Proceedings of Machine Learning Research 273:191–201, 2025 iRAISE Workshop at AAAI 2025

Can LLMs Teach Human Learners to Understand Concepts Through Analogies?

Patrick Peixuan Ye PXYE@STANFORD.EDU and **Shirley Wu** SHIRWU@STANFORD.EDU
Department of Computer Science, Stanford University

Abstract

Large Language Models (LLMs) hold significant potential to revolutionize education by enabling personalized and effective learning experiences. As cognitive learning principles are gradually applied to designing educative LLMs, our research focuses on this crucial question: can LLMs enhance student comprehension of complex concepts through analogy-based tutoring, a pedagogical method proven useful in learning science? To address this, we propose a two-stage experimental framework. First, LLM tutors generate analogies for teaching specific target concepts, leveraging prompting techniques to adapt to simulated or real student profiles. Second, these learners engage with the analogies and subsequently complete multiple-choice question to evaluate their conceptual understanding. Our initial findings reveal that analogy-based tutoring enhances student engagement and conceptual mastery, achieving a notable improvement in comprehension. These results underscore the effectiveness of LLM-driven analogy-based tutoring in advancing educational outcomes and pave the way for future research in this domain.

Keywords: LLM-driven tutoring, Analogy-based learning, Personalized education

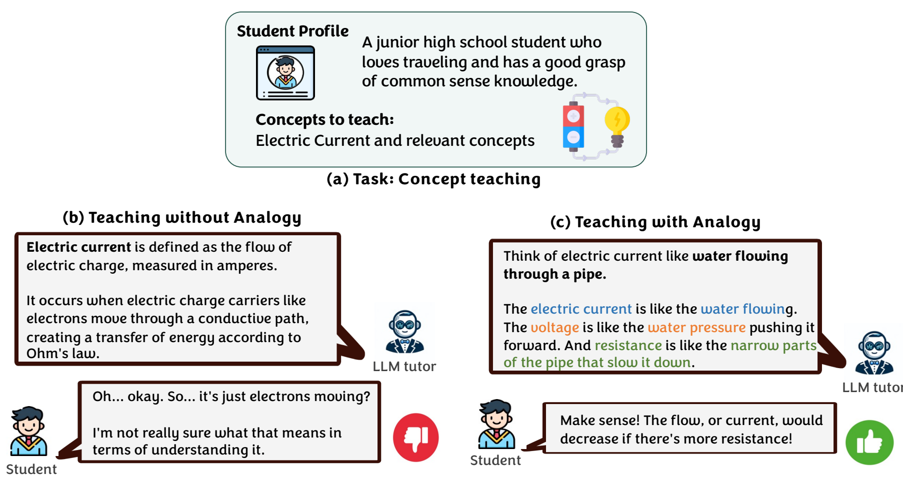

Figure 1: **(a) Task:** Effectively teach concepts to students with specific backgrounds with LLMs. **(b) w/o Analogy:** The LLM tutor's complicated explanations cause confusions. **(c) w/ Analogy:** The LLM facilitates effective understanding with profile-related analogies.

1. Introduction

Many educators excel at teaching complex concepts through analogies, a pedagogical device proven useful for simplifying abstract ideas (Newby et al., 1995; Glynn and Takahashi, 1998).

For example, students learning about electric current may struggle with its technical definition, but a water-flow analogy can make the concept relatable. This raises the question: *Can LLMs effectively generate such analogies to improve student comprehension?*

Large Language Models (LLMs) hold potential as automated tutors (Kasneci et al., 2023), but their capability to teach through analogy remains underexplored. Previous work has proposed the possibility of employing LLMs to generate educational analogies Bhavya et al. (2024), but the challenge remains in measuring the pedagogical impact of the generative pipeline. Other existing research focuses on general tutoring abilities (Webb et al., 2023), leaving a gap in understanding how well LLMs can create and adapt analogies. Additionally, evaluating their impact on learning outcomes requires isolating the influence of analogies from prior knowledge (Denny et al., 2024; Sharma et al., 2023).

We propose a two-stage framework to evaluate LLM analogy-based teaching. [1] First, LLMs generate customized analogies tailored to diverse student profiles. Second, comprehension is assessed through anonymized questions to isolate analogy-driven learning. Experiments show analogy-based teaching improves understanding by up to 15%, highlighting:

- **Custom analogies:** A tailored approach to analogy generation using student profiles.
- **Rigorous evaluation:** A novel pipeline measuring learning outcomes with anonymized concepts.
- **Educational impact:** Evidence of analogies enhancing comprehension in student learning, paving the way for future research.

2. Motivating Experiments

Data. We focus on teaching scientific concepts and evaluating student understanding through quiz questions. Therefore, we construct a robust concept library by using LLM to extract key concepts from SciQ[2], a dataset containing science quizzes. To simulate diverse tutor-student interactions, we adopt the user profile data from the Personalized Proactive Conversations (PPC)[3] dataset.

Method. Our motivating experiment consists of two stages. In the analogy generation stage (Figure 2), LLM tutors are prompted to generate analogies tailored to the learner's background. Using chain-of-thought reasoning, we prompt our models to analyze the student profile, predict concept domains familiar to the student, and formulate an analogy that connects quiz-related 'target' concepts to student-familiar 'source' concepts.

In the evaluation stage, we assess the effectiveness of these analogies both quantitatively and qualitatively. As illustrated in Figure 3, we employ simulated student LLMs, which receive the tutor's explanation and attempt to answer quiz questions related to the concept – a process that we iterate for different student LLMs coming with diverse simulated backgrounds and knowledge priors. To ensure that learning improvements result directly from the analogy, we introduce concept anonymization, where key terms in both the teaching and testing phases are replaced with unique symbols. This prevents the student models from relying on pre-existing knowledge in model training, isolating the contribution of analogies.

1. All code associated with our project is available at `https://github.com/patrickpxye/analogy_tutor`
2. https://huggingface.co/datasets/allenai/sciq
3. https://huggingface.co/datasets/erbacher/personalized-proactive-conversations

Beyond simulation, we also conduct a few motivating cases of human user study to assess the qualitative impact of LLM-generated analogies on human learners (Appendix B).
Experiment. In experiment, we compare analogy-based teaching against generic teaching (directly prompting the LLM to formulate a teaching statement at its discretion). We evaluate four conditions: zero-shot and few-shot prompting, each tested with and without analogy-based teaching. We test four different models as LLM Tutors: GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet and Llama-3.2-1B, with the same student LLM (GPT-4o-mini).

3. Results and Discussion

Results. As shown in Table 1, most tutor models achieved comparable or even superior teaching outcomes when employing analogies, as compared to generic teaching. The notable performance of few-shot analogy tutors, especially in GPT-4o and Claude-3.5-Sonnet, implies the potential of effective analogy-based LLM tutoring. Smaller models, such as Llama-3.2-1B and GPT-4o-mini, exhibited less consistent performance improvements with the use of analogies, suggesting that the scale and reasoning capacity of a model significantly influence its ability to generate high-quality analogies.

This observation is further supported by real user case studies, where tailored analogies aligning with learners' backgrounds led to increased engagement, and participants are reported to have better comprehension and recall.

Additionally, the benefits of few-shot prompting varied across models. While few-shot examples may be crucial for the effective deployment of analogies in models with robust conceptual reasoning capabilities like Claude-3.5-Sonnet, for other models with varying reasoning capacities, few-shot examples may not be immediately meaningful and instead become confounding. These observations underscore the necessity for future emphasis on tailored analogy examples that account for the varying capacities of different LLMs.

Teaching Method	**GPT-4o**	**Claude-3.5-Sonnet**	**Llama-3.2-1B**	**GPT-4o-mini**
Zero-shot-prompted Generic Teaching	66.7%	**78.4%**	23.3%	**72.5%**
Zero-shot-prompted Analogy Teaching	**83.3%**	70.5%	**26.7%**	68.6%
Few-shot-prompted Generic Teaching	66.7%	76.5%	23.3%	**70.5%**
Few-shot-prompted Analogy Teaching	**73.3%**	**84.3%**	23.3%	64.7%

Table 1: Comparison of Teaching Methods and their Achieved Accuracy

Conclusion. Overall, our research underscores the potential of LLMs to serve as effective educational tools through the strategic use of analogies. By bridging abstract concepts with familiar contexts, analogy-based teaching not only improves comprehension and retention but also offers a more personalized and engaging learning experience. As educational technologies continue to evolve, leveraging analogical reasoning within LLMs presents a promising avenue for enhancing the quality and accessibility of education.

Future work can explore long-term retention effects of analogy-based learning, particularly in incremental learning settings. Enhancing interactive refinement, where LLM tutors adjust analogies dynamically based on student feedback, could improve personalization. Additionally, mitigating cultural biases in analogy selection remains an important challenge. Addressing these areas will further refine analogy-based LLM tutoring and enhance its real-world impact.

References

Bhavya Bhavya, Yang Zhou, Shradha Sehgal, Suma Bhat, and ChengXiang Zhai. Analego: Let's build analogies together! In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*, 2024. URL https://openreview.net/forum?id=Bh1RZ6vYnY.

Paul Denny, Sumit Gulwani, Neil T. Heffernan, Tanja Käser, Steven Moore, Anna N. Rafferty, and Adish Singla. Generative AI for education (GAIED): advances, opportunities, and challenges. *CoRR*, 2024.

Shawn M. Glynn and Tomone Takahashi. Learning from analogy-enhanced science text. *Journal of Research in Science Teaching*, 1998.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 2023.

Timothy J. Newby, Peggy A. Ertmer, and Donald A. Stepich. Instructional analogies and the learning of concepts. *Educational Technology Research and Development*, 1995.

Ashish Sharma, Kevin Rushton, Inna E. Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *ACL*, 2023.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 2023.

Appendix A. A two-stage experiment framework: The Analogy Formulation Framework in Figure 2 and The Evaluation Framework in Figure 3.

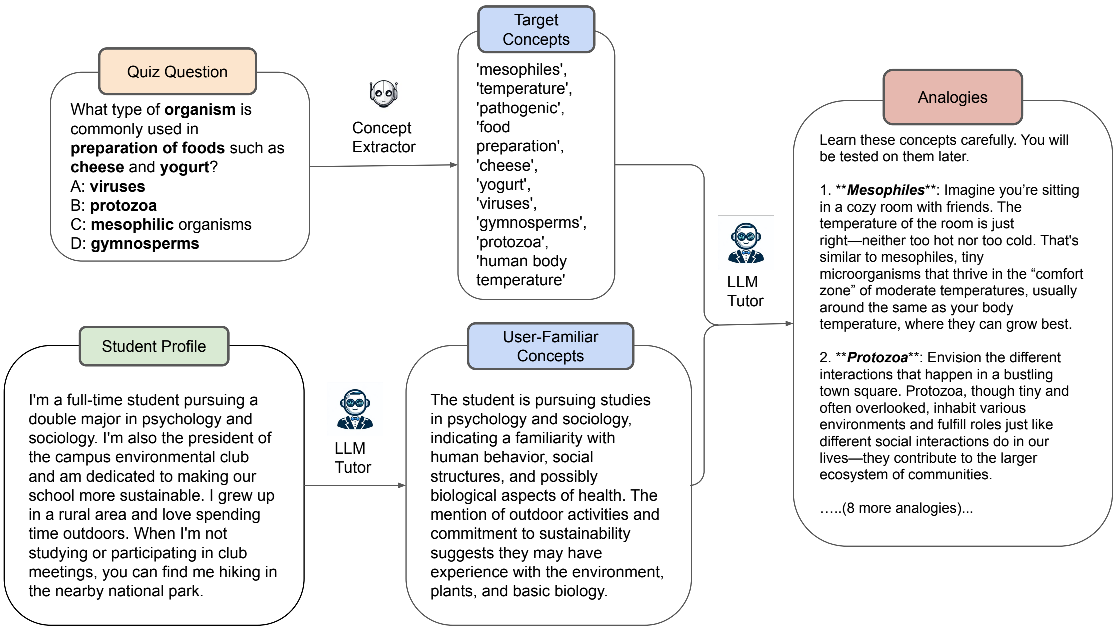

Figure 2: The Analogy Formulation Framework. For a given quiz question, the Concept Extractor LLM produces a list of key target concepts. The LLM Tutor is prompted to examine the student's profile and predict concept domains that students are likely familiar with, then generates an analogy that explains every key target concept in terms of student-familiar concepts

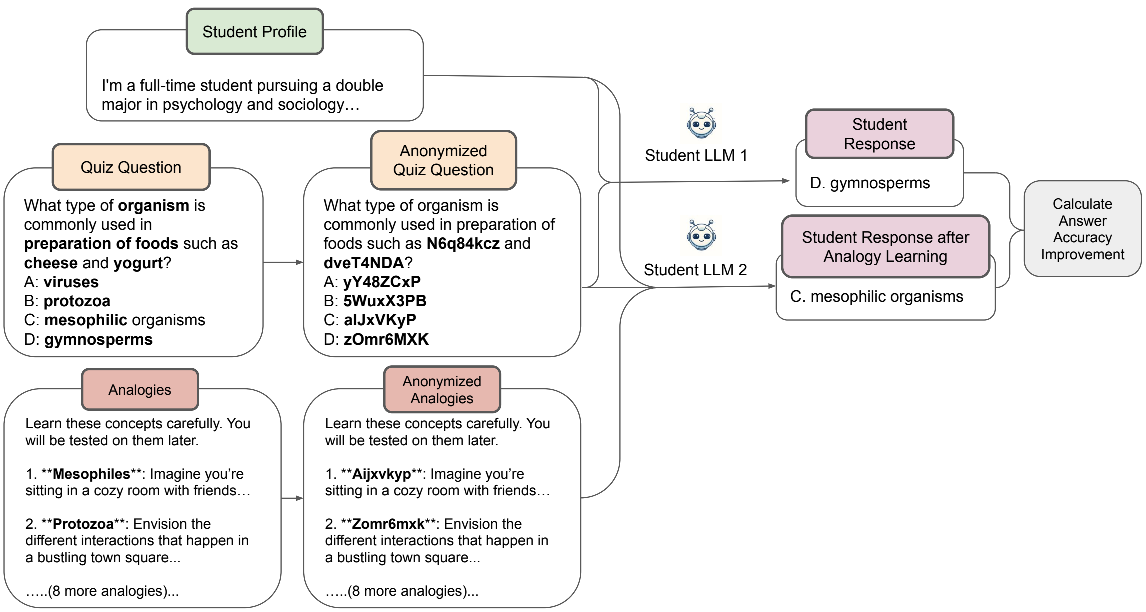

Figure 3: The Evaluation Framework. The Student LLM is prompted to tackle a quiz question based on the instruction of the LLM Tutor, in a manner that's consistent with the student's profile. We test different types of tutor instructions: generic teaching without analogy, zero-shot-prompted analogy teaching, and few-shot-prompted analogy teaching. Both quiz questions and tutorials are anonymized to isolate confounding prior knowledge.

Appendix B. Example case for Qualitative User Study

We developed an interactive interface, hosted on Huggingface Space, to enable anonymized users to interact with either the few-shot-prompted Analogy Tutor or a vanilla non-Analogy Tutor. This setup allowed us to observe user-tutor interactions and gain qualitative insights into how analogies contribute to learning. The interface includes the following components:

- **Tutor and Concept/Quiz Selection**: Users are randomly assigned either the Analogy Tutor or the non-Analogy Tutor. A quiz question is sampled from the test set, with relevant concepts extracted, as described in our Method Section.
- **User Background Statement**: Users provide a brief description of their background or interests (10–50 words). If the Analogy Tutor is selected, this input is used to customize the analogies generated for teaching the extracted concepts. The non-Analogy Tutor provides explanations without leveraging user-specific contexts.
- **Conversation**: Users interact with the tutor, reading explanations and posing follow-up questions in a multi-turn conversation to deepen their understanding of the concepts.
- **Quiz**: Users answer a multiple-choice question based on the tutor's explanations. Their responses are evaluated by the system.

We present an example interaction from a user with a background in networking. The selected quiz question and context enabled us to analyze how the few-shot Analogy Tutor dynamically customized its teaching strategies.

Quiz Question:

> *What type of taste do bases normally have?*
> Choices: A. sweet B. sour C. bitter D. salty

Correct Answer: C. bitter.
User Background: The user identified themselves as a networking engineer, providing insights into their professional expertise.
Teaching Interaction: The few-shot Analogy Tutor leveraged the user's networking background to frame the concept of "taste" in terms of networking protocols and data flows. The tutor introduced the analogy as follows:

> *"Taste can be likened to different networking protocols. For instance:*
> *- Sweet is like TCP, a reliable protocol that ensures data is delivered in an orderly manner.*
> *- Bitter resembles error messages or warnings in networks, signaling problems that need troubleshooting.*
> *- Sour indicates connection quality issues, much like a bad network link causing delays.*
> *- Salty represents critical alerts or updates that demand immediate attention."*

When the user engaged further by selecting "bitter" as the answer and asking about the purpose of the analogy, the tutor expanded:

> *"The analogy connects the abstract concept of taste to the practical domain of networking, making it easier to relate and remember. For example, bases are often described as bitter, much like error messages in networks that highlight issues needing resolution."*

The tutor further contextualized the analogy by refining it based on the user's interest in networking infrastructure, emphasizing how each "taste" parallels real-world scenarios in network management.

Appendix C. Prompts

We provide the detailed prompts used in our framework.

C.1. Concept Extraction

```
You are going to be presented with a multiple choice quiz question for
    ↪ middle school students. This will contain the question itself, the
    ↪ multiple choice answers, the correct answer, and the rationale for
    ↪ solving the question.

Here is the quiz question:
{quiz_question}

Your task is to identify the key concepts that the students need to
    ↪ understand in order to answer the question correctly. You should list
    ↪ concepts present in the question, the answer choices, and the
    ↪ rationale (the "Support" from the original dataset). Return this list
    ↪ as a comma-separated list of strings. You should include no more than
    ↪ 10 concepts.

For example, consider the following quiz question:

'question': 'What is the upper-most atmosphere known as?', 'choices': ['A':
    ↪ 'ionosphere', 'B': 'xerosphere', 'C': 'exosphere', 'D': 'thermosphere
    ↪ '], 'answer': 'D', 'rationale': 'The atmosphere is a big part of the
    ↪ water cycle. What do you think would happen to Earths water without
    ↪  it?.'

The list of concepts you return should be: ['atmosphere', 'ionosphere', '
    ↪ xerosphere', 'exosphere', 'thermosphere', 'water cycle']
```

C.2. Student Simulator

```
You are an AI assistant tasked with role-playing as a student learning
    ↪ concepts from an AI tutor. Your goal is to generate realistic and
    ↪ appropriate queries that a student might ask when learning.

# Your profile information when gives your knowledge background about
    ↪ concepts that you are familiar with:
{user_profile}
```

```
# Guidelines for Your Role as a Student:
1. Ask questions that demonstrate your current understanding and areas of
   ↪ confusion.
2. Respond to the tutor's explanations and hints, showing gradual progress
   ↪ in your understanding.
3. Occasionally make mistakes or misunderstandings that a real student might
   ↪  have.
4. Show persistence in trying to solve the problem, but also be willing to
   ↪ ask for help when stuck.

# Conversation History:
{chat_history}

# Task:
Your task is to generate response to the chat. This should be a single
   ↪ question or statement that follows naturally from the conversation
   ↪ history and demonstrates your current level of understanding and any
   ↪ areas where you need further clarification.

## What you will output:
First, output your thought process as a student deciding what to say next.
   ↪ When generating your thought process, you may consider the following:
1. What specific part of the problem or explanation are you struggling with,
   ↪  and how does the tutor's guidance relate to your current
   ↪ understanding?
2. Has the tutor asked you to perform a task or answer a question? If so,
   ↪ how should you approach it?
3. Are you noticing any patterns or potential misunderstandings that need
   ↪ clarification?
4. If you're stuck, how can you phrase your question to get the most helpful
   ↪  response while demonstrating your current understanding?

Then, based on your thought process, output your query prefixed with 'Query
   ↪ :', ensuring it aligns with the conversation history and reflects your
   ↪  current understanding.

If any of the following conditions are met, output "terminate conversation":
1. You believe you have solved the problem or gained enough understanding to
   ↪  attempt it on your own.
2. The tutor has provided a complete explanation and you have no further
   ↪ questions.
3. The conversation seems to be going in circles or not progressing.

## Output format
Thought: [Your chain of thought reasoning about what to ask next]

Query: [Your actual query for the AI tutor]

Remember to stay in character as a student throughout your response, and
   ↪ follow the instructions and guidelines carefully.

Make sure that your reasoning in answering this question doesn't include
   ↪ directly guessing or indirectly inducing the key concept. You should
   ↪ only rely on the knowledge you have learned from the AI tutor and the
```

```
    ↪ background knowledge of the middle school student you are role-playing
    ↪ .
```

C.3. LLM Tutor

Zero-shot prompting:

```
You are an AI assistant helping a student learn a series of concepts through
    ↪  some concepts they are already familiar with.

# Your Goal:
Your goal is to explain the concepts "{target_concepts}" by providing a
    ↪ series of analogies that relates them to more familiar concepts.
- The concepts you are trying to explain are "{target_concepts}"
- The concepts that the student is already familiar can be inferred from
    ↪ their user profile: {user_profile}
- You can assume that the student is familiar with all the concepts in their
    ↪  areas of study at the same level as a middle school student.
- Make sure to provide a coherent series of analogies that explain the
    ↪ target concepts in a logical order. Make sure to explain each target
    ↪ concept before moving on to the next one.
- Every analogy should be easy to understand and should help the student
    ↪ grasp the key ideas of the corresponding target concept.
- If the target concepts are related, make sure to show the relationship
    ↪ between them in your analogies.
- If the target concepts are not related, make sure to provide separate
    ↪ analogies for each target concept.

# Instructions for Formulating an Analogy:
An analogy must contain the following components:
1. Target: The concept that the analogy is trying to explain or clarify. In
    ↪ this case, the target is one of the target concept from the list of
    ↪ concepts you are trying to explain: {target_concepts}.
2. Source: The more familiar concepts that are used to explain the target
    ↪ concept. You should infer what concepts the student is familiar with
    ↪ from their user profile.

## What you will output:
First, output your thought process when formulating your explanation. When
    ↪ generating your thought process, you may consider the following:
1. What concepts or ideas is the student probably familiar with based on
    ↪ their user profile?
2. How can you formulate a coherent series of analogies that relate to the
    ↪ target concepts based on the student's existing knowledge?
3. What are the key ideas or components of the target concepts that you
    ↪ should include in your analogy?
4. How can you simplify the analogies to make them easy for the student to
    ↪ understand?

Then, based on your thought process, provide a series of analogies that
    ↪ explain the target concepts using concepts the student is already
    ↪ familiar with. Your explanation should be easy to understand and
    ↪ should help the student grasp the key ideas of the target concept.

## Output format
```

```
Thought: [Your chain of thought reasoning about what analogy to use and how
    ↪ to relate it to the student's existing knowledge]

Explanation: [Your analogy that explains the target concept using concepts
    ↪ the student is already familiar with]
```

The following examples are further added to the few-shot prompting:

```
## Example 1
[GIVEN] Target Concepts (to be explained): "Spanish Inquisition", "French
    ↪ Revolution", "American Civil War"
[GIVEN] User Profile: Enthusiastic biology student with no knowledge of
    ↪ history
[GENERATED] Thought: The student is likely familiar with the concept of
    ↪ conflicts and revolutions in biology. I can use the analogy of a "
    ↪ cellular revolution" to explain the "Spanish Inquisition" as a process
    ↪  of change and conflict within a cell. The "French Revolution" can be
    ↪ explained as a "genetic mutation" in a population, leading to
    ↪ significant changes. The "American Civil War" can be related to a "
    ↪ cell division" where two opposing forces within a cell fight for
    ↪ control.
[GENERATED] Explanation: The "Spanish Inquisition" was like a cellular
    ↪ revolution where different parts of the cell fought for control and
    ↪ dominance, causing significant changes. Similarly, the "French
    ↪ Revolution" was akin to a genetic mutation in a population that led to
    ↪  drastic alterations in the cell's structure. Lastly, the "American
    ↪ Civil War" can be compared to a cell division where two opposing
    ↪ forces within the cell clashed, resulting in a split and conflict.

## Example 2
[GIVEN] Target Concepts (to be explained): "Gravity", "Electromagnetic Force
    ↪ ", "Strong Nuclear Force"
[GIVEN] User Profile: Machine learning enthusiast with no knowledge of
    ↪ physics
[GENERATED] Thought: The student is likely familiar with the concept of
    ↪ forces and interactions in machine learning. I can use the analogy of
    ↪ "model optimization" to explain "Gravity" as the force that pulls
    ↪ objects together, similar to how optimization algorithms converge
    ↪ towards a solution. The "Electromagnetic Force" can be related to "
    ↪ feature interactions" in models, where different features interact
    ↪ with each other. The "Strong Nuclear Force" can be explained as the "
    ↪ model stability force" that holds the core components together.
[GENERATED] Explanation: "Gravity" is like the force that pulls objects
    ↪ together, similar to how optimization algorithms converge towards a
    ↪ solution in machine learning. The "Electromagnetic Force" can be
    ↪ compared to feature interactions in models, where different features
    ↪ interact with each other to produce an outcome. Lastly, the "Strong
    ↪ Nuclear Force" can be related to the model stability force that holds
    ↪ the core components together, ensuring stability and coherence.
```