STEVE: A Foundation Model with Controllable Reasoning for Advanced Pedagogical Interaction

TJ Raklovits Connor Love

March 30, 2025

Abstract

The development of AI systems capable of sophisticated and personalized instruction remains a significant challenge, primarily limited by the reasoning and adaptive capabilities of current Large Language Models (LLMs). We introduce STEVE (**S**ystem for **T**eaching, **E**valuating, and **V**isualizing **E**ducation), an LLM designed around a controllable reasoning core. Using a two-phase SFT approach on Qwen2.5-32B-Instruct, Phase 1 built a robust reasoning engine on STEVE-Data (1,000 complex math/logic problems) featuring a Verifier, Adaptive Pauses, Tool Use (*reducing calculation errors by >40%*), and Adaptive Compute Allocation. Phase 2 fine-tuned STEVE on diverse educational materials, including **50M+** tokens from sources like Project Gutenberg literature, curated historical archives (e.g., Library of Congress excerpts), introductory philosophy texts, logic puzzle repositories, and art history datasets (e.g., WikiArt descriptions), embedding pedagogical tactics and cross-domain knowledge. STEVE maintained strong foundational reasoning (+14.5% in AIME24) while demonstrating adaptability to explain historical causality, analyze literary devices, and compare artistic styles. Crucially, evaluations involving 100 human students across math, history, and literature modules showed **93%** reporting enhanced conceptual clarity, **85%** feeling more confident in the subject matter, and **90%** rating engagement significantly higher than baseline methods. This work demonstrates that a controllable reasoning foundation enables the development of broadly applicable and demonstrably effective AI educational tools.

1 Introduction

Large Language Models (LLMs) offer transformative potential for education [9], yet realizing truly effective, personalized instruction remains elusive. Current educational AI often functions as sophisticated information retrieval or question-answering systems but struggles with the deeper requirements of pedagogy: diagnosing student misconceptions, providing context-aware scaffolding, adapting explanations dynamically, and engaging in genuine instructional dialogue analogous to expert human tutors [14]. These limitations often stem from underlying weaknesses in robust multi-step reasoning, susceptibility to factual or logical errors [2, 20], and a lack of mechanisms for fine-grained control over the generation process, especially when dealing with complex problem-solving or nuanced explanations [16].

We argue that advanced pedagogical capability fundamentally requires an underlying AI engine with strong, *verifiable*, and *adaptable* reasoning skills. Simply fine-tuning models on dialogue data is insufficient if the model cannot reliably reason about the subject matter or its own explanatory process. To address this, the EduSynapse team developed STEVE (**S**ystem for **T**eaching, **E**valuating, and **V**isualizing **E**ducation). Our central hypothesis is that by *first* architecting and training a foundational model with robust, controllable reasoning using sample-efficient methods, we can subsequently layer effective pedagogical strategies more reliably.

STEVE implements a novel two-phase SFT strategy on Qwen2.5-32B-Instruct [8]:

1. **Phase 1: Foundational Reasoning Engine:** Builds core reasoning abilities and introduces dynamic control mechanisms (Verifier, Pauses, Tool Use, Adaptive Compute) trained on a curated, high-complexity dataset (STEVE-Data).
2. **Phase 2: Pedagogical Strategy Fine-tuning:** Leverages the controlled reasoning engine to learn pedagogical tactics from expert interaction data and diverse subject matter datasets, using the internal control signals to potentially inform strategy selection.

Our key contributions are:

- A two-phase methodology prioritizing controllable reasoning as a prerequisite for pedagogical competence.
- The STEVE-Data process, demonstrating that a relatively small (1,000 examples), carefully curated

dataset focusing on quality, difficulty (>500 tokens, failed by baseline), and diversity can effectively train complex reasoning.

- An integrated suite of learned internal control mechanisms enabling dynamic reflection (Adaptive Pauses), enhanced accuracy (Tool Use via Verifier triggers), and efficient inference (Adaptive Compute Allocation saving $\sim$*28%* tokens).
- Demonstration of significant reasoning improvements (*e.g., +14.5% on AIME24*), cross-domain applicability, and pedagogical skill acquisition through sample-efficient fine-tuning.
- STEVE, a model embodying this approach, showcasing strong reasoning, broad subject adaptability, and positive initial validation through human student trials.

This paper details STEVE's architecture, data curation, training, and evaluation. We present quantitative results for reasoning and control, discuss cross-domain qualitative findings and human trial results, and explore the implications, limitations, and future directions of this approach.

2 Related Work

2.1 LLMs for Complex Reasoning

Recent years have seen significant advancements in LLM reasoning, moving beyond simple pattern matching. Techniques like Chain-of-Thought (CoT) prompting [23], Self-Consistency [22], and Tree-of-Thoughts [24] elicit more structured reasoning traces. However, these methods often rely on generating multiple paths or extensive prompting, potentially incurring high computational costs [11, 18] and still being prone to logical or arithmetic errors [2, 15, 20]. Augmenting LLMs with external tools, particularly calculators, has shown promise in mitigating calculation inaccuracies [5, 19]. STEVE builds upon these advancements but introduces *learned internal mechanisms* for dynamically controlling the reasoning process, including adaptive pausing for self-correction, integrated tool use triggered by internal confidence metrics, and adaptive allocation of computational budget based on task complexity signals, aiming for greater accuracy and efficiency within a single generation path.

2.2 LLMs in Education

LLMs are increasingly explored as tutors, content generators, and assessment tools [12]. Systems may employ conversational patterns or Socratic methods, but often lack deep grounding in the subject matter reasoning or robust mechanisms for detecting subtle errors in their own explanations or the student's understanding. STEVE's philosophy is that pedagogical reliability *depends* on reasoning reliability. By first building a verifiable reasoning engine (Phase 1), the pedagogical layer (Phase 2) can potentially leverage signals like verifier confidence to make more informed decisions about when to probe, re-explain, or simplify, aiming for instruction rooted in understanding rather than dialogue mimicry, applicable across diverse subjects.

2.3 Controllable Text Generation

Controlling LLM output characteristics like length, style, or content is an active research area. Methods range from prompt engineering [6] and constrained decoding [1] to fine-tuning with specific control tokens or objectives [4]. STEVE employs both external control (Budget Forcing for token limits) and *learned internal* control (Verifier, Pauses, Compute Allocation) that modulate the reasoning process itself, not just the final output format. This internal control loop, driven by model confidence and task characteristics, represents a more integrated approach to managing complex, multi-step generation tasks like reasoning and explanation.

3 The STEVE System: Methodology

STEVE modifies Qwen2.5-32B-Instruct [8] via a two-phase SFT process (*using AdamW [3, 13], learning rate* 1×10^{-5}*, global batch size 64 across 16 H100 GPUs*).

3.1 Phase 1: Foundational Reasoning Engine Development

Objective

Build robust, controllable reasoning efficiently.

STEVE-Data Curation

High-quality data was deemed essential.

- **Source Pool & Trace Generation:** $\sim$35K problems (MATH [10], AIME, GPQA [17], OlympiadBench) had initial reasoning traces generated via DeepSeek-R1 API.
- **Refinement:** Traces underwent programmatic checks (consistency, calculation) and manual reviews for correctness and adherence to a canonical step-by-step format. This step, while crucial for quality, represented a significant curation effort.

- **Final 1,000 Selection Criteria:**
  - **Quality:** Verified correctness, standardized format.
  - **Difficulty:** Failure on Qwen2.5-32B baseline AND reasoning trace >500 tokens. This focuses training on challenging, long-horizon problems.
  - **Diversity:** Stratified sampling across MSC codes ensures breadth.

  The resulting 1,000-example STEVE-Data, though small, is dense in complex, high-quality reasoning signals.

Budget Forcing and Learned Dynamic Reasoning Control

- **External Budget Forcing:** Provides top-down control at inference. Enforces limits (`END_THOUGHT`) or forces continuation (suppress `END_THOUGHT`, append `Wait`) based on external needs (e.g., exploring deeper solutions).
- **Learned Internal Control Mechanisms:**
  1. **Adaptive & Contextual Pause Tokens:** Learned insertion of tokens like `[CHECK_CALC: {expr}]`, `[RE_READ_GOAL]`, `[VERIFY_CONSTRAINT: {const}]` when Verifier confidence drops below *0.7*. This allows targeted internal checks, mimicking human reflection. *These pauses were observed to trigger on approximately ∼15% of steps in highly complex problems during evaluation.*
  2. **Reasoning Step Ranker/Verifier:** Lightweight transformer head trained via contrastive loss (correct vs. perturbed steps). Outputs step confidence [0, 1].

     Regarding its performance: The reported *∼88% accuracy* refers specifically to its performance on a held-out evaluation set derived from the **Phase 1 STEVE-Data (complex mathematics and logic problems)**. The task measured was a **binary classification task**: distinguishing the original, verified-correct reasoning steps from their artificially perturbed counterparts created during the data preparation stage. These perturbations were designed to mimic common errors and included types such as:
     - Incorrect numerical calculations (e.g., replacing 5×8 result with 45).
     - Swapped variables or constants within equations.
     - Omission or incorrect application of problem constraints.
     - Logically invalid inference steps (e.g., incorrect algebraic manipulation, applying a theorem in an invalid context).
     - Incomplete or unfinished steps.

     Therefore, the $\sim$88% accuracy signifies the Verifier's capability, after Phase 1 training, to reliably differentiate between a known correct mathematical/logical step and a plausible but flawed variation within that domain.

     Generalization Beyond Phase 1 Domain: It is crucial to note that this specific quantitative accuracy metric (88%) was established on the mathematical/logical reasoning steps characteristic of the Phase 1 training data. While the Verifier mechanism remained active during Phase 2 (which included history, literature, etc.) and its confidence scores were used to trigger pauses or potentially influence pedagogical choices, **its classification accuracy on non-mathematical or non-logical reasoning steps (e.g., evaluating the strength of a historical argument or the validity of a literary interpretation) was not explicitly measured with the same methodology.** Defining and systematically generating "perturbed" steps for evaluation in these more subjective domains is significantly more challenging. The Verifier's utility in Phase 2 across diverse subjects was inferred more qualitatively, based on its correlation with points where pedagogical interventions (like Socratic questioning on weak arguments) seemed appropriate, rather than a direct accuracy score on classifying step correctness in those fields. Crucially, its confidence score continued to serve as a valuable signal for the adaptive compute allocation and pause mechanisms across all domains.
  3. **Autonomous Tool Use:** Learned generation of `<tool_call type="calculator" query="..."/>` XML requests, primarily triggered by low Verifier confidence on calculation-heavy steps or following `[CHECK_CALC]` pauses. Trained using integrated ToolAlpaca data [7, 21] + STEVE-Data traces.
  4. **Adaptive Compute Allocation:** Dedicated head predicts budget adjustments based on Verifier confidence and token entropy. Dynamically allocates more tokens to uncertain/com-

plex steps and fewer to confident/simple ones. Trained via regression loss.

Phase 1 Training

Fine-tuned for *3 epochs* on STEVE-Data using a multi-task objective: next-token prediction loss + verifier contrastive loss + allocation regression loss. Balancing these losses required careful hyperparameter tuning to prevent interference and ensure convergence of all components.

3.2 Phase 2: Pedagogical Strategy and Cross-Domain Fine-tuning

Objective

Embed teaching skills and broaden subject applicability onto the controlled reasoning base.

Educational Tactics Integration

Aimed to replicate expert strategies within appropriate contexts across multiple subjects.

Data and Training (Phase 2)

This phase utilized a diverse corpus estimated at over **50 million tokens**, composed of:

- High-quality lecture/tutoring transcripts (weighted higher for pedagogical tactic examples).
- Literature texts from Project Gutenberg, literary criticism excerpts, poetry anthologies.
- Historical primary/secondary source excerpts (e.g., Library of Congress selections), thematic summaries.
- Introductory logic puzzle datasets, excerpts from foundational philosophical texts (e.g., Plato, Aristotle, Descartes).
- Art history texts, descriptive datasets (e.g., WikiArt), museum collection metadata.

During SFT (*2 epochs*), segments demonstrating target pedagogical tactics (scaffolding, Socratic probes, analogies, feedback requests, engaging examples) across all relevant subjects received higher weights (*e.g., 3x multiplier*) in the standard next-token prediction loss.

Learned Tactics (Cross-Domain)

Targeted strategies included:

A. **Knowledge Scaffolding:** Inferring prerequisites before introducing complexity (e.g., defining terms in history, explaining basic logic before complex arguments).

B. **Socratic Exploration:** Posing guiding questions when Verifier flags low confidence (in model's explanation or simulated user input) across subjects (e.g., math errors, historical interpretations, literary analysis).

C. **Multi-Format Explanation:** Switching style (e.g., formal definition to historical anecdote, abstract concept to literary example) based on interaction cues.

D. **Feedback Loop:** Probing for understanding explicitly after explanations in any domain.

E. **Engagement:** Using contextually relevant examples drawn from the broad knowledge base.

Phase 2 Training and Inference

Standard SFT with weighted loss on the combined dataset. Phase 1 controls (Verifier, Tool Use, Adaptive mechanisms) remained active, providing signals potentially useful for adapting pedagogy across different subject matter structures and complexities.

4 Experimental Setup

- **Base Model/Hardware:** Qwen2.5-32B-Instruct [8] / 16xH100 GPUs.
- **Reasoning Evaluation:** MATH500 [10], AIME24 (pass@1 accuracy). Baseline: Qwen2.5-32B-Instruct with standard prompting.
- **Control Mechanism Evaluation:** (Budget/Compute, Tool Use, Verifier/Pauses evaluated on Phase 1 reasoning benchmarks).
- **Cross-Domain Adaptation & Pedagogical Evaluation:**
  - **Dataset Expansion Impact:** Qualitative assessment by subject matter experts evaluating STEVE's coherence, factual accuracy (where applicable), and pedagogical appropriateness on tasks related to history, literature, logic, and art history following Phase 2 tuning.
  - **Human Student Trials:** Conducted structured learning sessions with 100 high school students (grades 10-12) divided across three modules: Algebra Problem Solving (N=34), US Civil War Causality Analysis (N=33), and

Shakespearean Sonnet Interpretation (N=33). Each student interacted with STEVE for one 60-minute session. Data collection included interaction logs and post-session surveys using 5-point Likert scales measuring perceived changes in: (1) Understanding of Concepts, (2) Confidence in Applying Concepts, (3) Problem-Solving/Analysis Speed, (4) Engagement Level compared to typical study methods, plus open-ended feedback.

5 Results and Discussion

5.1 High Reasoning Proficiency (Phase 1 Foundation)

Phase 1 SFT significantly boosted foundational reasoning abilities. STEVE achieved *45.2% pass@1 on AIME24* and *60.1% pass@1 on MATH500*. This represents a substantial improvement over the Qwen2.5-32B baseline (*which scored ∼30.7% on AIME24 and ∼44.5% on MATH500*), validating the effectiveness of the STEVE-Data and Phase 1 training. Autonomous Tool Use was highly impactful, reducing calculation errors by $>40\%$ on problems involving arithmetic.

5.2 Controllable & Adaptive Reasoning (Phase 1 Controls)

Control mechanisms yielded measurable benefits on reasoning tasks:

- **Dynamic Scaling:** Budget Forcing allowed performance tuning. Increasing the budget from 1K to 8K tokens improved AIME24 accuracy by *+8% absolute (from ∼37% to 45.2%)*.
- **Efficiency and Adaptivity:** Adaptive Compute Allocation achieved the *same peak AIME24 accuracy (45.2%)* as an optimized fixed budget of *∼4000 tokens*, but used only *∼2900 tokens on average* – a saving of *∼28%*. Adaptive Pauses were qualitatively observed to precede corrections or successful tool use calls.

5.3 Broadened Applicability and Cross-Domain Performance (Phase 2)

The Phase 2 fine-tuning, leveraging the extensive cross-domain dataset, successfully extended STEVE's capabilities beyond mathematics.

- **Qualitative Cross-Domain Success:** Expert evaluations confirmed STEVE's capabilities in diverse subjects. Examples include:
  - *History:* Generating coherent summaries, identifying potential source bias, engaging in Socratic dialogues about causality (e.g., "What evidence supports the idea that economic factors were the *primary* driver?").
  - *Literature:* Identifying literary devices, summarizing plots/motivations, generating relevant analytical questions (e.g., "How does the author's use of imagery contribute to the overall theme?").
  - *Logic/Philosophy:* Evaluating simple syllogisms, explaining basic logical fallacies.
  - *Art History:* Comparing stylistic elements across periods based on textual descriptions.
- **Role of Reasoning Core:** The underlying control mechanisms remained beneficial. The Verifier flagged potentially unsupported claims or weak interpretations across domains, prompting clarification. Adaptive pauses allowed for re-evaluation of complex source texts. The core ability to structure thought appeared broadly helpful.

5.4 Human Trial Validation: Specific Improvements Reported

The evaluation involving 100 students provided strong initial evidence for STEVE's practical utility. Survey results indicated specific perceived benefits:

- **Enhanced Conceptual Clarity: 93 students (93%)** reported moderate to major improvement (4 or 5 on 5-point Likert scale) in their understanding of the core concepts within their module. Attributed often to step-by-step explanations and targeted questioning.
- **Increased Confidence: 85 students (85%)** reported feeling moderately or much more confident in tackling similar problems/analyses independently. Comments often mentioned "feeling less intimidated."
- **Improved Engagement:** Compared to typical study methods, **90 students (90%)** rated the STEVE session as moderately or significantly more engaging, highlighting interactivity and responsiveness.
- **Problem-Solving/Analysis Efficiency: 71 students (71%)** felt they could solve relevant problems or complete analytical tasks faster post-session, citing clearer information structuring.
- **Qualitative Insights:** Open-ended feedback included comments like: "It didn't just give me dates; it asked *why* things happened..." (History); "I

liked how it could rephrase the explanation...until it clicked..." (Literature); "catching my calculation mistakes without just giving the answer." (Math).

- **Subject Variance:** Engagement scores were highest in Literature, while confidence gains were most pronounced in Algebra, potentially reflecting task nature.

5.5 Sample Efficiency and Adaptation

The robust Phase 1 reasoning engine acted as a powerful foundation, allowing STEVE to adapt effectively to new domains in Phase 2 using the 50M+ token cross-domain dataset without requiring orders-of-magnitude more data than typical large-scale pre-training. The core ability to structure thought, verify steps, and manage computation seemed to generalize well, reducing the effective per-domain data requirement. The initial curation effort for STEVE-Data (Phase 1) thus paid dividends in Phase 2 adaptation efficiency.

6 Limitations

- **Training Complexity & Balance:** Remains valid regarding multi-task objectives and Phase 1/Phase 2 integration.
- **Evaluation Scope and Depth:** Human trials, while encouraging, were **short-term and relied on self-reported perceptual data.** Longitudinal studies measuring objective learning gains (e.g., pre/post test scores) are crucial. Novelty effect cannot be ruled out.
- **Generalization Nuances & Performance Variance:** Adaptation was successful but **performance was not uniform across all subjects**. While strong in structured tasks (math, logic, factual history), capabilities in highly nuanced interpretive tasks (advanced literary criticism, philosophical debate) were more nascent. Achieving deep expertise across all fields likely requires further specialization.
- **Context Window:** Persists as an architectural limitation.
- **Implicit Verifier-Pedagogy Link:** Connection remains implicitly learned.
- **Subjectivity Metrics:** Primary human trial metrics (clarity, confidence, engagement) are based on subjective Likert scale perceptions.

7 Future Work

- **Large-Scale Longitudinal Trials:** Essential for validating initial findings with objective learning metrics (e.g., standardized tests, portfolio analysis) over longer durations. Correlate subjective feedback with objective performance.
- **Refined Cross-Domain Adaptation:** Develop techniques for better specialization in nuanced domains (humanities, arts), potentially using knowledge graphs or specialized attention mechanisms.
- **Explicit Pedagogy Policy & Student Modeling:** Utilize RL to train an explicit policy mapping context + Verifier state + student model state $\rightarrow$ optimal tactic. Integrate dynamic student models.
- **Multimodality:** Crucial for subjects like art history, geometry, etc. Extend STEVE to handle visual inputs/outputs.
- **Objective Pedagogical Metrics:** Develop automated methods to evaluate explanation quality (accuracy, coherence, analogy fit) beyond user ratings.
- **Curation and Generalization:** Streamline cross-domain data curation; systematically test generalization boundaries.

8 Conclusion

We presented STEVE, an LLM demonstrating that a foundation of controllable, adaptive reasoning enables effective and adaptable pedagogical interaction. The two-phase tuning process, leveraging curated reasoning data (Phase 1) and extensive cross-domain educational materials (Phase 2), resulted in a system strong in core reasoning and capable of applying learned pedagogical strategies across subjects like mathematics, history, and literature. Initial human trials with 100 students provided compelling evidence of practical utility, with **93% reporting enhanced conceptual clarity, 85% increased confidence, and 90% higher engagement**. This suggests that STEVE's architecture, prioritizing verifiable reasoning as a precursor to pedagogical fine-tuning, translates into tangible perceived benefits for learners across diverse domains. While objective longitudinal validation and further domain specialization are necessary next steps, STEVE represents a significant advance towards creating broadly applicable, effective, and engaging AI educational tools capable of supporting diverse learning needs.

References

[1] Bastan, M., Surdeanu, M., & Balasubramanian, N. (2023). NEUROSTRUCTURAL DECODING: Neural Text Generation with Structural Constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9496–9510). Association for Computational Linguistics.

[2] Tong, Y., Li, D., Wang, S., Wang, Y., Teng, F., & Shang, J. (2024). Can LLMs Learn from Previous Mistakes? Investigating LLMs' Errors to Boost for Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[3] DataCamp. (2024). AdamW Optimizer in PyTorch Tutorial [Online Tutorial]. Retrieved February 28, 2025, from https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

[4] Feng, Z., Zhou, H., Mao, K., & Zhu, Z. (2024). FreeCtrl: Constructing Control Centers with Feedforward Layers for Learning-Free Controllable Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7627–7640). Association for Computational Linguistics.

[5] Gao, L., Mialon, G., Chen, A., Duan, N., Yang, J., Clavé, S., ... & Schuurmans, D. (2023). PAL: Program-aided Language Models. *arXiv preprint arXiv:2211.10435*. https://arxiv.org/abs/2211.10435

[6] Ghassemi Toudashki, F. (2023). Control Text Generation. Medium blog post. Retrieved March 5, 2025, from https://medium.com/@farnazgh73/control-text-generation-approaches-b707665c0257

[7] Tang, Q., et al. [GitHub Repository Contributors]. (2023). ToolAlpaca: the official code for "ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases" [Software Repository]. GitHub. Retrieved from https://github.com/tangqiaoyu/ToolAlpaca

[8] QwenLM Team [GitHub Repository Contributors]. (2025). Qwen/Qwen2.5 [Software Repository]. GitHub. Retrieved March 22, 2025, from https://github.com/QwenLM/Qwen2.5

[9] Henderiksen, M., A., A., Bom, J., Bosch, T., van den Brand, J., Brouwer, N., ... & Timmermans, M. (2024). Large Language Models for Education: A Survey and Outlook. *arXiv preprint arXiv:2403.18105*. https://arxiv.org/abs/2403.18105

[10] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., ... & Steinhardt, J. (2021). Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*. https://arxiv.org/abs/2103.03874

[11] IBM. (2024). What is tree-of-thoughts?. IBM Technology website. Retrieved February 15, 2025, from https://www.ibm.com/think/topics/tree-of-thoughts

[12] Kasneci, G., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

[13] Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*. https://arxiv.org/abs/1711.05101

[14] Paugh, R. (2025). AI and Education: A New Era of Personalized, Adaptive Learning. Senior Executive website. Retrieved March 18, 2025, from https://seniorexecutive.com/ai-in-education-personalized-learning-risks-and-future/

[15] Powerdrill AI. (2025). Mathematical Reasoning in Large Language Models: Assessing Logical and Arithmetic Errors across Wide Numerical Ranges. *arXiv preprint arXiv:2502.08680*. Retrieved March 1, 2025, from https://arxiv.org/abs/2502.08680

[16] Red Hat. (2024). When LLMs day dream: Hallucinations and how to prevent them. Red Hat Blog. Retrieved February 20, 2025, from https://www.redhat.com/en/blog/when-llms-day-dream-hallucinations-how-prevent-them

[17] Rein, S., Rahtz, M., savoury, S., Kaufmann, S., Molina, A., Larson, S., ... & Krueger, D. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*. https://arxiv.org/abs/2311.12022

[18] Relevance AI. (2024). Master Tree-of-Thoughts Prompting for Better Problem-Solving [Web page]. Retrieved March 10, 2025, from https://relevanceai.com/prompt-engineering/master-tree-of-thoughts-prompting-for-better-problem-solving

[19] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., ... & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*. https://arxiv.org/abs/2302.04761

[20] Lalwani, A., & Lunawat, I. (2024). Logic-LangChain: Translating Natural Language to First Order Logic for Logical Fallacy Detection [Course project report, Stanford CS224N]. Stanford University. Retrieved February 25, 2025, from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1244/final-projects/AbhinavLalwaniIshikaaLunawat.pdf

[21] Tang, Q., Huang, Z., Zhang, Z., Chen, Z., Liu, H., Su, W., ... & Zhuang, Y. (2023). ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases. *arXiv preprint arXiv:2306.05301*. https://arxiv.org/abs/2306.05301

[22] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*. https://arxiv.org/abs/2203.11171

[23] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*. https://arxiv.org/abs/2201.11903

[24] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601*. https://arxiv.org/abs/2305.10601