770**
	Out-of-domain	24,000	0.5286	0.5489	0.5476	0.5571	0.5307	<u>0.5580</u>	**0.5779**
	Overall	28,000	0.5457	0.5633	0.5622	0.5710	0.5478	<u>0.5724</u>	**0.5899**
Qwen2.5-1.5B	In-domain	4,000	0.7360	0.7419	0.7462	0.7430	0.7468	<u>0.7501</u>	**0.7563**
	Out-of-domain	24,000	0.7130	0.7208	<u>0.7233</u>	0.7189	0.7194	0.7216	**0.7325**
	Overall	28,000	0.7163	0.7237	<u>0.7262</u>	0.7223	0.7228	0.7257	**0.7356**
Qwen2.5-7B	In-domain	4,000	0.8788	0.8840	0.8763	0.8794	0.8817	<u>0.8850</u>	**0.8879**
	Out-of-domain	24,000	0.8651	0.8691	0.8654	0.8767	0.8692	<u>0.8733</u>	**0.8864**
	Overall	28,000	0.8671	0.8712	0.8669	0.8770	0.8710	<u>0.8750</u>	**0.8866**

Table 1: Main performance comparison on the EduData benchmark across three backbone models. The results for each backbone model represent the average performance across seven separate fine-tuning experiments, where each experiment specialized the model on one of the seven subjects. All results are based on fine-tuning the backbone models on a 10% data subset selected by each method. The top-performing method is bolded, and the runner-up is underlined.

Method	In-domain	Out-domain
Random	0.6519	0.5286
w/o Probing	0.6691	0.5531
Heuristic (Accuracy)	0.6586	0.5494
Heuristic (Difficulty)	0.6687	0.5562
w/o Curriculum	0.6743	0.5601
CASS	**0.6770**	**0.5779**

Table 2: Ablation study of the key components within the CASS framework. All experiments were conducted on the LLaMA3-8B model.

step, having them answer the questions on the training set via behavioral probing. Subsequently, we use this interaction data to train and estimate all parameters of a Multidimensional Item Response Theory model. In our experiments, the latent dimensionality of the MIRT model was set to 200. We then compute the Fisher Information for all data points in the auxiliary pool and select the top 10% that are ranked highest. Finally, the specialized model is further fine-tuned on this curated 10% subset using our staged, curriculum-based approach with LoRA to enhance its performance.

Performance Comparison (RQ1)

Table 1 presents the main performance comparison between CASS and various baseline data selection methods across different backbone models and tasks. The results lead to two primary findings:

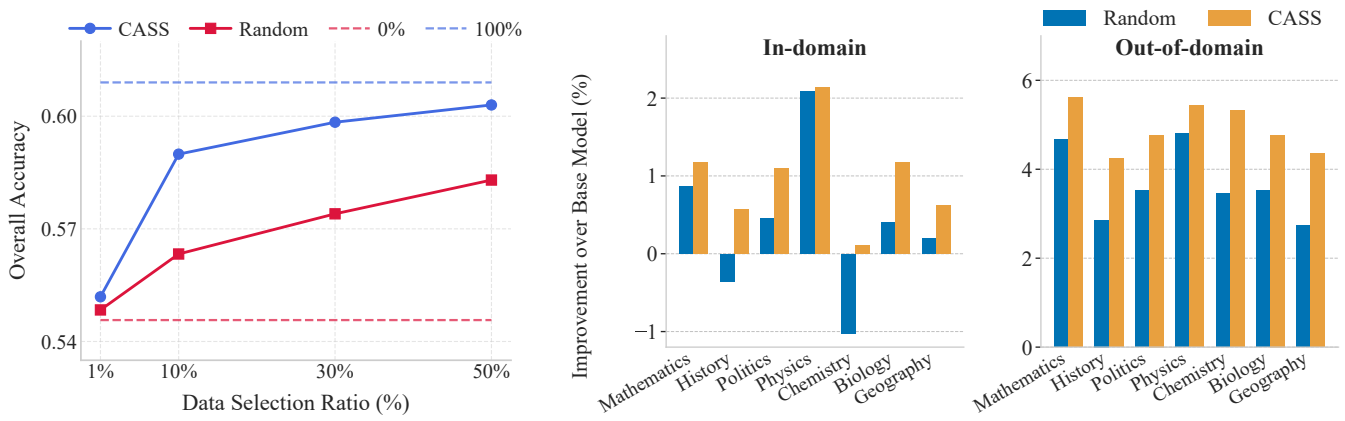
First, CASS consistently and significantly outperforms all baseline methods across all settings. As hypothesized, this advantage is particularly pronounced on the more challenging out-of-domain generalization tasks, demonstrating the effectiveness of our cognitive-driven strategy in fostering higher-order thinking skills and enabling knowledge transfer to novel problems.

Second, the performance hierarchy of the baselines clearly reveals the limitations of non-cognitive approaches when selecting data to foster generalization in educational LLMs. Representation-based methods (e.g., DEFT, BM25, DSIR) prioritize selecting semantically similar data points, which overlooks the importance of building connections across different knowledge concepts, thus proving less effective at fostering the cross-scenario generalization that defines true learning. This limitation persists even in the strongest baseline, LESS. Although its use of training dynamics is a more powerful paradigm, it attempts to quantify information gain through gradient alignment, which is a fundamentally different and less direct measure than assessing a data point's true cognitive contribution. CASS's consistent advantage stems from directly addressing this core issue. Instead of relying on such cognitively-agnostic signals, our framework builds a principled cognitive model of the LLM's state to more accurately identify data that maximizes genuine cognitive gain.

Ablation Study (RQ2)

To quantify the impact of each key component within CASS, we conduct a comprehensive ablation study. We compare the full CASS framework against several variants, with each variant removing or simplifying one key component. The details of these variants are as follows:

- **CASS w/o Probing**: This variant removes the behavioral probing step and instead builds the MIRT model from the LLM's raw performance on the standard, unperturbed dataset.
- **Heuristic (Accuracy)**: This variant replaces the MIRT-based diagnosis with a heuristic based on hard-example selection. It selects the subset of questions on which the model achieved the lowest raw accuracy.
- **Heuristic (Difficulty)**: This variant uses the MIRT diagnosis but replaces the Fisher Information criterion with a simpler heuristic that selects the data points with the highest estimated difficulty scores.
- **CASS w/o Curriculum**: This variant uses the full CASS data selection process but fine-tunes on the selected sub-



(a) Analysis by selection ratio.

(b) Analysis by subject.

Figure 2: Performance analysis of CASS and Random Selection on the LLaMA3-8B base model. (a) The left panel shows overall accuracy as a function of the data selection ratio, with 0% (base model) and 100% (full data) performance shown for reference. (b) The right panel shows the performance improvement over the base model across different subjects for both in-domain and out-of-domain tasks.

set in a random order, removing the easy-to-hard curriculum strategy.

The results are illustrated in Table 2. From these findings, we can draw several key conclusions. First, the performance degradation across all variants compared to the full CASS model highlights that each component in our framework is essential and contributes positively to the final performance. Furthermore, the greatest impact on performance was observed in the Heuristic (Accuracy) variant. This is because removing the MIRT-based diagnosis—the core component of our framework—and replacing it with a simple heuristic like the model's raw accuracy means the selection process fails to identify data that can optimally expand the model's cognitive boundary. Beyond the core MIRT diagnosis, the performance drops among the other variants reveal a clear hierarchy of importance: the Fisher Information selection criterion is the most critical component, followed by the behavioral probing step, and finally the adaptive curriculum. We also observe that the performance degradation for all ablated variants is more pronounced on the out-of-domain task. This is attributed to the fact that these cognitive components are specifically designed to cultivate transferable knowledge, making them indispensable for achieving robust generalization.

Analysis of Performance Across Subjects and Selection Ratios (RQ3)

To further investigate the efficiency and robustness of our framework, we analyzed CASS's performance across different selection ratios and subjects as shown in Figure 2.

As shown in Figure 2a, the performance curve demonstrates that the 10% selection budget represents an excellent trade-off between performance and efficiency. It captures the vast majority of achievable gains while using only a fraction of the computational resources required for larger subsets. In terms of practical application, this high data efficiency enables a much faster and more iterative development cycle for creating and iterating on specialized educational models.

Second, Figure 2b reveals the necessity of a cognitively-grounded selection strategy by demonstrating that a naive approach like Random Selection can be counterproductive, sometimes resulting in negative performance improvement for in-domain tasks (e.g., in History and Chemistry), whereas CASS provides consistent, positive gains across all subjects. This difference in robustness occurs because CASS employs a principled cognitive diagnosis to identify data that is truly valuable for generalization, whereas naive approaches like random selection lack such a mechanism and risk exposing the model to harmful or irrelevant examples.

Conclusion

In this work, we first analyzed the limitations of existing data selection methods for specializing educational LLMs, arguing for a shift towards a cognitively-grounded paradigm. Subsequently, we introduced CASS, a novel framework that uses a Multidimensional Item Response Theory model to diagnose an LLM's cognitive state, and then employs Fisher Information to select data that provides maximum informational value. Comprehensive experiments demonstrated the efficacy of our model, showing that models trained with data selected by CASS significantly outperform baseline methods in target-domain accuracy and demonstrate superior generalization. Crucially, our cognitive-driven approach proves more effective at identifying data with true pedagogical value, leading to more reliable generalization. In the future, we plan to enhance the efficiency of our framework by exploring diagnostic methods that can reliably estimate cognitive parameters without requiring large-scale interaction data. We also intend to broaden its applicability by extending our cognitive selection paradigm to more complex, open-ended, and multi-modal educational tasks.

Acknowledgments

This work was supported by the Science and Technology Innovation 2030-Major Project (2022ZD0120204), the National Key Research and Development Program of China (Grant No. 2024YFC3308200), the National Natural Science Foundation of China (62525606, 62477044, 62567004), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039) and the Fundamental Research Funds for the Central Universities. Zhenya Huang gratefully acknowledges the support of the Young Elite Scientists Sponsorship Program by CAST (No. 2024QNRC001)

References

Albalak, A.; Elazar, Y.; Xie, S. M.; Longpre, S.; Lambert, N.; Wang, X.; Muennighoff, N.; Hou, B.; Pan, L.; Jeong, H.; et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Atkinson, A. C. 2011. Optimum experimental design. In *International Encyclopedia of Statistical Science*, 1037–1039. Springer.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Bransford, J. D.; Brown, A. L.; Cocking, R. R.; et al. 2000. *How people learn*, volume 11. Washington, DC: National academy press.

Bransford, J. D.; and Schwartz, D. L. 1999. Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, 24(1): 61–100.

Chen, M.; Roberts, N.; Bhatia, K.; Wang, J.; Zhang, C.; Sala, F.; and Ré, C. 2023. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36: 36000–36040.

Dong, W.; Zhao, Y.; Sun, Z.; Liu, Y.; Peng, Z.; Zheng, J.; Zhang, Z.; Zhang, Z.; Wu, J.; Wang, R.; et al. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*.

Hambleton, R. K.; and Swaminathan, H. 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

Ivison, H.; Smith, N. A.; Hajishirzi, H.; and Dasigi, P. 2023. Data-Efficient Finetuning Using Cross-Task Nearest Neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, 9036–9061.

Killamsetty, K.; Durga, S.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, 5464–5474. PMLR.

Li, J.; Hu, R.; Huang, K.; Zhuang, Y.; Liu, Q.; Zhu, M.; Shi, X.; and Lin, W. 2024. Perteval: Unveiling real knowledge capacity of llms with knowledge-invariant perturbations. *Advances in Neural Information Processing Systems*, 37: 10679–10706.

Liu, J.; Huang, Z.; Xiao, T.; Sha, J.; Wu, J.; Liu, Q.; Wang, S.; and Chen, E. 2024. SocraticLM: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37: 85693–85721.

Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; and Hu, G. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1): 100–115.

Ly, A.; Marsman, M.; Verhagen, J.; Grasman, R. P.; and Wagenmakers, E.-J. 2017. A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80: 40–55.

Maghsudi, S.; Lan, A.; Xu, J.; and van Der Schaar, M. 2021. Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine*, 38(3): 37–50.

Marjieh, R.; Sucholutsky, I.; van Rijn, P.; Jacoby, N.; and Griffiths, T. L. 2024. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1): 21445.

Mencattini, T.; Minut, A. R.; Crisostomi, D.; Santilli, A.; and Rodolà, E. 2025. MERGE^3: Efficient Evolutionary Merging on Consumer-grade GPUs. *arXiv preprint arXiv:2502.10436*.

Mischler, G.; Li, Y. A.; Bickel, S.; Mehta, A. D.; and Mesgarani, N. 2024. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 6(12): 1467–1477.

Niu, Q.; Liu, J.; Bi, Z.; Feng, P.; Peng, B.; Chen, K.; Li, M.; Yan, L. K.; Zhang, Y.; Yin, C. H.; et al. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*.

Polo, F. M.; Weber, L.; Choshen, L.; Sun, Y.; Xu, G.; and Yurochkin, M. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, 34303–34326.

Pratama, M. P.; Sampelolo, R.; and Lura, H. 2023. Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of education, language teaching and science*, 5(2): 350–357.

Reckase, M. D. 2006. 18 Multidimensional item response theory. *Handbook of statistics*, 26: 607–642.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118*.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Shaki, J.; Kraus, S.; and Wooldridge, M. 2023. Cognitive effects in large language models. In *ECAI 2023*, 2105–2112. IOS Press.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *arXiv preprint arXiv:2009.10795*.

Tuckute, G.; Kanwisher, N.; and Fedorenko, E. 2024. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(2024): 277–301.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Yin, Y.; Wang, S.; and Su, Y. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8312–8327.

Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P. S.; and Wen, Q. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36: 74764–74786.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. LESS: selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, 54104–54132.

Xie, S. M.; Santurkar, S.; Ma, T.; and Liang, P. S. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227.

Zhang, Z.; Wu, L.; Liu, Q.; Liu, J.; Huang, Z.; Yin, Y.; Zhuang, Y.; Gao, W.; and Chen, E. 2024. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences*, 67(5): 152106.

Zhuang, Y.; Liu, Q.; Pardos, Z.; Kyllonen, P. C.; Zu, J.; Huang, Z.; Wang, S.; and Chen, E. 2025. Position: AI Evaluation Should Learn from How We Test Humans. In *Forty-second International Conference on Machine Learning Position Paper Track*.