



# Transforming Confusion into Diffusion: Advancing Machine Learning Education via Bottom-Up Instruction

Carlos Cotrini\*  
ETH Zürich  
Zürich, Switzerland  
ccarlos@inf.ethz.ch

Sverrir Thorgeirsson\*  
ETH Zürich  
Zürich, Switzerland  
sverrir.thorgeirsson@inf.ethz.ch

Jesus Solano  
ETH Zürich  
Zürich, Switzerland  
jesus.solano@inf.ethz.ch

Zhendong Su  
ETH Zürich  
Zürich, Switzerland  
zhendong.su@inf.ethz.ch

## Abstract

Balancing conceptual depth with practical skill development is a persistent challenge in advanced machine learning (ML) education, where powerful frameworks can obscure underlying mathematical and computational principles. To address this, we define a new principled approach that we call full-stack machine learning (FSML), which emphasizes the construction of large language models and diffusion models from scratch. To evaluate the effectiveness of FSML, we conducted a classroom-based randomized controlled trial (N=208) in which FSML-based instruction was compared against a popular library-based instructional approach. We measured students' conceptual understanding through a specialized assessment and administered a survey capturing knowledge-gap awareness, curiosity, and cognitive load. We found that students who received FSML instruction performed approximately 10% better than control participants in a quiz on transformers and stable diffusion ( $p = 0.006$ ). They also showed increased curiosity and more positive affective responses, suggesting deeper engagement with ML fundamentals. Our findings indicate that our full-stack approach to ML education can improve student learning outcomes, potentially reshaping curricula for ML and other advanced computing topics.

## CCS Concepts

• **Social and professional topics** → **Computing education**; • **Human-centered computing** → **Empirical studies in HCI**; *User studies*.

## Keywords

machine learning education, tertiary education, cognitive load theory, stable diffusion

## ACM Reference Format:

Carlos Cotrini, Sverrir Thorgeirsson, Jesus Solano, and Zhendong Su. 2026. Transforming Confusion into Diffusion: Advancing Machine Learning Education via Bottom-Up Instruction. In *Proceedings of the 57th ACM Technical*

\*Co-primary author



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGCSE TS 2026, St. Louis, MO, USA*  
© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2256-1/26/02  
<https://doi.org/10.1145/3770762.3772595>

*Symposium on Computer Science Education V.1 (SIGCSE TS 2026), February 18–21, 2026, St. Louis, MO, USA. ACM, New York, NY, USA, 7 pages.*  
<https://doi.org/10.1145/3770762.3772595>

## 1 Introduction

Computer science (CS) educators face many dilemmas that are familiar to their colleagues from other disciplines, one of which is how to balance high-level conceptual approaches to the subject matter with foundational skill development. While top-down approaches that involve authentic and meaningful tasks can improve the motivation and situational interest of students [3, 12], they may leave students short of practical proficiency when not paired with consistent practice in fundamentals—a problem well known, for example, in functional literacy education [15, 21] and mathematics [13, 31]. Consequently, many educators favor a blend of top-down and bottom-up methods to capitalize on the strengths of both, although the best way to integrate them in computing education remains a topic of ongoing discussion [39].

In machine learning (ML) education, this question is particularly salient for several reasons. For one, modern ML frameworks allow students to build powerful models quickly with minimal knowledge of the underlying methods [22], which may drive student motivation and interest, but can have a negative impact on their grasp of the underlying computational, mathematical, and statistical concepts. High-level, black-box approaches may also mask foundational knowledge gaps that can surface later when students are confronted with complex AI literacy issues, such as algorithmic bias and model explainability [19]. Second, while the rapid evolution of CS as a discipline is known to place a significant strain on CS educators in their effort to stay current [1], innovations in ML have been particularly fast, making it especially challenging for pedagogical adaptations to keep pace. As a result, course materials that were state-of-the-art just a year ago can suddenly seem outdated or incomplete, making it difficult for educators to determine whether to emphasize procedural, framework-specific skills or focus on the enduring conceptual and mathematical underpinnings of ML.

In this work, we focus on the topic of *transformers* [40] and *stable diffusion* [28], as these breakthroughs led to popular models like ChatGPT and DALL-E. We present a new pedagogical approach to teaching machine learning, called *full-stack machine learning* (FSML), designed not only for specialized courses in machine learning, but also for engineering degree programs at the university

level. It only requires knowledge of linear algebra, statistics, and machine-learning programming at the bachelor’s level. It applies cognitive load theory to balance the need for conceptual depth and practical skills by teaching students how to build complete and complex ML models from scratch with minimal-to-zero reliance on external tooling and libraries. To our knowledge, this is the first framework that enables students to construct working large language models and diffusion models within a single semester. To evaluate the effectiveness of our approach, we conducted a randomized controlled trial with 208 participants recruited from a course on machine learning for undergraduate and graduate students. Our control intervention was a tutorial developed by Wang [41] from Harvard on implementing and understanding stable diffusion from scratch. Our experimental intervention is here: <https://zenodo.org/records/15767655>. Our research questions were the following:

- RQ1** Do students given an FSML-based intervention demonstrate deeper conceptual understanding of related ML principles compared to students given a traditional library-based ML intervention?
- RQ2** How do students’ cognitive mechanisms, including cognitive load, and observed learning differ under the FSML approach when compared to traditional instruction?

Our hypothesis was that students would experience a better conceptual understanding of the material after exposure to the FSML-intervention. We also hypothesized that, compared to peers receiving library-based instruction, FSML students would experience lower cognitive load and improved cognitive factors (e.g., greater interest, increased self-efficacy, and stronger motivation to continue exploring advanced ML topics).

## 2 Background on Cognitive Load Theory

Cognitive Load Theory (CLT) is an influential educational theory on how information should be presented to help people learn [24, 33, 36]. CLT posits that working memory can only process a small amount of information at once; consequently, when instructional materials exceed this capacity, the ability to understand and retain new ideas deteriorates [25, 36]. CLT has had substantial influence on the design of teaching strategies, emphasizing how content should be structured and presented so that learners can make the most of their cognitive resources [16, 30].

Apart from the working memory constraints, a central concept within CLT is that information can be mentally organized into “schemas,” which bundle multiple elements into a single chunk [34]. As learners’ understanding deepens, these schemas reduce the load on working memory by treating interconnected information as a single unit. The theory originally categorized cognitive load into three distinct types: (1) intrinsic load, reflecting the inherent complexity of the material itself; (2) extraneous load, resulting from presentation or design choices that do not support learning goals; and (3) germane load, relating to the mental effort that fosters the formation and refinement of schemas [4, 35, 37]. For instance, if learners must juggle unneeded details or process redundant information, their extraneous load can rise, diverting resources away from productive schema building. By contrast, materials that scaffold content in a logical, minimally distracting way help learners

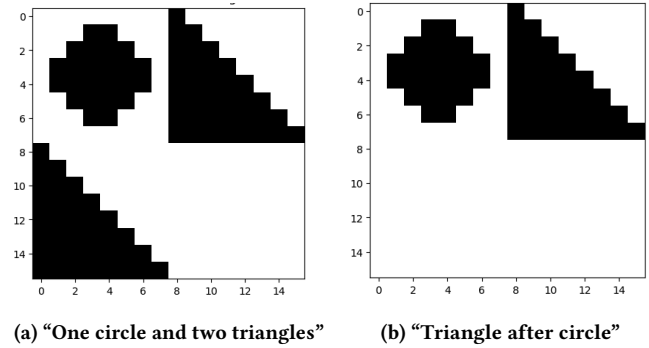


Figure 1: Examples of text-image pairs used for training.

focus on the essential aspects of what they need to master, thereby supporting germane load.

In response to critical views on how the components of cognitive load should be handled, Sweller has proposed that germane load might not be a separate contributor to overall cognitive load, but instead redirects extraneous cognitive resources to support schema formation [36]. Duran et al. [10], in their review of cognitive load theory in computing education, label the older view, which treated germane load as its own distinct additive factor, as “old CLT,” and the modified perspective—where germane load is understood more as a reallocation of resources rather than an independent contributor as the “new CLT.” They note that the “new CLT” in many ways returns to the essence of Sweller’s earliest ideas from the 1980s by simplifying the model and reducing conceptual overlap between load types.

## 3 Approach

We designed a tutorial that teaches how to implement a transformer and a stable diffusion model from scratch, using our full-stack machine learning approach (FSML). We designed the tutorial using a hierarchical, bottom-up instructional approach informed by cognitive load theory of schema acquisition. Given the large number of complex components in these two architectures and to avoid a high intrinsic cognitive load, we decomposed each architecture into its fundamental components, such as attention mechanisms, positional encodings, diffusion models, and UNets. We then decomposed each component further into subcomponents and created progressive schemas so that students would build upon what they previously learned when studying a new component. We also introduced these components in a scaffolded progression. We start with the simplest and most interpretable components and then gradually add more parts to them to reach more complex components.

The tutorial comes with code examples illustrating how to implement everything from scratch in PyTorch. To avoid requiring GPU resources, we restrict the training set to a very simple fragment of English and very simple images describing basic geometric figures. Figure 1 illustrates some text-image pairs. We emphasize that the same implementation could be extended by just adding more layers in the transformers and the diffusion model. This would allow the student to train models capable of understanding more complex English sentences and producing richer images.

We then structured the tutorial in three parts. We first explain how to implement attention mechanisms, then how to implement a transformer, and we conclude with a presentation of stable diffusion. An overview of all the components covered in the tutorial is in Figure 2.

### 3.1 Attention mechanisms

A transformer [40] consists of an encoder and a decoder that embed text into vectors and vectors back into text, respectively. The encoder and the decoder work by creating a vector embedding for each token in the text, adding positional encodings to it, and then passing it through a sequence of attention mechanisms. Figure 2 contains an overview of the transformer architecture.

To reduce the cognitive load behind attention mechanisms, we organized them in a hierarchy, where we first implement a basic attention mechanism and then gradually add components to it until we obtain all attention mechanisms that compose a transformer.

### 3.2 Transformers

Once we presented all attention mechanisms, the tutorial demonstrates how to implement a transformer. We provide step-by-step instructions on how to implement the positional encodings, which are necessary to give the embeddings information on their location in a sentence. We again start from basic components and show how to aggregate them to arrive at the matrix of positional encodings.

### 3.3 Stable diffusion

After transformers, we move to diffusion processes. By following a bottom-up approach, we manage to deliver an instructional approach that does not require the usual mathematical framework for diffusion processes based on stochastic differential equations, Langevin dynamics [14], or the heat equation. We start with a basic diffusion process trained on the two-dimensional Swiss roll [26]. Afterwards, we give an overview of a more complex diffusion model where instead of a feed-forward network in the reverse process, we use a UNet [29]. In the middle of the UNet, we introduce an image-to-text cross attention mechanism, which is needed for conditioned text-to-image generation. This is one of the most complex attention mechanisms, but at this point of the tutorial, thanks to our bottom-up approach, students build upon their experience implementing the attention mechanisms of a transformer.

Once all these components have been implemented, we conclude by showing students how the components can be put together to produce a model that takes sentences from our English fragment and outputs images illustrating those sentences.

### 3.4 Control tutorial

We compared our tutorial with a tutorial produced by Wang [41] from Harvard. To our knowledge, this is the most comprehensive tutorial on stable diffusion that teaches the basics of transformers and stable diffusion. We give an overview of the main differences between the two tutorials below.

*Diffusion models.* Their tutorial provides a thorough mathematical background on diffusion models. Our tutorial only explains how diffusion models learn to reverse a transformation of samples

from an arbitrary distribution to a standard Gaussian distribution, reducing extraneous load.

*Attention mechanisms.* Their tutorial provides one coding exercise where students learn to implement from scratch a masked and a cross-attention mechanism, which are the most complex attention mechanisms. We instead provide a scaffolded sequence of coding exercises, where students implement a sequence of gradually more complex attention mechanisms.

*Flow diagrams for attention mechanisms.* We present diagrams illustrating the different components and operations of the attention mechanism and show how they are extended to cross-attention mechanisms. The diagrams illustrate each of the steps required in a code implementation of an attention mechanism.

## 4 Method

After receiving ethics approval from our institution, we conducted a study in a graduate-level course on advanced machine learning in which the students were instructed with the experimental tutorial and the control tutorial described in the previous section. The participants were expected to have some prior background in analysis, statistics and numerical methods, as a prerequisite for registering in the course. The experiment was structured so that each student was instructed using both tutorials but in random order, with the learning outcomes measured with an *in itinere* assessment that took place after they had completed one tutorial but before they began the second one (see Figure 3). Of the 443 students in the class, 222 and 221 students were randomly assigned to experimental and control conditions, respectively, which refers to which tutorial they followed prior to taking the quiz. Our goal was to use the results to measure which tutorial induced better learning outcomes.

After the students had finished both tutorials, they were given the option to take the same assessment again. The purpose of this was not to gather additional data for our study, but to ensure that the course was fair to everyone regardless of which experimental condition they were assigned. If a student chose to solve the assessment twice, then whichever grade was higher contributed 5% towards a student's final grade in the course.

After both tutorials were finished, we invited our study participants to complete a survey about their cognitive mechanisms pertaining to each intervention. The purpose of this was to gain a better understanding of which factors could explain any group difference in the quiz results. We presented students with eight pairs of statements as given in Sinha et al.'s work [32] on cognitive load, state curiosity, knowledge-gap awareness, and cognitive dissonance. These questions were in turn collected from Leppink et al. [17], Naylor [23], Glogger et al. [11] and Levin et al. [18], respectively. Like in Sinha et al.'s work, the statements were presented together with a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". To limit survey fatigue, we did not administer all five statements for each scale; instead, we chose one or two statements that were the most representative ones.

All students were given an information sheet about the study and invited to participate. No financial or grade-based compensation was offered. To participate, students could fill out a checkbox to that effect on the initial assessment that they were given. In accordance with the instructions that we were given from our institution's

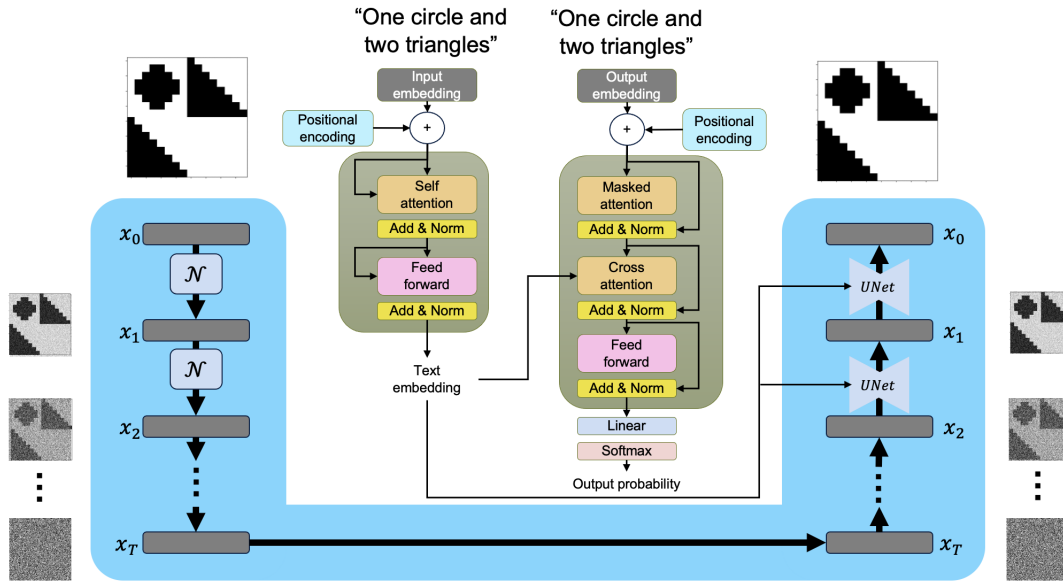


Figure 2: Architecture of a transformer and a stable diffusion.

ethics board, the course teacher was not informed which students chose to participate in the study so as not to influence their decision to do so.

We used Student’s t-tests to evaluate our hypotheses in response to RQ1 and RQ2. Specifically, we employed a two-sample t-test for the quiz scores to compare the performance between an experimental and a control group, and for the survey results, we used a paired t-test. Following established statistical recommendations [27, 38], we did not perform tests of normality, as t-tests are known to be robust against moderate normality deviations, and such tests can produce misleading outcomes if used solely to justify subsequent significance testing.

We conducted power analysis to determine the required sample size for both the between-subjects quiz-score comparison and the within-subjects survey comparison, using  $\alpha = 0.05$  and 80% power under a one-sided hypothesis. One-sided hypothesis testing was chosen since a difference in the opposite direction (lower scores) would have the same practical consequence as no difference at all—namely, continued use of the current curriculum—so only superiority is of interest.<sup>1</sup> Assuming a medium effect size ( $d = 0.5$ ), we obtained a requirement of 51 participants per group. For the survey, which contains paired data where each participant serves as their own control, the necessary sample size using the same parameters is 27 participants.

### 4.1 Instructional conditions

Students were given two lectures over two weeks on the theory of backpropagation, large language models and diffusion models by the main instructor. Each lecture lasted 3 hours and the content was the same for all students. In each week, there was also a two-hour

tutorial given by a teaching assistant. The tutorial was delivered in the form of recorded videos. Half of the students were given the control tutorial and the other half the experimental tutorial. In addition, the students were given code examples for each tutorial. One week after the last tutorial, a one-hour quiz was administered, which evaluated their skills in implementing different components of transformers and stable diffusion. The quiz contributed 5% of the final grade of the course. To further incentivize participation in the quiz, the instructor announced that one of the questions in the final exam would be copied from the quiz. Our quiz is available online [2].

### 4.2 Quiz

We designed a quiz that would test the students’ understanding of transformers and stable diffusion at different levels. The quiz serves the following objectives. First, we want to clearly discriminate between those who understand the low-level implementation details behind transformers and stable diffusion and how the different components of these models work together to perform text-to-image translation. Second, we want to verify that students achieve conceptual understanding, by evaluating if they understand *why* components like attention mechanisms, positional encodings, and diffusion models achieve their functionalities in their pipelines. Third, we want to verify that students achieve a procedural understanding by being able to reproduce parts of the implementation of these components that are crucial for their goals.

To verify conceptual understanding, we add questions where we include diagrams of the transformer and stable diffusion architectures, but with some crucial components intentionally removed. The students must then identify what the missing components are. We also include questions where students must apply what they

<sup>1</sup>We follow the rule-of-thumb provided by Bland and Altman on deciding on a one-sided test: “In general a one sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all” [6].

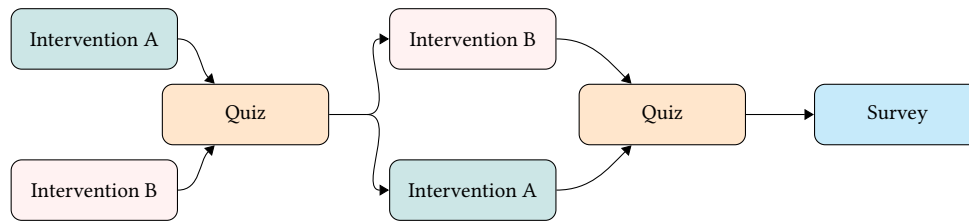


Figure 3: The research design that we used for our study.

learned in new areas, for example, we show them an incomplete diagram of the vision transformer architecture [9] and ask questions where participants must identify missing components.

To assess procedural understanding, we presented the code implementation of the main components of transformers and stable diffusion and asked students to complete them via multiple-choice questions. The goal here is to verify that they are capable of implementing the main parts of attention mechanisms, diffusion models, and cross-attention mechanisms.

*Quiz validation.* A research assistant independent of the authors was given the quiz together with the control tutorial and was asked to verify that answers to all quiz questions could be found in the control tutorial. In addition, he was asked to solve the quiz after seeing the tutorial material. He completed the quiz within one hour. The research assistant works in a different department from the authors and was not involved in the study in any way.

*Structure of the quiz.* The quiz is divided into sections as follows:

Topic	# Questions
Transformer architecture	3
Attention mechanism implementation	4
Diffusion models	2
Cross-attention	7
UNets	2
Stable diffusion and vision transformer	8
Diffusion models for bitstrings	2

## 5 Results

### 5.1 Learning outcomes

Of the total of 443 students in the course, 219 chose to take the initial graded assessment, of which 208 students consented to release their data. According to a free-form question on age and gender, 22 of the participants were women and 151 were men, and 35 chose not to answer. The average age of the participants was 24.0. 91 students were from the experimental group (44%) and 117 (56%) from the control group, exceeding the number we expected to need according to the power analysis (51 in each group). To assess whether the deviation from the expected balance between the two groups might reflect random variation in a finite population, we conducted a hypergeometric test ( $[N, K, n, x] = [443, 221, 208, 117]$ ). The result ( $p = 0.013 < 0.05$ ) indicated that this imbalance was unlikely to have arisen by chance alone.

The students from the experimental group and control group achieved an average grade of 20.31 (SD = 4.99) and 18.65 (SD = 4.94), respectively, out of a maximum score of 28. A one-sided Student’s  $t$ -test indicated a significant advantage for the experimental condition over the control:  $t(206) = 2.54, p = 0.006$ , an effect that would also remain significant with a two-tailed test ( $p = 0.012$ ). The coefficient of variation was 0.25 and 0.27, indicating moderate (between 0.1 and 0.3) variability of the test scores. Cohen’s  $d$  was 0.355 (SE: 0.142, with a 95% CI of  $[0.123, \infty]$ ), suggesting a small-to-moderate effect size according to Cohen’s benchmarks [8]. We offer a 95% confidence interval for the effect size, noting that when a one-tailed  $t$ -test is used, the corresponding 95% confidence interval for  $d$  is one-sided; its open end is therefore unbounded [7] and written as  $\pm\infty$  in the statistical software we used (JASP [20]).

We also took a look at the outcome of the individual questions on the assessment to get a more granular understanding of the student performance. Our data indicates that students in the experimental condition gained a better understanding of the technical implementation details about stable diffusion and transformers, and also positional encodings. The outcome was similar on the architectural concepts about transformers, stable diffusion, and vision transformers, but the experimental group performed worse at explaining diffusion processes.

### 5.2 Cognitive factors

62 study participants chose to answer our follow-up survey on their perceptions of the two tutorials, again exceeding the expected number we needed according to our power analysis. To answer RQ2, we conducted a Student’s paired-samples  $t$ -test. Our results can be found in Table 1. We found that across all measures, our experimental intervention outperformed the control intervention ( $p < 0.05$ ) with small-to-moderate effect sizes (Cohen’s  $d \in [0.23, 0.41]$ ), accounting for the two negative statements (6 and 7). We controlled the false-discovery rate of all nine confirmatory tests (quiz and cognitive measures) at 0.05 with the Benjamini–Hochberg procedure [5]. All remained significant after adjustment (including the quiz-score comparison ( $p = .006, q = .018$ )).

## 6 Discussion

Our first research question (RQ1) asked whether students taught with the FSML intervention would exhibit deeper conceptual understanding of machine learning principles than those taught under a more traditional, library-based approach. Our experimental data support this hypothesis. Students who were taught using the FSML tutorial outperformed the control group on our specialized

Paired Measure	Mean (SD)		Paired Student's <i>t</i> -test						
	Exp.	Ctrl.	<i>p</i>	<i>q</i>	<i>d</i>	SE	95% CI <sub>L</sub>	95% CI <sub>U</sub>	
(1) Germane load (1)	3.61 (1.06)	3.18 (1.00)	.007	.018	0.320	0.172	0.104	∞	
(2) Germane load (2)	3.53 (0.94)	3.26 (1.01)	.028	.037	0.248	0.146	0.035	∞	
(3) Low extraneous load	3.16 (1.20)	2.90 (0.97)	.033	.037	0.238	0.126	0.025	∞	
(4) State curiosity	3.36 (1.24)	3.00 (1.15)	.037	.037	0.231	0.166	0.018	∞	
(5) ML state curiosity	3.66 (1.09)	3.37 (1.15)	.029	.037	0.246	0.136	0.033	∞	
(6) Knowledge-gap awareness (1)	2.71 (1.05)	3.15 (0.77)	.001	.009	-0.409	0.152	-∞	-0.190	
(7) Knowledge-gap awareness (2)	2.69 (0.92)	3.00 (0.79)	.009	.018	-0.311	0.149	-∞	-0.096	
(8) Cognitive dissonance	3.02 (1.09)	2.71 (0.95)	.010	.018	0.306	0.126	0.091	∞	

**Table 1: Paired measures of participants' cognitive mechanisms for the experimental (Exp.) and control (Ctrl.) conditions. *p* is the unadjusted *p*-value of the paired *t*-test. *q* is the Benjamini–Hochberg FDR-adjusted value (false discovery rate) computed over all nine confirmatory tests (quiz score + eight cognitive measures,  $\alpha = 0.05$ ). *d* is Cohen's effect size; SE its standard error; CI<sub>L</sub> and CI<sub>U</sub> the bounds of its 95% confidence interval.**

assessment of transformer and diffusion-model fundamentals. This difference was statistically significant and corresponded to a small-to-moderate effect size (Cohen's  $d=0.355$ ). Notably, those in the FSML group excelled particularly in questions targeting low-level understanding of attention mechanisms and positional encodings. This suggests that building these components from scratch helps students form more coherent mental schemas compared to following a top-down, library-driven approach.

Our second research question (RQ2) asked how the FSML approach might affect key cognitive variables such as germane cognitive load, extraneous cognitive load, and affective responses (such as curiosity). Here again, our findings favor the FSML intervention. Survey results revealed that participants experienced higher germane load, lower extraneous load, and greater curiosity when implementing models from scratch, consistent with the tenets of cognitive load theory. These outcomes align with our intent that FSML's incremental, bottom-up design reduces distracting details (thereby lowering extraneous load) and fosters deeper schema acquisition (evidenced by higher germane load). Additionally, participants reported higher interest and lower knowledge-gap awareness when working with the FSML material, implying that they felt sufficiently equipped to tackle the complexities of model-building and thus maintained a more positive affective state.

Taken together, both research questions yielded results that support the effectiveness of the FSML approach. Our intervention appears to cultivate stronger conceptual mastery and more constructive cognitive and emotional engagement, relative to a more conventional, library-based approach.

## 7 Threats to Validity

There are some threats to the internal validity of the study. Although participants were randomly assigned to each intervention, the resulting sample sizes were uneven. A hypergeometric test suggested that this imbalance was unlikely to be solely due to chance. The imbalance itself may slightly skew observed differences between the two groups. For example, if the smaller group had students who were more prepared or more motivated, that alone could inflate observed advantages. Second, the FSML approach could have

benefited from a novelty effect, as students may have found our approach more exciting simply because it was new or different from past assignments. Confirming that these gains persist beyond initial exposure would require follow-up studies.

Construct validity threats: Our quiz may not capture the full range of competencies relevant to advanced machine learning. Ideally, our study would have made use of a validated, external instrument to measure conceptual understanding, similarly to how we could find external questions to assess the cognitive constructs. Furthermore, students' performance on specific quiz items could reflect short-term recall, or test-taking strategies rather than genuine conceptual depth. While item-level analysis suggests that many of the questions were suitably challenging and discriminative, further psychometric validation could support claims that the quiz truly captures conceptual understanding.

External validity threats: The study was conducted in one university course with students who had a fairly uniform background (meeting the same prerequisites in math and ML programming). Results may not generalize to different populations, such as non-traditional students or in online learning environments. Additionally, variations in instructor style or departmental culture could moderate the effectiveness of the FSML approach elsewhere.

## 8 Conclusion

In this paper, we introduced a new approach to machine learning education that we call full-stack machine learning (FSML). The approach involves teaching students to construct large language models and diffusion models from scratch using principles from cognitive load theory. After conducting a randomized controlled trial in a machine learning class with 208 students, we found that students who were exposed to the FSML approach performed better on a test measuring their conceptual knowledge, and experienced better outcomes on a test measuring cognitive factors than students who were assigned a traditional, library-based instructional method. While our results have implications for advanced machine learning instruction, our approach could also be extended to other advanced computing topics, and future empirical studies on the topic would be a valuable focus for future research.

## References

- [1] 2024. *Challenges and Opportunities for Computer Science*. Association for Computing Machinery, New York, NY, USA.
- [2] 2025. Quiz on transformers and stable diffusion. doi:10.5281/zenodo.15053918
- [3] David Paul Ausubel. 2012. *The acquisition and retention of knowledge: A cognitive view*. Springer Science & Business Media.
- [4] Jens F Beckmann. 2010. Taming a beast of burden—On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and instruction* 20, 3 (2010), 250–264.
- [5] Yoav Benjamini and Yoel Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.
- [6] J. Martin Bland and Douglas G. Altman. 1994. Statistics Notes: One and two sided tests of significance. *Bmj* 309, 6949 (1994), 248.
- [7] George Casella and Roger Berger. 2024. Statistical Inference. In *Statistical Inference* (2nd ed.). CRC Press, Chapter 9.
- [8] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* (2020).
- [10] Rodrigo Duran, Albina Zavgorodniaia, and Juha Sorva. 2022. Cognitive load theory in computing education research: A review. *ACM Transactions on Computing Education (TOCE)* 22, 4 (2022), 1–27.
- [11] Inga Glogger-Frey, Katharina Gaus, and Alexander Renkl. 2017. Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction* 51 (2017), 26–35.
- [12] Suzanne Hidi and K Ann Renninger. 2006. The four-phase model of interest development. *Educational psychologist* 41, 2 (2006), 111–127.
- [13] James Hiebert. 2007. THE EFFECTS OF CLASSROOM MATHEMATICS TEACHING. *Second handbook of research on mathematics teaching and learning: A project of the national council of teachers of mathematics 1* (2007), 371.
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [15] Ling-Ying Huang. 2014. Learning to Read with the Whole Language Approach: The Teacher's View. *English Language Teaching* 7, 5 (2014), 71–77.
- [16] Paul A Kirschner, John Sweller, and Richard E Clark. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist* 41, 2 (2006), 75–86.
- [17] Jimmie Leppink, Fred Paas, Tamara Van Gog, Cees PM van Der Vleuten, and Jeroen JG Van Merriënboer. 2014. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and instruction* 30 (2014), 32–42.
- [18] Daniel T Levin, Caroline Harriott, Natalie A Paul, Tao Zhang, and Julie A Adams. 2013. Cognitive dissonance as a measure of reactions to human-robot interaction. *Journal of Human-Robot Interaction* 2, 3 (2013), 3–17.
- [19] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.
- [20] Jonathon Love, Ravi Selker, Maarten Marsman, Tahira Jamil, Damian Dropmann, Josine Verhagen, Alexander Ly, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. 2019. JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software* 88 (2019), 1–17.
- [21] Barbara Matson. 1996. Whole Language or Phonics? Teachers and Researchers Find the Middle Ground Most Fertile. The Great Reading Debate. *Harvard Education Letter* 12, 2 (1996), 1–5.
- [22] Koby Mike and Orit Hazzan. 2022. Machine learning for non-majors: A white box approach. *Statistics Education Research Journal* 21, 2 (2022), 10–10.
- [23] Frank D Naylor. 1981. A state-trait curiosity inventory. *Australian Psychologist* 16, 2 (1981), 172–183.
- [24] Elizabeth Owen and John Sweller. 1985. What do students learn while solving mathematics problems? *Journal of educational psychology* 77, 3 (1985), 272.
- [25] Fred GWC Paas and Jeroen JG Van Merriënboer. 1994. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review* 6 (1994), 351–371.
- [26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cour-napeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2024. sklearn.datasets.make\_swiss\_roll. [https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make\\_swiss\\_roll.html](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_swiss_roll.html). Accessed: 2025-03-11.
- [27] Dieter Rasch, Klaus D Kubinger, and Karl Moder. 2011. The two-sample t test: pre-testing its assumptions does not pay off. *Statistical papers* 52 (2011), 219–231.
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752* (2022).
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 234–241.
- [30] Wolfgang Schnotz and Christian Kürschner. 2007. A reconsideration of cognitive load theory. *Educational psychology review* 19 (2007), 469–508.
- [31] Alan H Schoenfeld. 2004. The math wars. *Educational policy* 18, 1 (2004), 253–286.
- [32] Tanmay Sinha, Manu Kapur, Robert West, Michele Catasta, Matthias Hauswirth, and Dragan Trninić. 2021. Differential benefits of explicit failure-driven and success-driven scaffolding in problem-solving prior to instruction. *Journal of Educational Psychology* 113, 3 (2021), 530.
- [33] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.
- [34] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.
- [35] John Sweller. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review* 22 (2010), 123–138.
- [36] John Sweller, Jeroen JG van Merriënboer, and Fred Paas. 2019. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review* 31 (2019), 261–292.
- [37] John Sweller, Jeroen JG Van Merriënboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* (1998), 251–296.
- [38] Thomas Douglas Victor Swinscow, Michael J Campbell, et al. 2002. *Statistics at square one*. Number Ed. 10. Bmj London.
- [39] Sverrick Thorgeirsson, Tracy Ewen, and Zhendong Su. 2025. What Can Computer Science Educators Learn From the Failures of Top-Down Pedagogy?. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (Pittsburgh, PA, USA) (SIGCSETS 2025). Association for Computing Machinery, New York, NY, USA, 1127–1133. doi:10.1145/3641554.3701873
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [41] Binxu Wang. 2022. Understanding Stable Diffusion from 'Scratch'. <https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch/materials/stable-diffusion-scratch>. Accessed: 2025-03-11.