Teaching Gradient Descent Through Analogies, Step by Step

Evaluating and using analogies to teach concepts in Machine Learning to Computer Science students

Thomas Koppelaar[1]

Supervisor(s): Gosia Migut [1], **Ilinca Rentea** [1], **Yuri Noviello** [1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Thomas Koppelaar
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Rentea, Yuri Noviello, David Tax

Abstract

Machine Learning is becoming a standard part of Computer Science curriculums at universities. This paper aims to contribute to the education of Machine Learning in Computer Science, specifically through teaching concepts related to Gradient Descent (GD) through analogies. First, concepts related to Gradient Descent were collected through the use of academic textbooks, and analogies were created based on the definitions found. These analogies were then evaluated by experts, scoring the analogies on Target Concept Coverage, Mapping Strength, and Metaphoricity. The analogies that scored highest on a mean average were then used in an A/B survey distributed amongst Computer Science students that had not followed any Machine Learning course. One group was given the concept definitions, the other both the definitions and the analogies. The learning proficiency was measured, and no statistically significant result was found. In the end, this research explores the possibilities of creating analogies to explain machine learning concepts, and provides a modular framework for evaluating quality and measuring effectiveness of analogies.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are applied in many important sectors of the world, such as healthcare, finance, transportation and education [1]. Machine Learning is being taught in more and more lecture halls every year. As [2] shows, it is a core part of many Computer Science curriculums at universities. With the increase of interest in and importance of the subject, it is critical to take a look at the pedagogical methods that are used to convey abstract knowledge to students. Furthermore, [1] states that "Establishing ethical guidelines and frameworks is essential to prevent misuse and ensure that AI is used responsibly". Education in ML, amongst other things, "will play a vital role in promoting ethical practices and informed decision-making" [3]. It is clear that education in Machine Learning is a research topic worth investigating. [1]

In a study published in 2020, [4] defines Notional Machines (NMs) and looks at their use in Computing Education. They present three categories of NMs: Machine-generated representations, Hand- made representations, and Analogy. They also state that "NMs often relied on analogy to make salient and visible some aspect of the largely hidden underlying system" [4].

As ML is a large topic with many concepts to cover, a smaller domain of topics was chosen to be the focus of this research, namely Gradient Descent (GD). GD is present in both ML and Deep Learning books used in textbooks [5] [6].

This study investigates the following research question:
How does the use of analogies in explaining Gradient Descent affect the learning proficiency for Computer Science students?

To address this question, the following subquestions were defined:

- **SQ1: How do experts in Machine Learning evaluate different analogies?**
- **SQ2: What knowledge do Computer Science students gain from learning about Gradient Descent using analogies?**
- **SQ3: How do Computer Science students evaluate their engagement with the topic when using analogies to teach Machine Learning?**

The background section of this paper gives an overview of the relevant existing work this paper builds upon. The methodology involves the steps required to setup the experiment that can answer the research (sub)questions. The survey results are then analysed to assess differences in expert evaluation, as well as learning outcomes. The results are summarized and evaluated whether analogical instruction leads to improved comprehension of Gradient Descent. Ethics and Responsible Research are then discussed, to give an overview of ethical considerations that were taken into account for this research. Finally, the discussion reflects on the implications for ML education and suggests directions for future research, including refining analogy design and extending the approach to other ML concepts.

2 Background

2.1 Notional Machines

In a study from 1981, [7] states that "Novices should be introduced to programming through languages that embody simple notional machines with the facilities for making certain of the actions of the notional machine open to view". [4] gives a historic overview of various NMs found in literature [2]. Following that, [8] notes that NMs are made to help students understand a concept. They note that analogies "provide scaffolding to help refine the learner's mental model" [8], due to the student's mental model being incomplete or inaccurate.

2.2 Machine Learning Education

It is clear that NMs are used to teach concepts within Computing Education, but not much research exists on the topic of ML education [9], let alone the use of analogies. [10] later outlines two initiatives to teach ML to K-12 classes, where one "provides professional learning opportunities" which "empowers educators to deliver quality [Computer Science] education. Finally, [11] mentions the use of analogies in teaching ML concepts. Pendyala contributes analogies that explain ML concepts, also highlighting the similarity between some of the concepts and the real-world/simpler

1. Also stated by Amy J. Ko: "We need to learn how to teach Machine Learning". https://medium.com/bits-and-behavior/we-need-to-learn-how-to-teach-machine-learning-acc78bac3ff8. Accessed June 2025.

2. An overview of these notional machines is available at https://notionalmachines.github.io/analogies.html.

analogies. He notes that "this fundamental way of learning remains unexplored to a significant extent in human learning of difficult topics" [11]. The work itself does not contain empirical analysis, and hopes that it can inspire new research in this field. This research aims to answer that call by providing and evaluating analogies that can be used to explain core ML concepts, specifically related to Gradient Descent.

3 Methodology

This research aims to have three main contributions:

1. A set of analogies that can be used to explain concepts related to Gradient Descent.
2. An evaluation of these concepts done by experts.
3. An overview of the effects of using these analogies to teach concepts related to Gradient Descent, by measuring the learning proficiency for novice students.

3.1 Selecting Concepts and Creating analogies

In order to generate analogies, a list of concepts must first be chosen that will be explained to the student. For this, [6] and [5] were used to gather concepts and their definitions. Then, recalling that [4] mentioned that most NMs were handmade rather than Machine-generated, analogies were created per concept. An overview of all analogies can be found in Appendix A.

Category	Text
Concept Definition	Optimization refers to the task of either minimizing or maximizing some function f(x) by altering x. When we are minimizing it, we may also call it the cost function, loss function, or error function.
Analogy	Imagine you're in a radioactive zone. We're using a geiger counter (function) to measure the radiation in different spots, Optimization refers to the task of either looking for a safe zone (minimization), or looking for high spots of radiation (maximization). When we are looking for a safe spot, we are minimizing the radiation we measure on our geiger counter through measurements and calibration (loss function).

Table 1: An example concept definition, alongside an analogy.

Before continuing, it is good to recognize that analogies or metaphors have a multitude of definitions in literature, and are sometimes used interchangeably. This research builds on the work of [12], where we use the definition present to define analogies as "a description of an object or event, real or imagined, using concepts that cannot be applied to the object or event in a conventional way". In order to create a clear overview of how an analogy was formed, a concept map is created [13]. For the purposes of this research, we define a concept map to be a table that maps properties of an analogy to the properties of a concept. This was made in order to help experts with reviewing the analogies.

Concept property	Analogy counterpart
f(x) or function	Geiger counter
Spots in the radioactive zone	Datapoints
Error / loss / cost	Radiation
Minimization	Looking for a safe zone
Maximization	Looking for highest radiation levels

Table 2: An example concept map of the analogy for Optimization and Loss / Error / Cost function.

3.2 Expert Evaluation

Participant selection

The study done in this section of the research targeted experts in ML, which was minimally defined as having completed a course on ML. Experts were surveyed anonymously via an online form, and were asked to select their level of knowledge from the following options: Having passed a course on ML in a CS Bachelor, having Teaching Assistant experience for an ML course, having passed an ML course in a Masters program, or being a lecturer/professor on an ML course. This ensured that participants in the survey had minimal familiarity with the concepts, but allowed for differentiation between different levels of expertise and their ratings. Experts were given a random order of analogies to rank, and were asked to review as many as they had the time for. This means that not all analogies were rated an equal number of times, nor that analogies were reviewed by the same experts.

Criteria for evaluating analogies

In the survey, analogies were presented alongside the concept definition, as well as their concept mapping. Experts were asked to review analogies on three metrics. These metrics were originally introduced by [14] and later updated [3]. The criteria and their definitions are as follows:

- **Target concept coverage (TCC):** How well the analogy covers the topics in the description.
- **Mapping Strength (MS):** The logical soundness and consistency of the correspondence between source and target concepts.
- **Metaphoricity (M):** Conceptual distance between the source and the target concept.

The experts were asked to rate each analogy on a Three-point Likert scale, with the options being "1 - Low", "2 - Mid" and "3 - High" [4].

3. These definitions https://sites.google.com/illinois.edu/analogyeval24/analogy-evaluation-criteria.

4. This scale corresponds with the three-point scale presented on the website in the previous footnote.

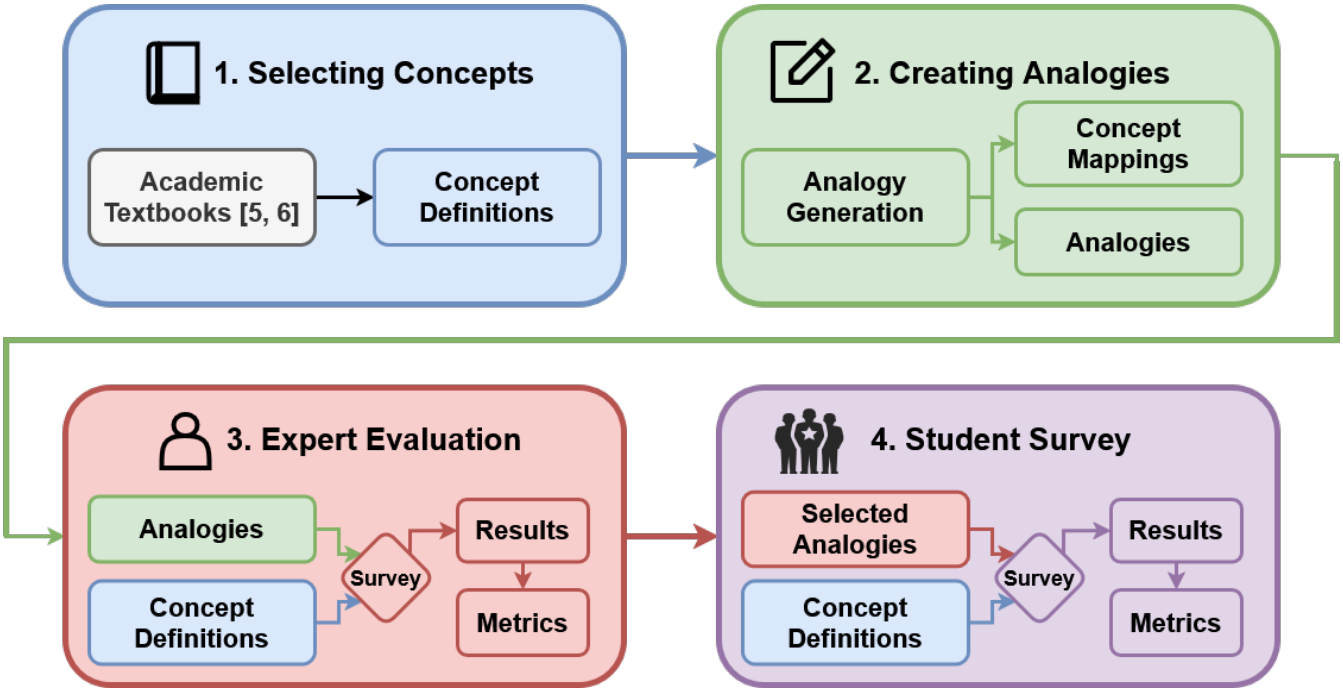

Figure 1: A full overview of the Methodology.

3.3 Student Survey

After the results from the expert evaluation are evaluated, a selection of analogies and concepts to be used in the student survey was made. The results of this selection as well as the metrics used are shown in Section 4.1.

Participant selection

In order to measure learning proficiency effectively, students that are novices in ML were selected to be the participants of the student survey. As [15] states, "students already functioning at a formal operational level may have an adequate understanding of the target and the inclusion of analogy might add unnecessary information". It follows from this that including students that have taken a course on ML would be detrimental to measuring the learning proficiency. Students were surveyed anonymously through an online form, through personal networks and chain-referrals.

Learning objectives

Bloom's Taxonomy [16] considers six categories in the cognitive process, where each category is considered more complex than the one below it. From this taxonomy, we look at two specific goals in the cognitive domain: Remembering and Understanding. These are the two lowest goals in the cognitive domain, and achieving these in a subject means that the student is able to recall concepts and their properties, as well as summarize and organize knowledge related to the concept. Due to the structure of the survey and the nature of analogies, the learning objectives were created to match the second category of Bloom's revised Taxonomy, Understand [16]. The third category in the taxonomy, Apply, is not applicable to the research question, as it sets the objective to apply or use the given knowledge in a given situation. The list of learning objectives can be found in the Appendix, section C.

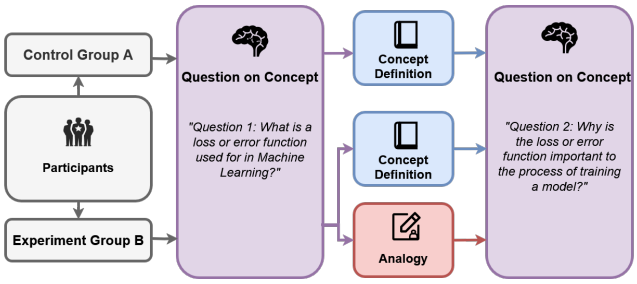

Figure 2: A diagram showcasing the setup for the first part of the student survey.

Survey structure

In order to measure the learning proficiency of the analogies used, some form of comparison must be made. This research employs an A/B survey: Participants were randomly put into either a control group (Group A) or an experiment group (Group B).

For each concept chosen to evaluate, a multiple-choice question was asked. Aside from thee possible answers, an option labelled "I don't know" was present. At the beginning of the survey, it was explained to the participant that this option was to be selected if the student could not explain for themselves why they would pick a certain answer. This was to dissuade participants from randomly guessing.

After the student picked an answer, the concept definition was shown. If the user was part of the experiment group, they would also be shown the relevant analogy. After this, the stu-

dent would be given another multiple-choice question, again with three possible answers and an option labelled "I don't know". This setup where each concept definition (and analogy, if the student was in the experiment group) was preceded and followed by a knowledge test was then repeated for every concept selected from the expert evaluation. The questions can be found in the Appendix, Section D.1.

Knowledge Gain and Engagement Evaluation

The knowledge gain is calculated by giving each answer a student gives either a 1 for a correct answer, and a 0 otherwise. If a student states that he does not know the answer, it is counted as incorrect. Then, the difference between the score of the question after the concept definition (and analogy, if applicable) and the score of the question before the concept definition is taken to be the knowledge gain for a student for that given topic.

The final part of the survey is done by both groups, and follows an adapted RIMMS structure [17]. The RIMMS is a 12-item questionnaire measuring self-reported Attention, Relevance, Confidence and Satisfaction. The adapted form that was chosen contained only 5 questions, in order to shorten the time required to fill in the survey. The idea was to make the survey more enticing to fill in, as it would take a student less time to participate in the survey. The questions can be found in the Appendix, Section D.2. They are general questions that ask if the student has enjoyed the learning process, if they felt like they were able to pay attention to the definitions and analogies, and if the analogies seemed helpful.

4 Results

4.1 Expert Evaluation

In total, 16 responses were gathered, of which 15 were participants claiming they had at least completed a course in ML. Thus only 15 responses were evaluated. The participant pool consisted of a professor in a course on ML, one student who had passed a Masters course on ML, three participants who had Teaching Assistance experience in an ML course and 10 students who had completed a bachelors course on ML. The data can be found in the Appendix, section B.

For each concept, an average score was calculated by taking the mean of each category an analogy was rated on, and then calculating the mean over the means. Furthermore, to determine the inter-rater-reliability, Krippendorf's Alpha [18] was calculated using an online tool called the K-Alpha Calculator [19]. Both values were rounded up to 3 decimal places.

From table 3 it is seen that the average score per concept is above 2, apart from A6. Furthermore, A1 had the highest average score of 2.667, followed by A3 with 2.389. A6 had an Average Score of 1.833. Looking carefully, A6 scored low on Mapping Strength and Target Concept Coverage, scoring 1.833 and 1.667 on average respectively. A1 was reviewed the most with 7 experts scoring the analogy, whilst the analogies for A2, A4 and A5 only got 3 reviews each. On average, each analogy was reviewed by 4.67 experts.

Calculating Krippendorf's Alpha returned lackluster results. An Alpha value of 1 indicates perfect agreement, whereas an Alpha value below 0.67 is an indication for poor agreement [19]. A negative value indicates systematic disagreement [19]. Gradient Descent had the highest Alpha value, simultaneously being the only value above 0.

Analogy	Average Score	K-Alpha
A1	**2.667**	-0.064
A2	2.222	-0.185
A3	**2.389**	**0.111**
A4	2.000	-0.111
A5	2.111	-0.233
A6	1.833	-0.064

Table 3: The average score and Krippendorf's Alpha (K-Alpha) for each analogy, given by the experts in the expert evaluation.

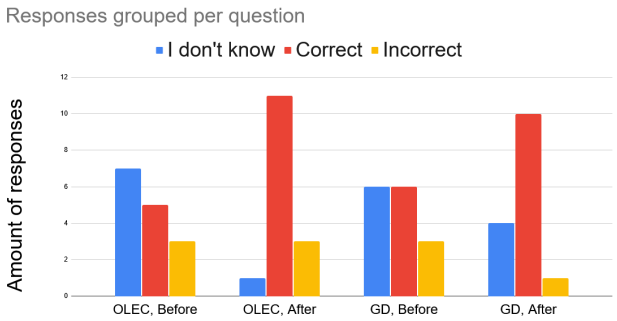

Figure 3: A bar chart with the responses from the student survey grouped per question. Responses from both groups are shown. OLEC = Optimization and Loss / Error / Cost function, GD = Gradient Descent.

Based on these results, two analogies were selected to be used in the next step of the research. The first concept that was chosen was A1, Optimization and Loss / Error / Cost function, as it had the highest average score. The second concept was A3, Gradient Descent, as it had a positive Krippendorf's Alpha, as well as the second-highest average score. The survey creation for the next step was previously described in Section 3.3.

4.2 Student Survey

In total, 15 students participated in the survey. Figure 3 shows the responses given per question. From initial observation, it is clear that overall, students performed better on the questions shown after the concept definition and analogy. With Optimization and Loss / Error / Cost function, 7 participants chose to say that they did not know the answer before being shown the concept definition. Afterwards, only one student filled this answer in again. It also should be noted that the students who got the second question incorrect, responded with "I don't know" or an incorrect answer on the first question. This means that the 7 students who answered correctly on the first question, also chose the correct answer on the second question for the questions related to optimization.

As for the responses on the questions related to Gradient Descent, there is a decrease in both incorrect answers and students stating that they do not know the answers, after they have been shown the concept definition and possibly the analogy. This time, 6 out 9 students who responded with "I don't know" or an incorrect answer, got the correct answer after being shown the concept definition. It is also interesting to note that out of the 4 students that stated that they did not know the answer to the second question, 2 students had correctly answered the first question related to Gradient Descent. Both of these students were in the control group, and were only shown the definition.

4.3 Knowledge Gain

Table 4 shows the average knowledge gain per concept for each group, as well as the results from performing an independent samples t-test. The results show no statistically significant differences between the control and the experiment group.

4.4 Engagement Evaluation

For unknown reasons, one participant had only filled in one of 5 questions present in the engagement evaluation at the end of the student survey. In order to gather more accurate insights, this submission was omitted from this part of the research. Regardless of this omission, no statistical significant difference was found in any category of engagement between the control and experiment group. On mean average, participants in the experiment group gave higher scores on Attention, Confidence and Satisfaction, whilst the control group gave higher scores on Relevance.

5 Discussion

5.1 Expert Evaluation

Figure [the one with krippendorfs alpha] immediately raises questions, especially when looking at the low or even negative values for Krippendorf's Alpha. This low agreement or systematic disagreement could be explained by various factors. Recall that on average each analogy was reviewed less than five times. This is a low sample size, meaning it is impossible to say what the agreement rate would look like if a larger number of participants had reviewed the analogies. Furthermore, each review was weighed equally to its peers, meaning that the participant who is a lecturer/professor in Machine Learning had an equal say in reviewing as someone who had passed their course. Furthermore, only a single round of reviews were done, without an expert reviewing the same analogy more than once. Overall, these results correlate with the findings done by He et al., where they state that their results "also highlight the subjective nature of the qualitative dimensions that characterize analogies" [20].

5.2 Student Survey

As shown in table 4, the results of the study showed no significant differences in learning proficiency between students who were shown analogies of concepts they were unfamiliar with, and those who only got a definition of a concept. Figure 3 shows that students who were exposed to the material performed better on the questions asked about the topics, which trivially correlates with our understanding of knowledge gain. It was not shown that the analogies used made significant impact. This may be due to the subjective nature of the task, as well as the method of teaching. [21] notes that actively engaging when using analogies to teach concepts is appreciated by students and leaves a stronger impression. A survey is an interface with an information stream that goes in one direction, thus lacking any form of interaction. This could explain the lack of difference, although further research with a higher participant is necessary to confirm this.

It should also be noted that all participants were first year students in Computer Science at the Technical University of Delft at the moment the survey was distributed. Considering the course has students from all over the world, a participant might have a different interpretation of what it means for something to be relevant, or for something to be satisfactory.

6 Responsible Research

This section discusses the ethics regarding the research.

6.1 Methodology and Reproducibility

This research has been conducted with the principles of Open Science in mind. The method of creating analogies has been explained in the methodology, and the analogies themselves can be found in the Appendix in Section A. As these analogies were created by hand and historically have mostly been made that way [4], it is trivial to create new analogies. For the expert evaluation, the survey setup alongside the metric definitions were given. From this, future researchers are able to recreate the survey as described in this paper. As the questions that were used for the student survey are present in Appendix D.1 and D.2, the student survey is reproducible. The procedure for selecting participants are present in section 3.2 and 3.3.

6.2 Data collection and privacy

All participants were asked voluntarily to participate in the surveys, and informed consent was required. No personally identifiable information (PII) was required in order to answer the research questions, therefore it was not collected in any survey. Participants in the expert evaluation were asked to select what level of expertise they had. However, as there was no PII collection, it is impossible for someone with access to the research data to link a submission to a person. The same goes for the student survey. The research is compliant with the Technical University of Delft's policies on data storage, and has been approved by the Human Resources and Ethics Committee of the faculty.

7 Conclusions and Future Work

This research serves as an exploration into the world of using analogies to teach ML concepts, as first started by [11], through the introduction of a framework through which analogies meant for teaching ML concepts can be reviewed and evaluated. Furthermore, analogies were created to teach concepts relevant to ML, by taking concepts specific to the

Analogy		μ	σ	t-test	p
Optimization	Control	0.375	0.518	-0.293	0.778
	Experiment	0.429	0.535		
Gradient Descent	Control	0.75	0.916	0.717	0.497
	Experiment	0.571	0.378		

Table 4: The mean knowledge gain per concept from the student survey, as well as the standard deviation. Values are rounded to 3 decimal places.

Category	Group	μ	σ	t-test	p
Attention	Control	1.625	1.708	-0.759	0.473
	Experiment	2.083	1.730		
Relevance	Control	3.25	1.165	0.202	0.846
	Experiment	3.167	1.602		
Confidence	Control	3.125	2.100	-0.505	0.629
	Experiment	3.5	1.517		
Satisfaction	Control	1.875	1.885	-0.438	0.674
	Experiment	2.167	1.329		

Table 5: The mean ARCS metrics from the responses of the student survey, as well as the standard deviation. Values are rounded to 3 decimal places.

topic of Gradient Descent [5]. Then, an expert review was done and highlighted the subjective nature of the task through a low value for Krippendorf's Alpha. Two of these analogies were then used in measuring learning proficiency for novice students, and no statistically significant result was found. In reviewing the self-reported engagement evaluation, no statistically significant result was found.

It is unclear whether or not the analogies created in this research improve the learning proficiency for explaining Gradient Descent for Computer Science students. To prove this, further research is required. For further research, it is recommended that a larger pool of participants is found for both the expert review, as well as the student survey. This could lead to producing statistically significant results, as a larger data set tends to produce less volatile results. As the student survey does not rely on the process that the expert evaluation takes, it is possible to solely focus on producing analogies that are highly rated by experts. There are two recommendations for this direction. First, the method of analogy generation could be researched more in-depth. As was done by [11], the analogies in this paper were created by hand. In a recent study, it was shown that AI chatbots can can outperform humans in creative thinking tasks [22]. This raises the question whether or not AI would be able to produce higher rated analogies than humans.

As briefly mentioned in the discussion, the analogies were only given one review per expert, and no suggestions for improvement were collected for the analogies. Research could be done to see if a general procedure for improving an analogy could be made, in order to provide a useful framework for teachers to create analogies for concepts that do not have them yet.

With regards to the student survey, the setup could be improved to mimic a traditional examination setup. This way, a more accurate method of determining whether or not a student has had an increase in learning proficiency could be employed. For this to be most effective, the research would ideally mimic the environment in which it is going to be used. In this case, that would be a classroom or an online course environment.

To summarize: This work is exploratory and aims to inspire others by providing a framework through which to evaluate analogies for use in education. With AI and Machine Learning becoming more and more prevalent in everyday life, it is important to recognize the ethical challenges engineers face today. Through the use of quality educational tools, we can empower a new generation of engineers to responsibly write a new chapter in the history of mankind.

References

[1] A. Mishra, "A comprehensive review of artificial intelligence and machine learning: Concepts, trends, and applications," *International Journal of Scientific Research in Science and Technology*, 2024.

[2] S. Mackay and A. Decker, "Computer science curriculum trends," *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, p. 1732–1733, Mar 2024.

[3] U. A. Usmani, A. Y. Usmani, and M. U. Usmani, "Ensuring trustworthy machine learning: Ethical foundations, robust algorithms, and responsible applications," *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 576–583, 2023.

[4] S. Fincher, J. Jeuring, C. S. Miller, P. Donaldson, B. du Boulay, M. Hauswirth, A. Hellas, F. Hermans, C. Lewis, A. Mühling, J. L. Pearce, and A. Petersen, "Notional machines in computing education: The education of attention," *Association for Computing Machinery*, p. 21–50, 2020.

5. These analogies, alongside other analogies created by researchers working in parallel to this project, will be made available through https://ml-teaching-analogies.github.io/.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[6] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2016.

[7] B. du Boulay, T. O'Shea, and J. Monk, "The black box inside the glass box: presenting computing concepts to novices," *International Journal of Man-Machine Studies*, vol. 14, no. 3, pp. 237–249, 1981.

[8] B. Munasinghe, T. Bell, and A. Robins, "Computational thinking and notional machines: The missing link," *ACM Trans. Comput. Educ.*, vol. 23, Dec. 2023.

[9] R. B. Shapiro and R. Fiebrink, "Introduction to the special section: Launching an agenda for research on learning machine learning," *ACM Trans. Comput. Educ.*, vol. 19, Oct. 2019.

[10] C. L. Pineda, A. A. K. Ashar, and J. Liu, "Fostering ai literacy: A survey of student perceptions and effective practices in k-12 machine learning," *2024 IEEE Frontiers in Education Conference (FIE)*, pp. 1–7, 2024.

[11] V. S. Pendyala, "Relating machine learning to the real-world: Analogies to enhance learning comprehension," *Springer International Publishing*, pp. 127–139, 2022.

[12] B. Indurkhya, *Characterizing Metaphor*. Dordrecht: Springer Netherlands, 1992.

[13] G. R. Watson, "What is... concept mapping?," *Medical Teacher*, vol. 11, p. 265–269, Jan 1989.

[14] B. Bhavya, C. Palaguachi, Y. Zhou, S. Bhat, and C. Zhai, "Long-form analogy evaluation challenge," in *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges* (S. Mille and M.-A. Clinciu, eds.), (Tokyo, Japan), pp. 1–16, Association for Computational Linguistics, Sept. 2024.

[15] D. F. Treagust, "The evolution of an approach for using analogies in teaching and learning science," *Research in Science Education*, vol. 23, pp. 293–301, 1993.

[16] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory Into Practice*, vol. 41, no. 4, p. 212–218, 2002.

[17] N. Loorbach, O. Peters, J. Karreman, and M. Steehouder, "Validation of the instructional materials motivation survey (imms) in a self-directed instructional setting aimed at working with technology," *British Journal of Educational Technology*, vol. 46, no. 1, pp. 204–218, 2015.

[18] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures*, vol. 1, no. 1, p. 77–89, 2007.

[19] G. Marzi, M. Balzano, and D. Marchiori, "K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient," *MethodsX*, vol. 12, p. 102545, 2024.

[20] G. He, A. Balayn, S. Buijsman, J. Yang, and U. Gadiraju, "It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 10, p. 89–101, Oct. 2022.

[21] S. Petchey, D. Treagust, and K. Niebert, "Improving university life science instruction with analogies: Insights from a course for graduate teaching assistants," *CBE—Life Sciences Education*, vol. 22, June 2023.

[22] M. Koivisto and S. Grassini, "Best humans still outperform artificial intelligence in a creative divergent thinking task," *Scientific Reports*, vol. 13, no. 1, 2023.

A Overview of concepts and analogies

See table 6.

B Expert Evaluation Results

See table 7.

C Learning Objectives for the Student Survey

- **Optimization and Loss / Error / Cost function:** The student is able to determine the importance of the use of a loss / error function in Machine Learning.
- **Gradient Descent:** The student can reason about the role of the gradient in Gradient Descent, and what happens when moving in the negative direction of the gradient.

D Student Survey Questions

D.1 Multiple-choice questions

Correct answers are emphasised.

1. What is a loss or error function used for in Machine Learning?
   - **A - Measuring the performance of a model.**
   - B - Decreasing the complexity of the data to better fit the model.
   - C - Decreasing the complexity of a model to perform better on the given data.
   - D - I don't know.
2. Why is the loss or error function important to the process of training a model?
   - A - It provides a method of quantifying model complexity.
   - **B - It defines the objective that the model seeks to optimize.**
   - C - It defines a function to optimize the usage of the model on given data.
   - D - I don't know.
3. What role does the derivative play in the Gradient Descent algorithm?
   - A - It identifies the size of a step the algorithm can take in order to determine a new position.
   - **B - It determines the direction in which the algorithm can move in order to determine a new position.**
   - C - It determines the loss or error of the current position compared to the new position.
   - D - I don't know.
4. Why does Gradient Descent update model parameters in the negative direction of the gradient?
   - A - A positive gradient represents a larger loss or error.
   - B - The negative of the gradient points towards the maximum of the loss function.
   - **C - Moving towards the negative of the gradient reduces the loss.**
   - D - I don't know.

D.2 Engagement Evaluation Questions

Categories the questions belong to are parenthesised and were not shown to the participant. All questions had 5 possible answers: 1. Not true, 2. Slightly true, 3. Moderately true, 4. Mostly true, 5. Very true.

1. The quality of the text helped to hold my attention. (Attention)
2. The variety of reading passages, exercises, illustrations, etc., helped keep my attention on the questions. (Attention)
3. The content of these questions will be useful to me. (Relevance)
4. As I worked with these questions, I was confident that I could learn how Gradient Descent works in Machine Learning. (Confidence)
5. I enjoyed working with these questions so much that I was stimulated to keep on working. (Satisfaction)

E Student Survey Answers

E.1 Learning Proficiency Answers

See table 8.

E.2 Engagement Evaluation Answers

See table 9.

ID	Concept	Definition	Analogy
A1	Optimization & Loss / error / cost function	Optimization refers to the task of either minimizing or maximizing some function f(x) by altering x. When we are minimizing it, we may also call it the cost function, loss function, or error function.	Imagine you're in a radioactive zone. We're using a geiger counter (function) to measure the radiation in different spots, Optimization refers to the task of either looking for a safe zone (minimization), or looking for high spots of radiation (maximization). When we are looking for a safe spot, we are minimizing the radiation we measure on our geiger counter through measurements and calibration (loss function).
A2	Gradient	We often minimize functions that have multiple inputs: $f : R^n \rightarrow R$. For the concept of "minimization" to make sense, there must still be only one (scalar) output. For functions with multiple inputs, we must make use of the concept of partial derivatives. The partial derivative $f \frac{\partial}{\partial x_i}(x)$ measures how f changes as only the variable x_i increases at point x. The gradient generalizes the notion of derivative to the case where the derivative is with respect to a vector: the gradient of f is the vector containing all the partial derivaties, denoted $\nabla_x f(x)$. Element i of the gradient is the partial derivative of f with respect to x_i.	We often have multiple sources of radiation that affect our readings. In order to know if we're in a safe zone, we simply want to know our radiation level as a single number. In order to do this, we can measure how the radiation changes if we only walk in one direction. We can then summarize this into a single reading for our current location (the gradient).
A3	Gradient Descent	We can decrease f by moving in the direction of the negative gradient. This is known as the method of steepest descent, or gradient descent. Steepest descent proposes a new point $x' = x - \epsilon \nabla_x f(x)$. Where ϵ is the learning rate, a positive scalar determining the size of the step.	Gradient descent is like moving in the direction of a safer area, to a new spot. The new spot is chosen by walking in the direction of safety for some an amount of time that we decide beforehand. Once we arrive at a place with less radiation, we take a new measurement and decide on where to move next.
A4	Critical Points	When f'(x) = 0, the derivative provides no information about which direction to move. Points where f'(x) = 0 are known as critical points. A local minimum is a (critical) point where f(x) is lower than all neighboring points, so it is no longer possible to decrease f(x) by making infinitesimal steps. A local maximum is a point where f(x) is higher than all neighboring points. Some critical points are neither maxima nor minima. These are known as saddle points.	In some places, the direction we need to move in is unclear. There are three such uncertain cases: 1. Every step we take brings us closer to radiation, so it may be that we've found the optimal location to stay. 2. Every step we take brings us further from radiation. 3. Every step we take doesn't change our radiation levels.
A5	Batch Gradient Descent	Note that the error function is defined with respect to a training set, and so each step requires that the entire training set be processed in order to evaluate ∇ f(x). Techniques that use the whole data set at once are called batch methods.	Our geiger counter takes measurements from its surroundings in order to calculate a direction that we need to move to. Batch gradient descent is like using all of our surroundings to calculate what direction we need to go to. This, of course, means our calculations are tied to how large our surroundings are. If our area grows, so does the time it takes to calculate the radiation and the direction we need to go into.
A6	Stochastic Gradient Descent	The computational cost of calculating the gradient descent is O(m), where m is the training set size. As the training set size grows to billions of examples, the time to take a single gradient step becomes prohibitively long. The insight of SGD is that the gradient is an expectation. The expectation may be approximately estimated using a small set of samples. Specifically, on each step of the algorithm, we can sample a minibatch of examples drawn uniformly from the training set. The minibatch size m' is typically chosen to be a relatively small number of examples, ranging from one to a few hundred. Crucially, m' is usually held fixed as the training set size m grows. We may fit a training set with billions of examples using updates computed on only a hundred examples.	We measure radiation based on our surroundings, so every measurement requires us to sample our surroundings. If we have a huge area to check, this would take a long time to process. However, if we configured our geiger counter to only (randomly) sample a couple of spots, we could still get useful measurements, meaning our time taken to measure stays consistent regardless of the area size that we're walking through. This also means that we don't fully measure all directions that we can walk in, meaning we might end up in a location that doesn't give us any safe direction to move in, even though there may be an even safer place somewhere else.

Table 6: An overview of all concepts with their analogies.

	OLEC			G			GD			CP			BGD			SGD		
Knowledge level	TCC	MS	M	TCC	MS	M	TCC	MS	M	TCC	MS	M	TCC	MS	M	TCC	MS	M
Bachelor	3	3	2	2	3	2							3	2	2	3	2	2
TA	3	2	1															
Bachelor	2	3	3				2	1	2									
Bachelor	3	3	2							3	2	3						
Bachelor	3	3	2				3	2	2							3	2	2
Bachelor	3	3	3	3	3	2	3	2	3	1	2	2				2	2	2
Bachelor																1	2	2
Master										1	2	2						
Bachelor													2	2	1			
Bachelor																		
Bachelor				1	1	3	3	2	2							1	1	2
TA							3	3	3									
TA							2	2	3									
Lecturer/Professor	3	3	3										2	2	3	1	1	2
Bachelor																		

Table 7: The rankings given to concepts by experts. 1 = Low, 2 = Mid, 3 = High. TCC = Target Concept Coverage, MS = Mapping Strength, M = Metaphoricity. OLEC = Optimization & Loss / Error / Cost Function, G = Gradient, GD = Gradient Descent, CP = Critical Points, BGD = Batch Gradient Descent, SGD = Stochastic Gradient Descent.

Group	What is a loss or error function used for in Machine Learning?	Why is the loss or error function important to the process of training a model?	What role does the derivative play in the Gradient Descent algorithm?	Why does Gradient Descent update model parameters in the negative direction of the gradient?
A	I don't know.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	Moving towards the negative of the gradient reduces the loss.
A	Decreasing the complexity of the data to better fit the model.	It defines a function to optimize the usage of the model on given data.	It determines the loss or error of the current position compared to the new position.	Moving towards the negative of the gradient reduces the loss.
A	Decreasing the complexity of the data to better fit the model.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	I don't know.
A	Measuring the performance of a model.	It defines the objective that the model seeks to optimize.	It determines the loss or error of the current position compared to the new position.	Moving towards the negative of the gradient reduces the loss.
A	I don't know.	It defines a function to optimize the usage of the model on given data.	I don't know.	Moving towards the negative of the gradient reduces the loss.
A	I don't know.	It defines a function to optimize the usage of the model on given data.	It determines the loss or error of the current position compared to the new position.	Moving towards the negative of the gradient reduces the loss.
A	I don't know.	It defines the objective that the model seeks to optimize.		Moving towards the negative of the gradient reduces the loss.
A	Measuring the performance of a model.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	I don't know.
B	Measuring the performance of a model.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	Moving towards the negative of the gradient reduces the loss.
B	I don't know.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	Moving towards the negative of the gradient reduces the loss.
B	I don't know.	I don't know.	I don't know.	I don't know.
B	Decreasing the complexity of the data to better fit the model.	It defines the objective that the model seeks to optimize.	I don't know.	Moving towards the negative of the gradient reduces the loss.
B	Measuring the performance of a model.	It defines the objective that the model seeks to optimize.	I don't know.	A positive gradient represents a larger loss or error.
B	I don't know.	It defines the objective that the model seeks to optimize.	I don't know.	I don't know.
B	Measuring the performance of a model.	It defines the objective that the model seeks to optimize.	It determines the direction in which the algorithm can move in order to determine a new position.	Moving towards the negative of the gradient reduces the loss.

Table 8: Responses on the Learning Proficiency part of the student survey.

Group	The quality of the text helped to hold my attention.	The variety of reading passages, exercises, illustrations, etc., helped keep my attention on the questions.	The content of these questions will be useful to me.	As I worked with these questions, I was confident that I could learn how Gradient Descent works in Machine Learning.	I enjoyed working with these questions so much that I was stimulated to keep on working.
A	Not true	Slightly true	Moderately true	Mostly true	Moderately true
A	Mostly true	Mostly true	Mostly true	Mostly true	Mostly true
A	Slightly true	Not true	Moderately true	Slightly true	Slightly true
A	Slightly true	Not true	Moderately true	Very true	Slightly true
A	Mostly true	Slightly true	Mostly true	Very true	Slightly true
A	Moderately true	Moderately true	Moderately true	Not true	Not true
A	Mostly true	Not true	Very true	Very true	Very true
A	Not true	Not true	Slightly true	Slightly true	Not true
B	Not true	Not true	Not true	Moderately true	Moderately true
B	Slightly true	Slightly true	Mostly true	Mostly true	Moderately true
B	Very true	-	-	-	-
B	Mostly true	Slightly true	Moderately true	Moderately true	Slightly true
B	Mostly true	Not true	Mostly true	Very true	Slightly true
B	Moderately true	Mostly true	Mostly true	Slightly true	Slightly true
B	Mostly true	Moderately true	Mostly true	Very true	Mostly true

Table 9: Responses on the adapted RIMMS part of the student survey.