EVALUATING THE CAPABILITIES OF LARGE LANGUAGE MODELS (LLMS) IN MIMICKING HUMAN LEARNING BEHAVIORS: AN INTERDISCIPLINARY APPROACH

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Bachelor of Arts

in

Mathematical Data Science

by Wenhua (Wendy) Liang

June 2025

Examining Committee:

Soroush Vosoughi, Ph.D.

Daniel Rockmore, Ph.D.

Peter Mucha, Ph.D.

Dean of the Guarini School of Graduate and Advanced Studies

Abstract

As large language models (LLMs) are increasingly used in research and demonstrate the potential to mimic aspects of human reasoning and behavior, a key question remains unexplored: how does interdisciplinary exposure affect their learning behavior? This study investigates whether prior training on secondary subjects (e.g., mathematics, economics) influences LLM performance on tasks in a primary domain. We evaluate how different subject pairings affect performance—whether they lead to improvements or declines—and examine how the order of exposure influences these outcomes. We also compare multiple LLM architectures to assess whether the observed patterns are consistent across models or architecture-specific. Additionally, we investigate the role of Chain-of-Thought (CoT) reasoning in facilitating interdisciplinary gains. Results show that interdisciplinary exposure leads to both gains and interference depending on the pairing and order, and that CoT selectively boosts performance in structured domains like math while degrading accuracy in others. This work presents a scalable framework for studying cross-domain generalization in LLMs, with experiments centered on Computer Science assessments but designed to extend easily to other subjects.

Keywords: Large Language Models, Interdisciplinary Learning, Knowledge Transfer, Chain-of-Thought, AI in Education, Cognitive Flexibility, Cross-Domain Performance

Contents

Section 0.1

Introduction

0.1.1. Background

The debate on the value of a liberal arts education, and more broadly, interdisciplinary learning, is not new. It reflects longstanding societal tensions around the purpose of education—whether it should serve as a vehicle for immediate economic productivity or as a foundation for cultivating thoughtful, adaptable citizens (Carmichael and LaPierre, 2014; Liu et al., 2022; Xu et al., 2022). Critics argue that in an economy increasingly driven by technological advancement, automation, and domain-specific expertise, the broad-based approach of liberal arts education may fall short in providing short-term, market-aligned skills. As such, there has been a growing emphasis on STEM-oriented curricula and vocational training. However, proponents of the liberal arts maintain that the value of interdisciplinary education lies in its capacity to foster critical thinking, ethical reasoning, creative problem-solving, and the ability to synthesize diverse perspectives—competencies that are essential not only for leadership and innovation, but also for navigating complexity in a globalized and uncertain world.

In parallel, the rise of large language models (LLMs) like GPT-4 has introduced a powerful new tool for simulating various aspects of human cognition. Researchers are beginning to explore whether LLMs can approximate or even extend human abilities in areas such as social reasoning, moral judgment, collaborative problem-solving, and generalization across domains (Chan et al., 2023; Gao et al., 2023; Leng and Yuan, 2024; Li et al., 2023; Webb et al., 2023). Yet despite the growing interest in LLMs as proxies for human intelligence, a critical gap remains: we still know little about how LLMs respond to interdisciplinary exposure, or whether they exhibit

behavioral patterns analogous to human learning when encountering multiple domains in sequence. In traditional education, students often transfer knowledge from one subject to another—for better or worse—depending on the relevance, structure, and sequencing of the content. Do LLMs exhibit similar patterns of transfer learning? Do certain combinations of subjects enhance reasoning in a target domain, while others cause cognitive interference?

This study takes a critical step toward understanding how LLMs respond to interdisciplinary exposure by investigating whether they exhibit learning patterns analogous to those observed in human education. Rather than claiming to replicate human cognition, we focus on characterizing how subject pairing, order of exposure, and reasoning strategies (e.g., Chain-of-Thought) affect LLM performance on Computer Science learning assessments. By simulating domain exposure through structured prompting, we examine how and when LLMs generalize across disciplines. In doing so, this work contributes to AI interpretability and educational research, offering a scalable and ethical framework to explore the potential of LLMs as tools for modeling learning behavior in complex, cross-domain settings.

Section 0.2

The Interdisciplinary Learning Task

To systematically evaluate interdisciplinary learning in LLMs, we examine how models apply knowledge given contextual information from one subject (s_1) to another subject (s_2) when sequentially prompted. This study measures whether prior exposure to s_1 changes how the model processes and generates responses in s_2, capturing any evidence of knowledge transfer. To ensure the effects of interdisciplinary learning are clearly isolated, we adopt an evaluation framework where knowledge application is assessed exclusively on one subject before and after exposure to another. We define

different model states to describe how the LLM has been exposed to subject-specific contextual information. The Raw Model State (M_0) represents the LLM before receiving any subject-specific context. The Single-Subject Model State (M_{s_1} or M_{s_2}) refers to the model after being prompted with context from only one subject. Lastly, the Interdisciplinary Model State (M_{s_1,s_2}) describes the model after sequentially receiving contextual information from two subjects.

The model's response function, conditioned on its state, is defined as:

$$f_M : X \to Y, \tag{1}$$

where $f_M(x)$ represents the model's response to problem x when in model state M.

To measure the effects of interdisciplinary exposure, we conduct evaluations in two controlled settings. In the single-subject baseline, the model is provided contextual information only from subject s_2, and its responses to problems from s_2 are recorded. The response function in this scenario is:

$$f_{M_{s_2}}(x) \to y, \tag{2}$$

where $x \in X_{s_2}$ represents a problem from subject s_2 and y is the model's response.

In the interdisciplinary exposure setting, the model is first provided context from subject s_1, followed by subject s_2. The response function in this case is:

$$f_{M_{s_1} \to M_{s_2}}(x) \to y, \tag{3}$$

where $x \in X_{s_2}$ represents a problem from subject s_2 and $M_{s_1} \to M_{s_2}$ indicates the transition from the single-subject state M_{s_1} to the final state M_{s_2} after sequential exposure.

The change in response patterns due to interdisciplinary exposure is quantified as:

$$\Delta Y = f_{M_{s_1} \to M_{s_2}}(X_{s_2}) - f_{M_{s_2}}(X_{s_2}), \tag{4}$$

where $f_{M_{s_2}}(X_{s_2})$ represents the model's baseline response when provided context from only subject s_2. A positive ΔY indicates that prior exposure to s_1 improves the model's performance in s_2, suggesting effective knowledge transfer. A negative ΔY implies interference, where training on s_1 hinders performance in s_2, possibly due to conflicting reasoning patterns. If $\Delta Y = 0$, interdisciplinary exposure has no measurable effect, indicating that knowledge from s_1 neither enhances nor disrupts performance in s_2.

0.2.1. Additional Tasks: Order and Reasoning Strategy Effects

Beyond measuring whether interdisciplinary exposure affects performance, we further investigate two key factors that may modulate these effects: the order in which subjects are presented and the use of explicit reasoning strategies like CoT. These analyses help clarify under what conditions transfer occurs and how prompting structure shapes the model's learning behavior.

Effect of Subject Order To evaluate the effect of subject order, we define two distinct interdisciplinary model states:

$$M_{s_1 \to s_2} \quad \text{and} \quad M_{s_2 \to s_1}$$

These represent the model after receiving context from both subjects, but in different sequences. To isolate the impact of ordering, we compare the model's response functions:

$$\Delta_{\text{order}} Y = f_{M_{s_1 \to s_2}}(X_{s_2}) - f_{M_{s_2 \to s_1}}(X_{s_2}) \tag{5}$$

A nonzero $\Delta_{\text{order}} Y$ implies that the sequence of subject exposure influences the model's reasoning, revealing order-sensitive learning dynamics.

Effect of Reasoning Strategy (Chain-of-Thought) We define the Chain-of-Thought (CoT) variant of the model's response function:

$$f_M^{\text{CoT}}(x) \quad \text{and} \quad f_M^{\text{noCoT}}(x)$$

These denote the model's responses with and without CoT prompting, respectively. The effect of reasoning strategy under interdisciplinary exposure is measured as:

$$\Delta_{\text{CoT}} Y = f^{\text{CoT}}_{M_{s_1 \to s_2}}(X_{s_2}) - f^{\text{noCoT}}_{M_{s_1 \to s_2}}(X_{s_2}) \tag{6}$$

A positive $\Delta_{\text{CoT}} Y$ indicates that CoT prompting enhances transfer in subject s_2, while a negative value suggests potential reasoning inefficiencies or interference.

We also define the interaction effect between ordering and CoT prompting:

$$\Delta_{\text{order, CoT}} Y = f^{\text{CoT}}_{M_{s_1 \to s_2}}(X_{s_2}) - f^{\text{CoT}}_{M_{s_2 \to s_1}}(X_{s_2}) \tag{7}$$

This measures whether CoT reasoning amplifies or dampens the ordering effect across interdisciplinary exposure.

Section 0.3

Data

To evaluate interdisciplinary learning in LLMs, we constructed a dataset comprising training and testing materials, leveraging Advanced Placement (AP) exam questions as a standardized benchmark. AP exams provide structured assessments across mul-

tiple subjects, ensuring consistency in cognitive skill evaluation and question complexity. This framework enables a controlled study of LLMs' ability to generalize across disciplines.

0.3.1. Training Data

For training, we used AP preparation materials, such as *5 Steps to a 5*, which provide structured, self-contained content aligned with AP curricula. We extracted these materials using Optical Character Recognition (OCR) and web scraping. Text and equations were processed for clarity, with equations converted into structured text representations. Tables were extracted in a structured format to preserve relational data, ensuring accessibility for text-based processing. Images were excluded unless they had attached textual descriptions, which were incorporated to retain relevant contextual information. This preprocessing ensured that the extracted content remained interpretable and aligned with AP standards.

0.3.2. Test Data

For evaluation, we used official AP sample test questions, focusing exclusively on multiple-choice questions (MCQs). MCQs were selected because they allow for objective grading and minimize ambiguity in assessing correctness. Additionally, AP MCQs cover a range of cognitive skills, from factual recall to applied reasoning, providing a structured way to analyze how prior exposure to one subject affects performance in another. Unlike free-response questions (FRQs), which require subjective grading and can introduce response variability, MCQs ensure a controlled and reproducible assessment framework.

Subjects with fewer figures were prioritized to minimize reliance on visual information. For graph-heavy subjects where inclusion was necessary, we attached textual descriptions summarizing key information. These descriptions were iteratively refined

and tested to ensure they were comprehensible to LLMs, preserving the integrity of the original content while allowing for accurate text-based evaluation.

Section 0.4

Methods

To evaluate interdisciplinary learning in LLMs, we employed two different methods: a LangChain-based retrieval system and OpenAI's Assistant API file search. The former involved manually structuring subject materials into a vector database and retrieving relevant content before generating responses. Later, we transitioned to OpenAI's Assistant API, which allowed for direct file-based retrieval, eliminating the need for manual chunking and improving efficiency. This transition was driven by improvements in speed, cost, and retrieval transparency, ensuring a more controlled evaluation of knowledge transfer effects. For consistency, all evaluations were run with temperature set to 0 to eliminate randomness and ensure deterministic outputs.

0.4.1. LangChain-Based Retrieval Approach

The LangChain-based approach required segmenting AP subject materials into smaller chunks due to LLM context window limitations. Each segment was embedded using OpenAIEmbeddings and stored in a FAISS vector database for retrieval. When presented with a test question, the system retrieved the most relevant document segments, formatted a structured prompt, and generated a response using the LLM. This method ensured that models referenced structured training materials but introduced inefficiencies due to manual preprocessing and retrieval overhead.

Algorithm 1 LangChain-Based Retrieval and Response Generation

Require: AP training material $S = \{s_1, s_2, ..., s_n\}$

Require: AP test questions X

Require: Pre-trained LLM M

Require: Maximum token limit T

Initialize vector store V

for each subject s in S **do**

    Split text from s into chunks $C = \{c_1, c_2, ..., c_m\}$ where $|c_i| \leq T$

    Compute embeddings for each chunk using OpenAIEmbeddings

    Store embeddings in FAISS vector store V

end for

for each test question $x \in X$ **do**

    Reset model state M

    Retrieve top k most relevant chunks $C' \subset C$ from V

    Construct input prompt P as:

        " Context: {C'}

        Question: {x}

        Answer: "

    Generate response $y = M(P)$

    Compare y with ground truth answer key

end for

Compute overall accuracy based on correct responses

While LangChain provided a structured retrieval process, it required significant preprocessing. Chunking content manually added complexity, and similarity-based retrieval introduced variability in responses. Moreover, test execution was computationally expensive, often requiring extended processing times and multiple API calls,

making large-scale evaluations inefficient.

0.4.2. OpenAI Assistant API File Search

To address these inefficiencies, we transitioned to OpenAI's Assistant API, which allows for full-document retrieval without manual chunking. Instead of segmenting materials, the entire AP textbook for each subject was uploaded as a file, and queries were answered based on the Assistant API's built-in search mechanism. This method removed the need for external vector indexing, reduced API calls, and provided a more transparent and reproducible retrieval process.

Algorithm 2 OpenAI Assistant API File Search (with per-question stateless querying)

Require: AP training material $S = \{s_1, s_2, ..., s_n\}$

Require: AP test questions X

Require: Pre-trained LLM M

Upload full-text files S to OpenAI Assistant API storage
Initialize API assistant instance A with file search enabled
for each test question $x \in X$ **do**
    Reset assistant context
    Query API A with:
    "Refer to uploaded materials and answer: x"
    Receive generated response y from API
    Compare y with ground truth answer key
end for
Compute overall accuracy based on correct responses

The OpenAI Assistant API provided multiple advantages over LangChain. First, it eliminated the need for text chunking and embedding-based retrieval, allowing for

more efficient and accurate content access. Second, its built-in file search ensured consistent and deterministic retrieval, reducing variations that might arise from vector-based similarity searches. Finally, this approach significantly reduced computational overhead and execution time, making large-scale testing feasible.

0.4.3. Evaluation Setup

For both evaluation settings, model responses were assessed using official AP answer keys. Each response was classified as correct or incorrect, and accuracy was calculated as the proportion of correct answers. To quantify the impact of interdisciplinary exposure, we measured performance differences in two ways.

First, we compared the model's performance after sequential exposure to s_1 and s_2 with its baseline state:

$$\Delta Y = f_{M_{s_1} \to M_{s_2}}(X_{s_2}) - f_{M_0}(X_{s_2}) \tag{8}$$

Second, we compared performance against the single-subject training condition:

$$\Delta Y = f_{M_{s_1} \to M_{s_2}}(X_{s_2}) - f_{M_{s_2}}(X_{s_2}) \tag{9}$$

Here, $f_{M_{s_1} \to M_{s_2}}(X_{s_2})$ represents model performance on subject s_2 after prior exposure to s_1, $f_{M_{s_2}}(X_{s_2})$ reflects performance after training only on s_2, and $f_{M_0}(X_{s_2})$ denotes performance from the unexposed raw model. A positive ΔY in either case suggests knowledge transfer, while a negative value indicates potential interference from prior subject exposure.

Handling Non-Responses and Random Outputs Initially, the model occasionally failed to generate a response or produced irrelevant text instead of selecting an answer. This was especially problematic in multiple-choice evaluations, where

the expected output was a single letter corresponding to one of the answer choices. To mitigate this issue, we refined the prompting strategy by explicitly instructing the model to select one of the given choices. This adjustment significantly reduced instances of non-responses and arbitrary outputs, ensuring more reliable and standardized evaluation.

Addressing Context Loss in Prompting A key observation during testing was that batch prompting—where all test questions were asked at once—significantly degraded accuracy. We hypothesize that this was due to the model losing track of earlier context as the prompt length increased, causing information relevant to later questions to become less salient. Additionally, some responses may have been contextually influenced by prior questions or answers, leading to inconsistencies in reasoning. To address this issue, we adopted a per-question prompting strategy in which the full contextual training process was repeated before each test question rather than performing it once for all questions. This ensured that each question was answered in a consistent and independent context, eliminating biases introduced by the order of questions. Through experimental validation, we found that this approach resulted in a measurable improvement in accuracy, producing results that were both more reliable and unbiased. Based on these findings, we adopted this new approach for all subsequent evaluations.

Algorithm 3 Previous Approach: Batch Prompting

Require: Contextual training data T, test questions X

Require: Pre-trained LLM M

Provide training context T once

Query M with all test questions $X = \{x_1, x_2, ..., x_n\}$

for each response y_i **do**

    Compare y_i with the correct answer

end for

Compute overall accuracy

Algorithm 4 Current Approach: Per-Question Contextual Prompting

Require: Contextual training data T, test questions X

Require: Pre-trained LLM M

for each test question x_i **do**

    Provide full training context T before asking x_i

    Query M with x_i and retrieve response y_i

    Compare y_i with the correct answer

end for

Compute overall accuracy

The batch prompting approach in Algorithm 3 provided training context once and then presented all test questions in sequence. However, this method led to context degradation, where later questions lost relevance due to an extended prompt length or potential interference from prior questions. Additionally, responses could be influenced by previous test answers rather than being based purely on the subject content.

In contrast, Algorithm 4 introduces a per-question contextual prompting strategy.

Instead of providing context once, we refresh the model's knowledge for each test query by reintroducing the relevant training data before every question. This ensures that each response is generated from a fresh, unbiased, and contextually accurate state.

Experimental results confirmed that the per-question approach significantly improved accuracy compared to batch prompting. By reducing context loss and preventing cross-question interference, this method yielded more consistent and interpretable results, making it the preferred strategy for evaluating interdisciplinary learning effects.

Statistical Testing To assess whether performance differences were statistically significant, we applied multiple hypothesis tests tailored to the structure of each comparison. For row-level comparisons between each interdisciplinary pairing and the baseline, we used Welch's two-sample t-test on five trial scores, chosen for its robustness to unequal variances and suitability for small samples. To evaluate overall trends in performance shift, we applied one-sample t-tests and Wilcoxon signed-rank tests to the distribution of average accuracy differences across pairings. These tested whether interdisciplinary exposure led to consistent improvement or interference relative to baseline conditions. For subject order effects, we applied Welch's t-tests between reversed pairings (e.g., `Math_CS` vs. `CS_Math`), followed by aggregate tests on directional differences. For CoT prompting, we used paired t-tests and Wilcoxon tests on accuracy scores across matched pairings with and without CoT. In all cases, we report p-values to verify robustness under different distributional assumptions.

Section 0.5

Results

This section evaluates the impact of interdisciplinary learning on LLM performance. Each evaluation condition is based on 60 multiple-choice test questions drawn from publicly available AP-style subject exams. For robustness, we ran five independent trials per condition and report the average accuracy.

We begin with a cross-model comparison, analyzing how different architectures (GPT-3.5 Turbo, GPT-4o Mini, and GPT-4o) generalize across training conditions, highlighting differences between LangChain-based and OpenAI Assistant retrieval. Next, we assess subject cross-impact, focusing on Computer Science (CS) as the test subject to determine which interdisciplinary pairings enhance or hinder performance, and how training order influences results. Finally, we examine the effect of Chain-of-Thought (CoT) prompting, identifying subjects that benefit from structured reasoning and cases where CoT unexpectedly reduces accuracy. These findings offer insights into optimizing interdisciplinary training for LLMs.

0.5.1. Cross-Model Performance Comparison

To evaluate the impact of interdisciplinary learning on LLM performance, we first analyze how different architectures respond to cross-subject exposure. Since LLMs vary in their capacity to generalize knowledge, we begin by comparing the performance of different models before examining the specific effects of interdisciplinary learning.

A key distinction emerges between LangChain-based retrieval and OpenAI Assistant retrieval for GPT-3.5 Turbo across different training conditions. As shown in Table 1, OpenAI Assistant achieves a higher accuracy only in the Raw Model condition (60.0% vs. 42.5% for LangChain). However, in all other training conditions, LangChain retrieval consistently outperforms OpenAI Assistant retrieval. For

Table 1: Performance Comparison of Models Across Different Training Conditions

Trained On	**LangChain**	**OpenAI Assistant**		
	GPT-3.5 Turbo	**GPT-3.5 Turbo**	**GPT-4o Mini**	**GPT-4o**
Raw	42.5%	60.0%	67.5%	67.5%
CS	45.0%	42.5%	65.0%	67.5%
CS-Econ	45.0%	37.5%	60.0%	70.0%
Econ-CS	47.5%	35.0%	65.0%	67.5%
CS-Psych	45.0%	42.5%	65.0%	70.0%
Psych-CS	47.5%	32.5%	62.5%	70.0%
CS-Latin	50.0%	30.0%	65.0%	67.5%
Latin-CS	47.5%	35.0%	67.5%	62.5%
CS-CompLit	45.0%	40.0%	60.0%	70.0%
CompLit-CS	45.0%	30.0%	67.5%	70.0%

Table 2: Accuracy is reported as the average across 5 trials on a set of 60 multiple-choice Computer Science questions. Each model was evaluated after training on either a single subject (e.g., CS) or a subject pair (e.g., CS-Econ) to assess the effect of interdisciplinary exposure. Results are shown for both LangChain-based retrieval and OpenAI Assistant API retrieval across three model variants.

instance, in Econ-CS, LangChain achieves 47.5%, whereas OpenAI Assistant records 35.0%. Similarly, in CS-Econ and CS-Psych, LangChain reaches 45.0%, compared to 37.5% for OpenAI Assistant. Despite these higher accuracy scores, LangChain retrieval introduces greater variability and lacks transparency, primarily due to its vector-based similarity search mechanism, which may retrieve inconsistent context. In contrast, OpenAI Assistant retrieval provides a more deterministic and reproducible evaluation framework, eliminating potential retrieval-based confounders. Due to these methodological advantages, OpenAI Assistant retrieval was adopted as the standard method for all subsequent analyses.

Within OpenAI Assistant retrieval, a clear ranking emerges: GPT-4o consistently outperforms both GPT-4o Mini and GPT-3.5 Turbo across nearly all training conditions. For example, in the CS-Econ pairing, GPT-4o achieves 70.0% accuracy,

compared to 60.0% for GPT-4o Mini and 37.5% for GPT-3.5 Turbo. In Psych-CS, the gap is similarly wide—GPT-4o reaches 70.0%, far ahead of Mini (62.5%) and Turbo (32.5%). These patterns suggest that more advanced architectures generalize more effectively across subject pairings.

Notably, GPT-4o is also the only model in our evaluations to consistently surpass the Raw baseline when trained on additional interdisciplinary content. This is especially significant given that Computer Science, our test subject, is likely well-represented in pretraining corpora. Smaller models may already perform near their ceiling on CS tasks, limiting the observable benefits of added context. In contrast, GPT-4o appears more sensitive to the structure and sequencing of interdisciplinary inputs, showing both stronger gains and sharper declines depending on the pairing.

Given these observations—and the methodological advantages of OpenAI Assistant's file-based retrieval—all subsequent analyses in this paper use GPT-4o as the evaluation model.

0.5.2. Subject Cross-Impact Analysis

To evaluate the effect of interdisciplinary learning on GPT-4o's ability to generalize knowledge, we systematically tested its performance across different training conditions, focusing on CS as the test subject. Prior experiments indicated that CS exhibits substantial variation when trained alongside different disciplines, making it an ideal candidate for analyzing interdisciplinary transfer effects. A more comprehensive preliminary graphic of subject impacts is provided in the Appendix (see Figure 5).

For each training condition, we conducted five independent trials to account for potential variability in model responses. This repeated evaluation ensures that any observed patterns are not due to stochastic fluctuations in the model's outputs but instead reflect systematic differences in knowledge transfer. The trials were then averaged to obtain a reliable measure of accuracy under each training condition.

Table 3 presents the results across all tested conditions. The Raw model represents GPT-4o's base performance without any additional subject training and achieved an accuracy of 70.5%, serving as a benchmark for evaluating the impact of interdisciplinary training. Interestingly, when the model was trained exclusively on CS, accuracy declined to 67.5%. This suggests that exposure to CS alone may constrain generalization. One possible explanation is that CS-only training leads the model to overfit narrow syntactic or procedural patterns, which do not translate well to broader problem-solving contexts.

Since CS exposure already appears to hinder performance, we use the CS-only model (67.5%) as the primary benchmark for most comparisons. To assess whether interdisciplinary training meaningfully alters performance, we conducted hypothesis tests on the average accuracy differences between subject pairings and the CS-only baseline. Across 42 subject pairings, interdisciplinary training yielded an average improvement of +1.36 percentage points over CS ($p < 0.001$), with 28 pairings showing improvement, 10 showing decline, and 4 neutral. These results were robust to non-normality, as confirmed by a Wilcoxon signed-rank test ($p < 0.001$).

When compared to the Raw model, however, the same subject pairings resulted in a mean accuracy decrease of -1.64 percentage points ($p < 0.001$). This contrast highlights that while interdisciplinary input helps counter the rigidity introduced by narrow CS-only training, it does not universally outperform a more balanced initial model state. Together, these findings suggest that the effectiveness of interdisciplinary learning is sensitive to baseline conditions and subject pairings—it can mitigate overfitting, but also introduces the risk of distraction or interference.

Table 3: GPT-4o CS Performance Across Different Training Conditions

Model: GPT-4o — Test Subject: CS

Training Subjects	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Avg. Accuracy
Raw	72.5	70.0	70.0	67.5	72.5	70.5
CS	67.5	67.5	65.0	70.0	67.5	67.5
CS_Econ	67.5	67.5	65.0	65.0	67.5	66.5
Econ_CS	67.5	67.5	65.0	70.0	67.5	67.5
CS_Psych	65.0	70.0	65.0	65.0	65.0	66.0
Psych_CS	67.5	65.0	67.5	67.5	65.0	66.5
CS_Latin	67.5	70.0	67.5	65.0	70.0	68.0
Latin_CS	65.0	70.0	65.0	65.0	65.0	66.0
CS_CalculusAB	67.5	70.0	67.5	67.5	65.0	67.5
CalculusAB_CS	67.5	55.0	67.5	62.5	70.0	64.5
CS_Stats	72.5	67.5	70.0	67.5	65.0	68.5
Stats_CS	60.0	70.0	70.0	67.5	62.5	66.0
CS_CompLit	67.5	67.5	65.0	65.0	65.0	66.0
CompLit_CS	65.0	65.0	65.0	70.0	70.0	67.0
Econ_Psych	67.5	70.0	65.0	70.0	62.5	67.0
Psych_Econ	77.5	70.0	72.5	70.0	72.5	72.5
Econ_Latin	70.0	70.0	67.5	75.0	70.0	70.5
Latin_Econ	70.0	70.0	67.5	72.5	70.0	70.0
Econ_CalculusAB	67.5	70.0	72.5	72.5	70.0	70.5
CalculusAB_Econ	67.5	75.0	62.5	67.5	65.0	67.5
Econ_Stats	72.5	67.5	75.0	70.0	65.0	70.0
Stats_Econ	70.0	70.0	72.5	67.5	65.0	69.0
Econ_CompLit	67.5	70.0	72.5	72.5	67.5	70.0
CompLit_Econ	67.5	75.0	70.0	67.5	72.5	70.5
Psych_Latin	72.5	70.0	72.5	70.0	72.5	71.5

Latin_Psych	70.0	65.0	65.0	70.0	70.0	68.0
Psych_CalculusAB	70.0	70.0	67.5	70.0	67.5	69.0
CalculusAB_Psych	67.5	67.5	70.0	65.0	67.5	67.5
Psych_Stats	70.0	72.5	70.0	70.0	72.5	71.0
Stats_Psych	67.5	72.5	70.0	75.0	72.5	71.5
Psych_CompLit	70.0	72.5	67.5	70.0	70.0	70.0
CompLit_Psych	72.5	65.0	70.0	70.0	72.5	70.0
Latin_CalculusAB	70.0	67.5	65.0	72.5	70.0	69.0
CalculusAB_Latin	72.5	65.0	70.0	75.0	72.5	71.0
Latin_Stats	72.5	70.0	70.0	70.0	72.5	71.0
Stats_Latin	75.0	72.5	72.5	70.0	67.5	71.5
Latin_CompLit	67.5	70.0	65.0	67.5	65.0	67.0
CompLit_Latin	70.0	72.5	67.5	70.0	70.0	70.0
CalculusAB_Stats	65.0	70.0	72.5	65.0	72.5	69.0
Stats_CalculusAB	70.0	72.5	67.5	67.5	67.5	69.0
CalculusAB_CompLit	67.5	72.5	70.0	70.0	67.5	69.5
CompLit_CalculusAB	70.0	67.5	67.5	67.5	72.5	69.0
Stats_CompLit	67.5	70.0	72.5	70.0	70.0	70.0
CompLit_Stats	70.0	75.0	70.0	62.5	72.5	70.0

Note: GPT-4o accuracy on 60 AP Computer Science questions across different interdisciplinary training conditions. Each row reports accuracy averaged over 5 trials using the OpenAI Assistant API. Models were trained on a single subject or subject pair.

Subjects That Improve CS Performance. A number of subject combinations led to statistically significant improvements (at $p < 0.05$) over the CS baseline (67.5%), though none showed significant gains relative to the Raw model (70.5%). The strongest gain was observed in **Psych-Econ**, which improved CS performance by 5.0 percentage points. Other significantly improving combinations include **Psych-Latin**, **Stats-Psych**, and **Stats-Latin**

(each with +4.0 points), followed by **Psych-Stats** and **Latin-Stats** (each with +3.5 points), and **Econ-CalculusAB** (+3.0 points). These results suggest that exposure to conceptually rich or complementary reasoning domains—particularly those rooted in social science, formal logic, or mathematical abstraction—may enhance performance on CS tasks, especially in comparison to more narrowly focused CS-only training.

Subjects That Reduce CS Performance. Conversely, several interdisciplinary subject pairings result in statistically significant performance declines (at $p < 0.05$) relative to the Raw model baseline (70.5%), though none of these declines are significant when compared to the CS-only baseline (67.5%). The most pronounced drops occur in **CS-Psych**, **Latin-CS**, and **CS-CompLit**, each reducing performance by 4.5 percentage points. Additional significant declines include **CS-Econ** and **Psych-CS** (−4.0 points), as well as **Latin-CompLit** (−3.5 points), and **Econ-CS**, **CS-CalculusAB**, and **CalculusAB-Psych** (each −3.0 points).

Notably, seven of these nine combinations involve CS, either as the first or second subject. The only non-CS-inclusive pairings to show significant performance drops are **Latin-CompLit** and **CalculusAB-Psych**, suggesting that interference effects are especially common in CS-related configurations.

For a complete visualization of all subject pairings and their performance effects—including which were significantly beneficial or harmful—see Figure 1.

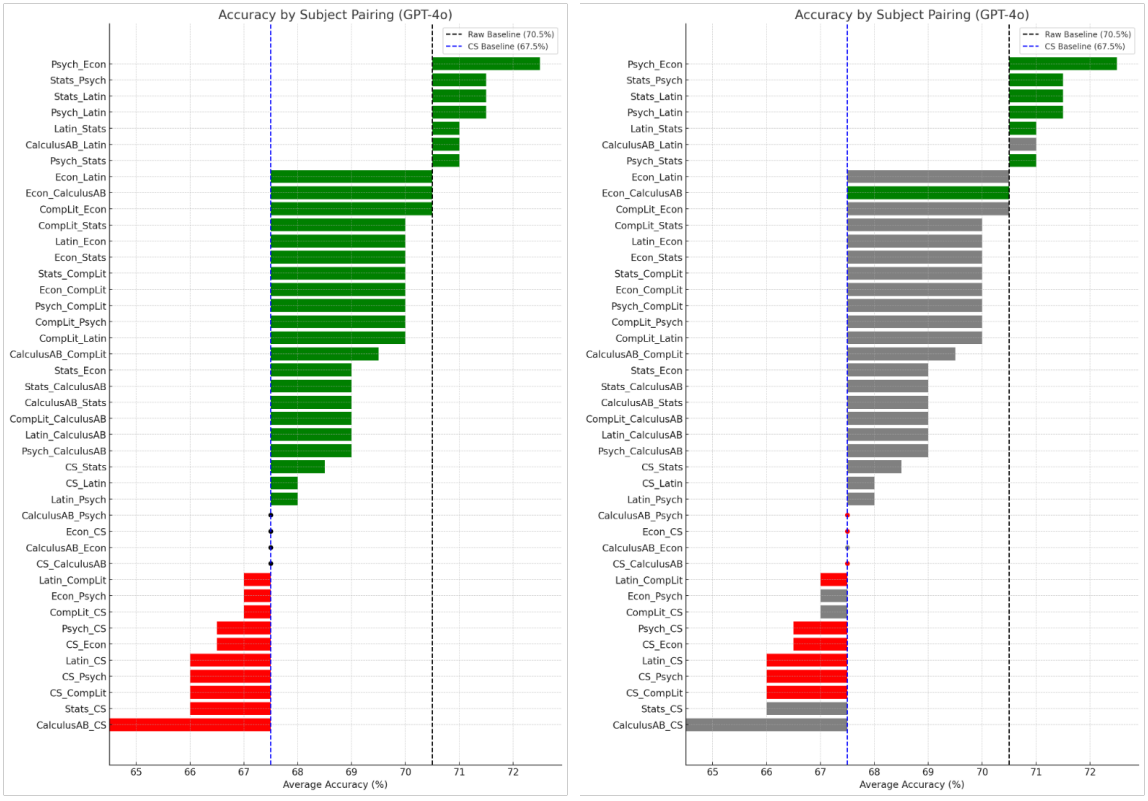


Figure 1: Impact of Interdisciplinary Training on GPT-4o's CS Task Accuracy. Both panels show average accuracy by subject pairing, sorted in descending order. **Panel (a)** presents the full ranking: green bars indicate performance increases and blue bars indicate decreases, regardless of significance. **Panel (b)** filters for statistical significance: green bars indicate improvements significant relative to the CS baseline, red bars indicate declines significant relative to the Raw baseline, and gray bars are not statistically significant. Dashed lines show the Raw and CS baselines.

Several of these combinations also reveal asymmetries depending on subject order—for example, **Latin-CS** (66.0%) significantly underperforms, while the reversed pairing **CS-Latin** (68.0%) leads to improvement. These findings point to subject order as a critical factor in determining whether interdisciplinary exposure helps or hinders performance, a dynamic we explore further in the next section.

Order of Subject Exposure Matters. To evaluate whether the sequence of exposure influences outcomes, we conducted pairwise comparisons of reversed subject orders (e.g., *CS-Econ* vs. *Econ-CS*). On average, the difference in accuracy between reversed pairs was only +0.14 percentage points, with no statistical significance detected ($p = 0.76$). This suggests that, overall, the order in which subjects are introduced does not systematically alter model performance.

Significant Order Effects. However, certain subject pairs exhibited statistically significant directional effects. For instance, **Psych-Econ** significantly outperformed **Econ-Psych** by 5.5 percentage points ($p = 0.025$), suggesting that introducing Psychology before Economics better facilitates downstream reasoning. Similarly, **Latin-CompLit** outperformed **CompLit-Latin** by 3.0 points ($p = 0.041$), and **Psych-Latin** outperformed **Latin-Psych** by 3.5 points ($p = 0.044$). These findings imply that knowledge transfer in LLMs may be asymmetric and that certain conceptual domains serve better as priming contexts for others. While not universal, these effects highlight the need to consider sequencing when modeling interdisciplinary learning in LLMs (Table 4).

Table 4: Impact of Subject Order on CS Performance (Significant Order Effects Highlighted)

Pair	**Accuracy (First)**	**Accuracy (Second)**	**Difference**	***p*-value**
Psych-Econ vs Econ-Psych	72.5	67.0	+5.5	0.025
CompLit-Latin vs Latin-CompLit	70.0	67.0	+3.0	0.041
Psych-Latin vs Latin-Psych	71.5	68.0	+3.5	0.044
Psych-CalculusAB vs CalculusAB-Psych	69.0	67.5	+1.5	0.174
CS-Latin vs Latin-CS	68.0	66.0	+2.0	0.182

These overall trends motivated a closer look at specific subjects where order effects were most pronounced. Psychology and Latin stood out as particularly informative case studies—each appeared in multiple pairings with varied outcomes depending on their order.

Tables 5 and 6 present all combinations where Psychology or Latin was involved, highlighting how performance differs when each subject is introduced first versus second.

Table 5: Impact of Training Order for Psychology-Related Pairs

Training Order	**Accuracy (%)**	**Reverse Order**	**Accuracy (%)**	**Difference**
Psych-CS	66.5	CS-Psych	66.0	+0.5 (Psych first better)
Psych-Econ*	**72.5**	**Econ-Psych**	**67.0**	**+5.5 (Psych first better)**
Psych-Stats	71.0	Stats-Psych	71.5	–0.5 (roughly the same)
Psych-CompLit	70.0	CompLit-Psych	70.0	0.0 (No difference)
Psych-Latin*	**71.5**	**Latin-Psych**	**68.0**	**+3.5 (Psych first better)**
Psych-CalculusAB	69.0	CalculusAB-Psych	67.5	+1.5 (Psych first better)

** Indicates statistically significant difference at $p < 0.05$*

Table 6: Impact of Training Order for Latin-Related Pairs

Latin First	**Accuracy (%)**	**Latin Second**	**Accuracy (%)**	**Difference**
Latin-CS	66.0	CS-Latin	68.0	–2.0 (Latin second better)
Latin-Econ	70.0	Econ-Latin	70.5	–0.5 (roughly the same)
Latin-CompLit*	**67.0**	**CompLit-Latin**	**70.0**	**–3.0 (Latin second better)**
Latin-Psych*	**68.0**	**Psych-Latin**	**71.5**	**–3.5 (Latin second better)**
Latin-Stats	71.0	Stats-Latin	71.5	–0.5 (roughly the same)
Latin-CalculusAB	69.0	CalculusAB-Latin	71.0	–2.0 (Latin second better)

** Indicates statistically significant difference at $p < 0.05$*

Notably, Psychology frequently yields better downstream performance when it appears first in the sequence, while Latin tends to result in higher accuracy when it is introduced second.

Stable Performance Across Similar Domains Some interdisciplinary pairs exhibit stable performance regardless of order, suggesting that their effects are bidirectional and mutually reinforcing. This is evident in the following cases:

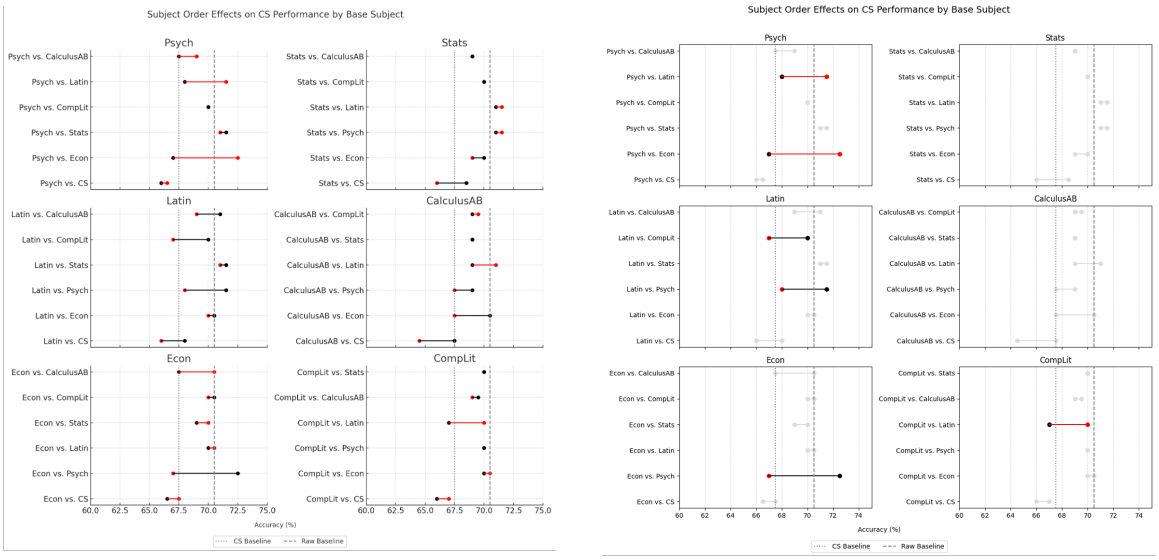
Table 7: Subject Pairs with No Performance Difference Based on Training Order

Training Order	**Accuracy (%)**	**Reverse Order**	**Accuracy (%)**	**Difference**
Stats-CompLit	70.0	CompLit-Stats	70.0	0.0 (No difference)
CalculusAB-Stats	69.0	Stats-CalculusAB	69.0	0.0 (No difference)
Psych-CompLit	70.0	CompLit-Psych	70.0	0.0 (No difference)

These results suggest that interdisciplinary training does not exert a uniform influence on CS performance. Rather, its impact depends on the specific subjects involved and the sequence in which they are introduced. Importantly, all comparisons are based on the model's accuracy on CS-related questions. As observed earlier, Psychology tends to yield more favorable outcomes when it is introduced before the paired subject, acting as an effective primer that enhances downstream reasoning in CS tasks. In contrast, Latin often leads to better CS performance when it is introduced after the paired subject, suggesting that Latin may benefit more from prior conceptual scaffolding. These subject-specific patterns underscore the role of directional influence in interdisciplinary learning and its implications for transfer effectiveness in large language models.

Some subjects, such as CalculusAB, exhibit mixed or neutral effects depending on the pairing—suggesting that not all domains function as consistent facilitators or recipients of transfer. Meanwhile, the presence of stable pairs, where training order has no impact, suggests that some interdisciplinary effects are bidirectional and mutually reinforcing. These patterns are illustrated in Figure 2, which shows both the full set of order comparisons (Panel a) and a filtered view of statistically significant effects (Panel b).

Overall, these findings point to the potential for directional knowledge transfer in LLMs—certain subjects may better prepare the model for reasoning in another domain, while others introduce interference. Although this paper does not primarily focus on knowledge transfer mechanisms, the results suggest an avenue for further exploration.



Legend: ● Header Subject First ● Header Subject Second

Figure 2: Subject Order Effects on CS Performance by Base Subject. Each plot groups subject pair comparisons by the first-listed (header) subject. Red dots indicate performance when the header subject is introduced first; black dots when it is introduced second. The connecting line is colored based on which order yields higher accuracy. Left: all comparisons. Right: only statistically significant differences.

0.5.3. Evaluating the Impact of Chain-of-Thought Prompting

To further explore reasoning structures, we examine the impact of *Chain-of-Thought* prompting on model performance. CoT prompting has been shown to enhance reasoning capabilities by encouraging step-by-step explanations rather than direct answer generation Wei et al. (2022). Prior research suggests that CoT is particularly beneficial for mathematical and symbolic reasoning Sprague et al. (2024), motivating our investigation into its effectiveness in this setting.

To analyze the effect of CoT prompting on interdisciplinary learning, we conducted two sets of evaluations. First, we ran five trials per subject pairing without CoT. Then, we implemented CoT using prompt refinement, instructing the model to provide explicit reasoning before arriving at a final answer. The results, presented in Table 9, reveal notable

trends in performance shifts across different subject pairings.

Table 8: Comparison of Accuracy With and Without Chain of Thought (CoT)

Subjects	**Without CoT (%)**	**With CoT (%)**	**Change (%)**
Raw	70.5	72.5	+2.0
CS	67.5	73.0	+5.5
CS_Econ	66.5	70.0	+3.5
Econ_CS	67.5	70.0	+2.5
CS_Psych	66.0	69.5	+3.5
Psych_CS	66.5	69.0	+2.5
CS_Latin	68.0	69.0	+1.0
Latin_CS	66.0	69.5	+3.5
CS_CalculusAB	67.5	70.5	+3.0
CalculusAB_CS	64.5	70.5	+6.0
CS_Stats	68.5	68.5	0.0
Stats_CS	66.0	69.5	+3.5
CS_CompLit	66.0	69.5	+3.5
CompLit_CS	67.0	69.5	+2.5
Econ_Psych	67.0	70.0	+3.0
Psych_Econ	72.5	69.0	-3.5
Econ_Latin	70.5	71.0	+0.5
Latin_Econ	70.0	69.0	-1.0
Econ_CalculusAB	70.5	73.5	+3.0
CalculusAB_Econ	67.5	69.5	+2.0
Econ_Stats	70.0	69.5	-0.5
Stats_Econ	69.0	69.5	+0.5

Econ_CompLit	70.0	70.0	0.0
CompLit_Econ	70.5	70.0	-0.5
Psych_Latin	71.5	68.5	-3.0
Latin_Psych	68.0	70.0	+2.0
Psych_CalculusAB	69.0	71.0	+2.0
CalculusAB_Psych	67.5	69.5	+2.0
Psych_Stats	71.0	70.0	-1.0
Stats_Psych	71.5	70.5	-1.0
Psych_CompLit	70.0	69.5	-0.5
CompLit_Psych	70.0	69.5	-0.5
Latin_CalculusAB	69.0	70.5	+1.5
CalculusAB_Latin	71.0	72.0	+1.0
Latin_Stats	71.0	68.5	-2.5
Stats_Latin	71.5	70.5	-1.0
Latin_CompLit	67.0	71.0	+4.0
CompLit_Latin	70.0	70.5	+0.5
CalculusAB_Stats	69.0	71.0	+2.0
Stats_CalculusAB	69.0	69.0	0.0
CalculusAB_CompLit	69.5	72.0	+2.5
CompLit_CalculusAB	69.0	71.5	+2.5
Stats_CompLit	70.0	68.0	-2.0
CompLit_Stats	70.0	68.0	-2.0

Note. Accuracy is reported as the average across five trials on 60 AP-style multiple-choice Computer Science questions. Each row shows the performance difference with and without Chain-of-Thought prompting. Positive changes indicate improved performance with CoT prompting.

Table 9: Comparison of Accuracy With and Without Chain of Thought (CoT)

Subjects	**Without CoT (%)**	**With CoT (%)**	**Change (%)**
Raw	70.5	72.5	+2.0
CS	67.5	73.0	+5.5
CS_Econ	66.5	70.0	+3.5
Econ_CS	67.5	70.0	+2.5
CS_Psych	66.0	69.5	+3.5
Psych_CS	66.5	69.0	+2.5
CS_Latin	68.0	69.0	+1.0
Latin_CS	66.0	69.5	+3.5
CS_CalculusAB	67.5	70.5	+3.0
CalculusAB_CS	64.5	70.5	+6.0
CS_Stats	68.5	68.5	0.0
Stats_CS	66.0	69.5	+3.5
CS_CompLit	66.0	69.5	+3.5
CompLit_CS	67.0	69.5	+2.5
Econ_Psych	67.0	70.0	+3.0
Psych_Econ	72.5	69.0	-3.5
Econ_Latin	70.5	71.0	+0.5
Latin_Econ	70.0	69.0	-1.0
Econ_CalculusAB	70.5	73.5	+3.0
CalculusAB_Econ	67.5	69.5	+2.0
Econ_Stats	70.0	69.5	-0.5
Stats_Econ	69.0	69.5	+0.5
Econ_CompLit	70.0	70.0	0.0
CompLit_Econ	70.5	70.0	-0.5

Psych_Latin	71.5	68.5	-3.0
Latin_Psych	68.0	70.0	+2.0
Psych_CalculusAB	69.0	71.0	+2.0
CalculusAB_Psych	67.5	69.5	+2.0
Psych_Stats	71.0	70.0	-1.0
Stats_Psych	71.5	70.5	-1.0
Psych_CompLit	70.0	69.5	-0.5
CompLit_Psych	70.0	69.5	-0.5
Latin_CalculusAB	69.0	70.5	+1.5
CalculusAB_Latin	71.0	72.0	+1.0
Latin_Stats	71.0	68.5	-2.5
Stats_Latin	71.5	70.5	-1.0
Latin_CompLit	67.0	71.0	+4.0
CompLit_Latin	70.0	70.5	+0.5
CalculusAB_Stats	69.0	71.0	+2.0
Stats_CalculusAB	69.0	69.0	0.0
CalculusAB_CompLit	69.5	72.0	+2.5
CompLit_CalculusAB	69.0	71.5	+2.5
Stats_CompLit	70.0	68.0	-2.0
CompLit_Stats	70.0	68.0	-2.0

Note. Accuracy is averaged over five trials on 60 AP-style CS multiple-choice questions. Rows show the performance change with Chain-of-Thought prompting. Green highlights indicate improvements (darker = larger gain), red indicates declines, and gray indicates no change.

General Improvement Across Most Conditions Notably, while CS underperformed the Raw baseline in the absence of CoT prompting, this trend reversed when CoT

was introduced: CS with CoT achieved an accuracy of 73.0%, exceeding the Raw baseline with CoT (72.5%). This shift suggests that CoT may enhance the model's ability to engage with and apply prior training more effectively. In this case, the structured nature of CoT could have supported deeper integration or retrieval of CS-relevant knowledge, even though the training materials remained unchanged. Additionally, every combination that included CS showed an improvement with CoT, further supporting this trend.

Overall, CoT yielded a statistically significant average accuracy increase of 1.23 percentage points across all subject pairings ($p = 0.0004$). At the 5% significance level, four training conditions demonstrated significant improvements: CS (+5.5 percentage points), CS-Econ (+3.5), Latin-CS (+3.5), and Psych-CalculusAB (+2.0). In contrast, only two pairings exhibited statistically significant declines: Psych-Latin (–3.0) and Latin-Stats (–2.5).

These patterns can be visualized in Figure 3, which shows the change in accuracy after applying Chain-of-Thought prompting across all subject pairings. Bars to the right indicate performance gains, while bars to the left indicate declines.

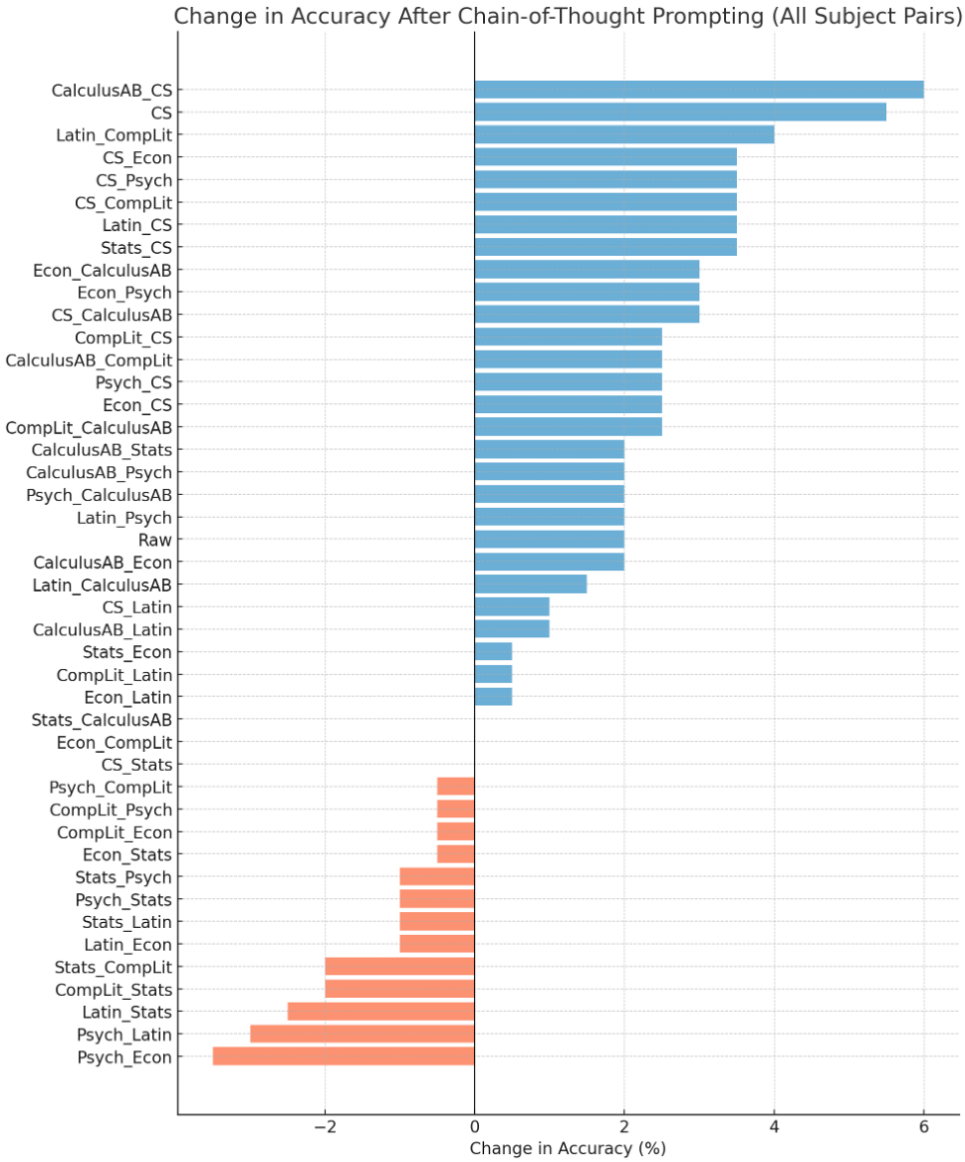


Figure 3: Change in Accuracy After Chain-of-Thought Prompting Across All Subject Pairings. Blue bars indicate performance increases; red bars indicate decreases. Results reflect the difference in average accuracy (in percentage points) between the CoT and non-CoT conditions for each training configuration.

Evaluating the Impact of CoT on Order Sensitivity To assess how CoT prompting influences the effect of subject order, we conducted two complementary analyses: one focusing on directional shifts, and the other on stability.

First, we compared the directional differences in performance between reversed subject pairings (e.g., *Psych-Econ* vs. *Econ-Psych*) before and after CoT prompting. The average change in order effect (CoT minus no-CoT) was minimal at +0.048 percentage points, with no statistically significant difference detected ($p = 0.930$, paired t-test; $p = 0.856$, Wilcoxon signed-rank). This indicates that CoT prompting does not reliably change which subject order yields better performance.

Next, we examined whether CoT reduces the overall sensitivity to order by analyzing changes in the absolute value of order effects. Here, CoT was associated with an average reduction of 0.619 percentage points in absolute value of the effect, suggesting a modest stabilizing effect. While this reduction was not statistically significant under the Wilcoxon test ($p = 0.124$), it was marginally significant under the paired t-test ($p = 0.085$), indicating weak evidence that CoT reduces variability introduced by subject order.

Together, these results suggest that CoT prompting does not alter the direction of order effects but may slightly diminish their magnitude, potentially improving model stability in interdisciplinary contexts. To further explore this stabilizing effect, we shift focus from directional ordering to overall performance outcomes across pairings under each condition.

To isolate the effect of Chain-of-Thought prompting independent of subject ordering, we visualized the change in CS performance across all pairings without emphasizing directional effects (i.e., whether the header subject was introduced first or second). As shown in Figure 4, accuracy under CoT (black) almost always consistently improves upon or remains comparable to the No-CoT condition (pink). In addition to a general rightward shift, many of the black segments are shorter, suggesting increased stability.

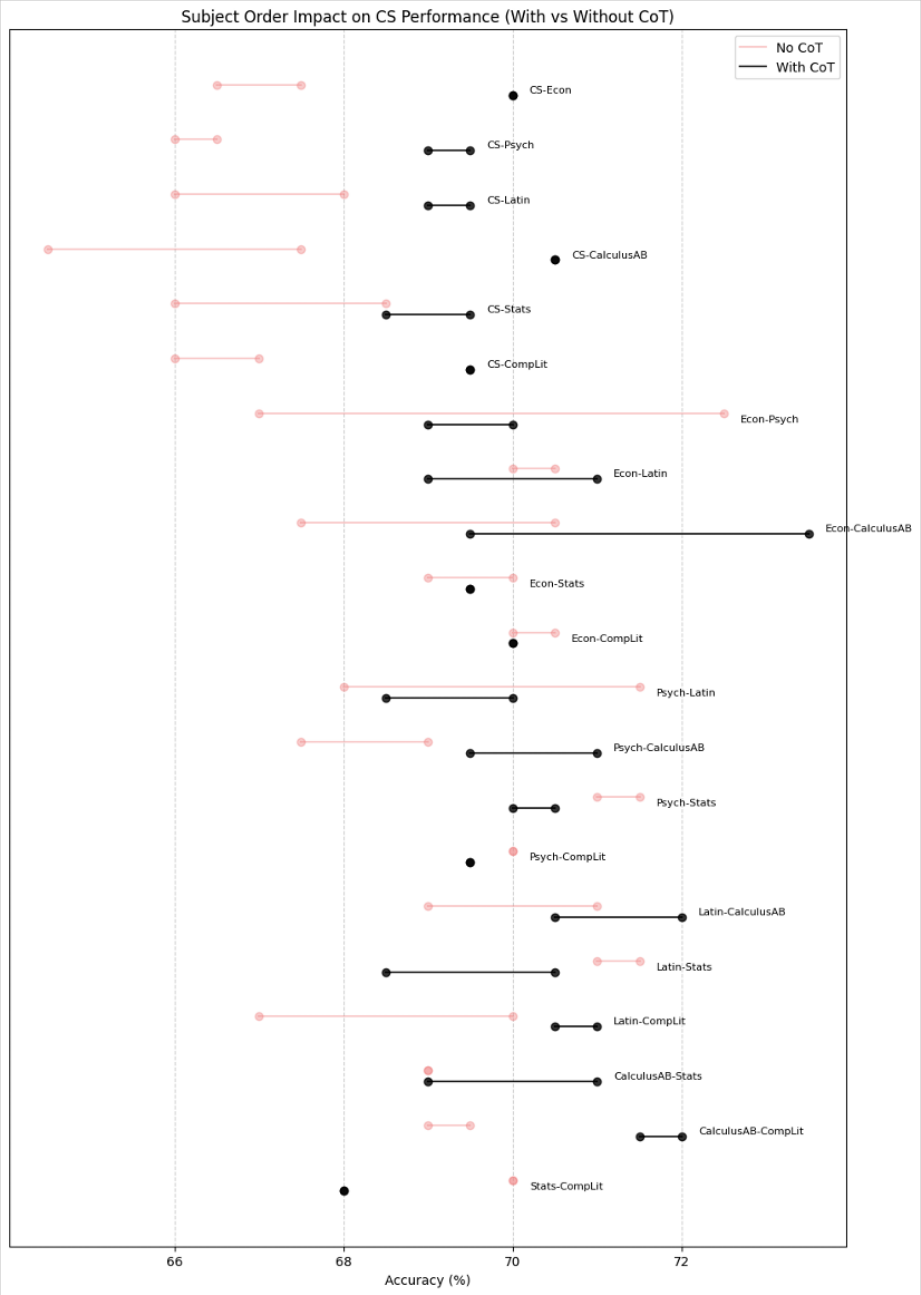


Figure 4: Comparison of CS Accuracy With and Without Chain-of-Thought Prompting Across Subject Pairings. Each line shows accuracy before (pink) and after (black) CoT prompting for a specific subject pairing. Directionality is omitted to focus on overall effect. CoT generally increases or stabilizes performance, with less variation across configurations.

Section 0.6

Future Directions

This study introduces a scalable framework for evaluating interdisciplinary learning in LLMs and reveals several avenues for future exploration. While our analysis centered on CS tasks, the findings suggest that interdisciplinary effects—both positive and negative—are highly dependent on the specific subject pairing, order of exposure, and reasoning strategy employed.

First, future research should investigate whether the observed patterns generalize beyond CS. While CS served as a controlled benchmark, evaluating additional target subjects such as history, physics, or biology would help determine whether certain domains are more sensitive to cross-subject transfer. Additionally, extending evaluations to non-multiple-choice formats (e.g., free response, open-ended reasoning) could test whether the observed trends persist under less structured conditions.

Second, our findings suggest that CoT prompting may enhance not only reasoning but also the model's ability to absorb and apply previously seen material. In particular, we observed that while CS alone underperformed the Raw baseline without CoT, it surpassed Raw with CoT—despite having identical training content. This suggests that CoT may help the model engage with training materials more effectively, potentially increasing the depth or salience of learned representations. Future work should investigate whether this effect generalizes across subjects and whether other reasoning strategies (e.g., few-shot exemplars, self-reflection prompts) yield similar benefits.

Third, the role of subject order in shaping learning outcomes deserves continued attention. While most pairs showed stable or minimal order effects, certain subjects—such as Psychology and Latin—displayed directional sensitivity depending on when they were introduced. Additional experiments are needed to test whether this asymmetry is driven by semantic properties of the subjects, cognitive load balancing, or representational overlap. Moreover, adaptive ordering strategies could be explored to determine whether models can

learn optimal subject sequences for maximal transfer.

Finally, deeper interpretability analyses are needed to understand the internal mechanisms driving these effects. Do models update internal attention patterns, intermediate representations, or token associations in ways that mirror human schema formation? Leveraging techniques such as probing classifiers, representation similarity analysis, and attention diagnostics could shed light on the nature of interdisciplinary integration in LLMs.

In sum, this study opens up rich questions at the intersection of education, cognitive science, and AI interpretability. As LLMs continue to be deployed in learning and decision-making settings, understanding how they process and integrate information across domains will be crucial for building reliable and pedagogically aligned AI systems.

Section 0.7

Conclusion

This study presents a novel framework for evaluating how LLMs respond to interdisciplinary exposure, with a particular focus on subject pairing, ordering effects, and the role of Chain-of-Thought prompting. By systematically varying the sequence and combination of subject matter provided to GPT-4o, we demonstrate that LLM performance is sensitive to both the content and structure of training inputs.

Our results reveal that interdisciplinary exposure can yield both gains and interference, depending on the conceptual relationship between subjects and the order in which they are introduced. Notably, certain subjects—such as Psychology—consistently improved downstream performance when introduced first, while others—such as Latin—tended to benefit more when introduced second. These findings highlight the directional nature of knowledge transfer in LLMs, suggesting that some domains serve as more effective primers than others.

We also find that CoT prompting significantly boosts performance across most subject combinations, particularly in structured domains like Computer Science. In some cases, CoT not only improved reasoning accuracy but also appeared to enhance the model's ability to

engage with training content—allowing CS to outperform a stronger Raw baseline when paired with CoT. While CoT did not eliminate the impact of subject ordering, it showed modest evidence of stabilizing performance by reducing the variability caused by sequencing.

Together, these findings suggest that interdisciplinary learning in LLMs is neither uniformly beneficial nor neutral—it is context-dependent, order-sensitive, and modifiable through prompt design. This work contributes to a deeper understanding of how LLMs integrate information across domains, and offers practical implications for how educational content and AI training pipelines might be structured to enhance generalization. Future work will further probe the cognitive parallels and mechanistic underpinnings of these effects, informing both AI alignment and the design of next-generation educational tools.

Section 0.8

Acknowledgements

I extend my gratitude especially to my advisors, Professors Soroush Vosoughi, Daniel Rockmore, and my Mathematical Data Science Thesis Advisor Peter Mucha, for their guidance and support throughout this thesis.

I am also deeply thankful to my friends and family, whose joy and encouragement inspired me along the way.

Section 0.9

Code and Reproducibility

All code, data, and experiment configurations used in this thesis are available on GitHub: `https://github.com/wliang-whl/Liberal-Art-Project.git`

Bibliography

Carmichael, T. and LaPierre, Y. (2014). Interdisciplinary learning works: The results of a comprehensive assessment of students and student learning outcomes in an integrative learning community. *Issues in Interdisciplinary Studies*, (32):53–78. Published by the Association for Interdisciplinary Studies.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201.

Gao, C., Lan, X., Lu, Z., Mao, J., Piao, J., Wang, H., Jin, D., and Li, Y. (2023). S3: Social-network simulation system with large language model-empowered agents. arXiv preprint arXiv:2307.14984.

Leng, Y. and Yuan, Y. (2024). Do llm agents exhibit social behavior? arXiv preprint arXiv:2312.15198.

Li, Y., Zhang, Y., and Sun, L. (2023). Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. arXiv preprint arXiv:2310.06500.

Liu, H.-Y., Hsu, D.-Y., Han, H.-M., Wang, I.-T., Chen, N.-H., Han, C.-Y., Wu, S.-M., Chen, H.-F., and Huang, D.-H. (2022). Effectiveness of interdisciplinary teaching on creativity: A quasi-experimental study. *International Journal of Environmental Research and Public Health*, 19(10):5875.

Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., and Durrett, G. (2024). To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. arXiv preprint arXiv:2212.09196.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xu, C., Wu, C.-F., Xu, D.-D., Lu, W.-Q., and Wang, K.-Y. (2022). Challenges to student interdisciplinary learning effectiveness: An empirical case study. *Journal of Intelligence*, 10(4):88.

Section .1

Appendix

.1.1. Subject Cross-Impact Visualization

To provide a more comprehensive overview of the preliminary interdisciplinary effects on GPT's ability to generalize knowledge, we include a visualization of subject impacts when preliminarily tested using GPT-3.5-Turbo.

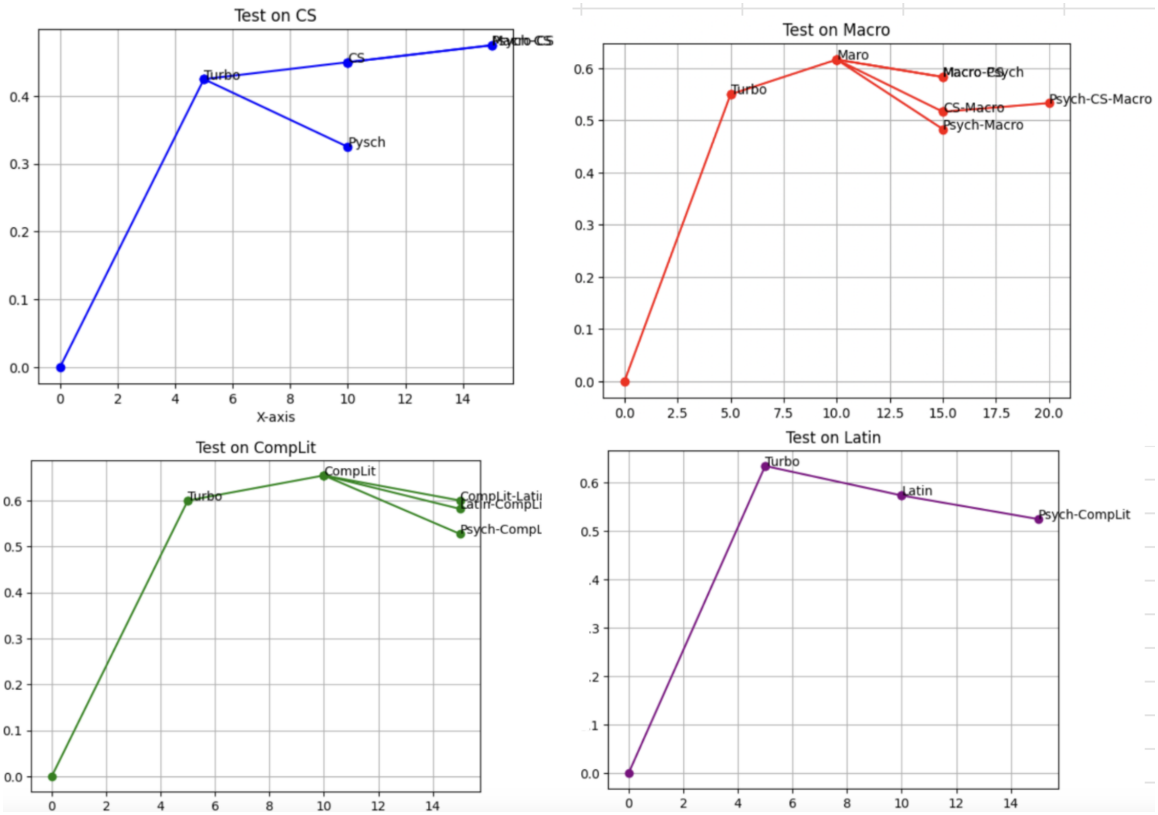


Figure 5: Subject cross-impact visualization across four test subjects (CS, Macroeconomics, CompLit, Latin). Among them, CS exhibits the greatest variation in response to interdisciplinary pairings, showing both strong gains and declines across conditions.