Pre-Analysis Plan

Experimental Evidence on the Learning Impact of Generative AI*

Zara Contractor Germán Reyes

April 2025

1 Introduction

Higher education plays a central role in human capital accumulation by fostering both domain-specific knowledge and transferable skills such as critical thinking and analytical reasoning. The emergence of generative artificial intelligence (AI) tools, such as ChatGPT and Claude, has the potential to transform how students acquire these skills. These tools generate human-like text by learning from vast datasets and responding to natural language prompts. Despite their rapid adoption and the growing debate about their educational implications, there little empirical evidence on how access to AI affects immediate learning outcomes and longer-term knowledge retention—both critical components in the formation of skills.

Research from cognitive psychology provides a useful framework for understanding how generative AI might impact learning. A large body of empirical evidence shows that learning improves significantly when individuals actively generate information rather than passively consume it (Jacoby, 1978; Slamecka and Graf, 1978).[1] This literature suggests that the effects of generative AI on learning is theoretically ambiguous as it depends on how AI alters the inputs into the student learning production function.

*Contractor: Middlebury College (zcontractor@middlebury.edu), Reyes: Middlebury College (greyes@middlebury.edu). For helpful discussions and comments, we thank Jeff Carpenter, Amy Collier, Peter Matthews, David Munro, Caitlin Myers, Andrea Robbett, and participants in the Cornell behavioral economics group. Financial support from the Middlebury College's Office of the Provost is gratefully acknowledged. The survey was reviewed and approved by the IRB at Middlebury College.

[1]For meta-analyses, see Bertsch et al. (2007) and McCurdy et al. (2020). For evidence from neuroscience, see Rosner et al. (2013).

On one hand, generative AI could enhance learning when used as a complement to students' own thinking rather than a substitute for it. Many educational uses of AI—such as prompting explanations of difficult concepts, identifying errors in code, reformulating definitions, providing feedback, refining arguments, or offering personalized clarification—may increase the quality or efficiency of learning inputs. These uses can promote deeper engagement with the material and support the cognitive processes that underlie durable skill acquisition.

On the other hand, generative AI tools may impair learning if students rely on them to bypass effortful thinking. A defining feature of generative AI is their ability to generate content at near-zero marginal cost, which may encourage passive consumption. If students adopt AI-generated responses without meaningful cognitive engagement, the effective inputs into the learning process—such as active retrieval, synthesis, and critical reflection—may be reduced, leading to weaker skill formation.

Given these competing mechanisms, the net effect of generative AI on learning is ultimately an empirical question. To estimate the causal impact of generative AI on student learning, we conduct a lab experiment with undergraduate students at Middlebury College.[2] Participants are randomly assigned to either an AI-allowed or AI-forbidden condition during a structured learning task on an unfamiliar academic topic. The study consists of two sessions held one week apart. In Session One, students complete a baseline knowledge test, engage in a 35-minute learning phase, and produce a written analysis, with only the treatment group allowed to use AI tools like ChatGPT. Session Two—one week later—measures knowledge retention through a written task and a multiple-choice test completed without access to AI. This design allows us to estimate both immediate and longer-term effects of AI access on learning outcomes, while rich behavioral and survey data enable exploration of underlying mechanisms and treatment effect heterogeneity.

The rest of this Pre-Analysis Plan is structured as follows. Section 2 describes the experimental design, including recruitment procedures, treatment implementation, outcome measures, and session structure. Section 3 outlines the statistical analysis plan, detailing our primary outcomes, hypothesis tests, empirical specifications, and approaches to han-

[2] While laboratory experiments necessarily involve some degree of artificiality, this design offers several key advantages for studying the impact of AI on learning. First, the controlled setting enables clean identification of causal effects by randomly assigning access to AI while holding other factors constant—something that is not feasible with observational data, where AI usage is endogenous. Second, the lab environment allows for consistent administration of tasks and standardized measurement of both immediate performance and longer-term retention. Third, it facilitates close monitoring of compliance with treatment instructions and the collection of detailed behavioral data, including patterns of AI use during the learning process.

dling multiple hypothesis testing. We also describe planned analyses of treatment effect heterogeneity, underlying mechanisms, robustness checks, and potential attrition.

2 Experimental Design

This section describes our experimental design and implementation procedures. The study design and procedures were approved by Middlebury College's Institutional Review Board. All experimental procedures were implemented using Qualtrics survey software.

2.1 Measuring Student Learning

A central methodological challenge in educational research is the multidimensional nature of student learning. Learning encompasses a wide array of cognitive skills, ranging from basic recall of facts to complex critical thinking and evaluation. Given the absence of a universally accepted measurement approach, our strategy for assessing learning outcomes is guided by Bloom's taxonomy (Bloom et al., 1956), a widely adopted educational framework that classifies cognitive skills into six hierarchical levels: knowledge, comprehension, application, analysis, synthesis, and evaluation.

We designed assessments targeting distinct cognitive skills. We use multiple-choice questions primarily to capture foundational cognitive skills—knowledge and comprehension—by objectively evaluating participants' recall and basic understanding of the assigned topics. Complementing this, we use a structured writing task specifically designed to measure participants' analytical skills, a higher-order cognitive dimension that requires students to identify conceptual relationships, distinguish essential components, and critically examine key ideas. Further, informed by existing empirical evidence on the importance of assessing both immediate learning and delayed retention, we measure skills immediately following the learning task (Session One) and again approximately one week later (Session Two). This allows us to distinguish between short-term learning gains and more durable, sustained learning outcomes.

2.2 Overview of the Experiment

The experiment consists of two in-person sessions conducted approximately one week apart at Middlebury College during April and May 2025. We recruited undergraduate participants through campus-wide email announcements and posted flyers, framing the study as an investigation into the determinants of learning. During recruitment, we emphasized that participation required attending two sessions in consecutive weeks and asked students to select their preferred dates and time slots from multiple options for each session.

We implemented several design features to minimize attrition. First, we structured compensation to incentivize completion of both sessions: participants receive \$5 for completing only Session One and \$50 if they complete both sessions, with all payments disbursed at the conclusion of Session Two. Second, at the end of Session One, participants receive written confirmation of their Session Two appointment time followed by email reminders before their scheduled session. Third, to accommodate potential scheduling conflicts, we provided some flexibility in the timing of the second session rather than requiring exactly seven days between sessions. Fourth, participants who miss their scheduled Session Two appointment receive a follow-up email near the end of the study with a Qualtrics link allowing them to complete the Session Two online.

2.3 Session One: Knowledge Production and Initial Learning

Laboratory Setup. The first session employs a between-subjects design with random assignment to either an AI-allowed or AI-forbidden condition. For each scheduled time slot, two concurrent sessions—one per treatment arm—are run in separate buildings across campus. Participants are randomly assigned to one of these two separate computer labs to prevent cross-contamination between treatment groups.

Treatment Implementation. The core experimental manipulation centers on access to generative AI tools. Participants in the *AI-allowed* condition receive explicit instructions that they may use any generative AI tools for the main task. This is explained in the following message:

> **How should I approach learning this topic?**
>
> - You are encouraged to use the same learning approach you would typically follow for a college assignment.
> - For example, if you typically use generative AI tools like ChatGPT, Claude, Gemini, or any other generative AI tools, feel free to use them here as well. **Using generative AI assistance is completely acceptable and will not affect your compensation**
> - You are welcome to use standard online resources like Google, Wikipedia, Middlebury College Library's website, or other websites to look up information

Those in the *AI-forbidden* condition are instructed not to use any generative AI tools. Specifically, they see the following message:

> **How should I approach learning this topic?**
>
> - You are encouraged to use the same learning approach you would typically follow for a college assignment.
> - **You are not allowed to use any generative AI tools like ChatGPT, Claude, Gemini, or any other generative AI tools**
> - You are welcome to use standard online resources like Google, Wikipedia, Middlebury College Library's website, or other websites to look up information

Compliance with the instructions is enforced through direct monitoring by laboratory staff, periodic captures of participants' writing interface, and participants' self-reported AI usage collected at the end of the study.

Initial Procedures. Each session begins with instructions. We tell participants that their primary task involves learning about a pre-specified topic and that they will demonstrate their understanding through a series of assessments. To ensure comprehension of the experimental procedures, participants must correctly answer a series of understanding checks, with laboratory staff providing additional clarification if needed.

We selected three topics rich in factual content about which most undergraduate students likely have limited prior knowledge. The three topics are: Blockchain, Carbon capture, and CRISPR. Participants are randomly assigned one of these three topics.

Baseline Test. Following the instructions, participants complete a five-question multiple-choice baseline test to assess their prior knowledge of the assigned topic before the learning phase. In all tests, the order of questions and answer options within each question is randomized. Participants are prohibited from using external resources during this test. This baseline measurement enables us to distinguish between pre-existing knowledge and knowledge acquisition in our subsequent analyses.

Learning Phase and Writing Assessment. Following the baseline test, participants receive instructions for the Learning Phase. In this phase, participants have 35 minutes to learn about their randomly-assigned topic and to demonstrate their understanding through a writing assessment. We encourage them to "use the same learning approach you would typically follow for a college assignment" and emphasize that their performance affects their payoffs (as described below).

The writing assessment consists of an analytical task (approximately 500 words) examining specific aspects of the assigned topic. Participants are asked to analyze relationships between concepts, compare different elements, and support their analysis with specific examples (see Appendix B). For each topic, we developed two different analytical prompts and randomly selected one prompt for Session 1 and the other one for the writing task in Session 2. During the 35-minute learning phase, participants have complete flexibility to divide their time between researching the topic, planning their analysis, drafting their response, or other activities like surfing the web. Participants can spend less than 35 minutes during the learning phase, but no more than that. At the 35-minute mark, the computer screen automatically advances.

Writing assessments are evaluated by independent graders who are blind to treatment condition and participant identity. Graders assess the overall quality of the written work on a 0-10 scale, with 10 representing excellent understanding and synthesis of the topic. In additional to overall quality, other evaluation dimensions include comprehension (accurate understanding of key concepts), analytical reasoning (depth of analysis and critical thinking), organization (logical flow and structure), and evidence use (effective integration of factual information). We round up to the nearest integer to create the final quality score. In addition to human ratings, we conduct a linguistic analysis of the essays, measuring text length, lexical diversity, syntactic complexity, and readability. This allows us to assess how AI access influences written expression beyond subjective quality ratings.

Post-Task Test. Following the learning phase, participants complete five multiple-choice questions designed to test factual knowledge and conceptual understanding of the assigned topic. Participants complete this assessment without access to any external resources, internet, or AI tools. This measure of learning, combined with the retention assessment in Session Two, allows us to differentiate between participants' initial knowledge acquisition and their longer-term retention of that knowledge.

Incentives. To incentivize effort we implement a lottery-based compensation structure. In addition to their base compensation, participants receive one lottery ticket for each correct response on the baseline test and post-task test. They also receive one lottery ticket for each point on their rounded quality score (0-10 possible tickets). At the conclusion of the experiment, we conduct a drawing in which 30 lottery tickets are randomly selected, with each winning ticket worth $100. Participants may win multiple times if multiple tickets of theirs are drawn.[3]

[3]Based on our recruitment target of 200 participants, we anticipate distributing approximately 20 lottery tickets per participant on average across all tasks, resulting in approximately 4,000 total tickets in

2.4 Session Two: Knowledge Retention

Session Two occurs approximately one week after the initial session, balancing our desire to measure meaningful retention effects against practical constraints of maintaining participant engagement. Participants have no access to any external resources during the session. The session consists of two assessments (shown in random order) and an exit survey.

Knowledge Retention Assessments. One of the two assessments of Session Two consists of a 15-minute analytical writing task on their originally assigned topic from Session One. This task uses the complementary analytical prompt we developed for each topic, examining different aspects or applications of the same subject matter. Independent graders evaluate these submissions using the same rating criteria and procedures established in Session One, with each point on the 0-10 quality scale translating to one lottery ticket.

The second assessment consists of a 10-item multiple-choice test covering key factual content from their assigned topic. We designed these questions to be conceptually similar but structurally distinct from those in Session One, ensuring that we measure genuine knowledge retention rather than memorization of specific question formats or answer patterns. Each correct answer earns participants one additional lottery ticket.

Exit Survey. Session Two concludes with a questionnaire capturing participants' AI proficiency and experimental engagement. We assess participants' AI experience through multiple dimensions: their self-reported AI skills (Likert scale), frequency of AI use for different academic tasks (never to daily), when they started using AI, and specific types of academic work for which they typically employ AI (e.g., brainstorming, writing, editing, research). The questionnaire also captures their self-reported use of AI during Session One, their prior knowledge of the assigned topic, and any study or review of the topic between sessions.

3 Statistical Analysis

This section outlines our analysis plan, including primary outcomes, hypothesis tests, and empirical specifications.

---

the pool. With 30 lottery prizes valued at $100 each, each ticket has a 0.75 percent chance of winning, yielding an expected bonus of approximately $15 per participant.

3.1 Balance and Randomization Verification

To verify the success of our randomization procedure, we will assess balance across treatment conditions through several approaches. Our primary analysis computes differences in means for all baseline characteristics, including demographic variables (age, gender, race/ethnicity), academic variables (major, year, GPA), and measures of prior AI experience.

In addition, we will conduct a joint test of balance by regressing treatment assignment on the full set of baseline characteristics:

$$\mathrm{AI}_i = \gamma_0 + X_i'\gamma + \nu_i,$$

where AI_i is an indicator for assignment to the AI-allowed condition and X_i represents the vector of baseline characteristics. We test the null hypothesis that all coefficients are jointly zero ($H_0 : \gamma = 0$) using an F-test.

In cases where we observe significant imbalance in any characteristic (defined as individual p-value below 0.05), we will report treatment effect estimates both with and without controlling for the imbalanced characteristics.

3.2 Primary Outcomes

Our analysis focuses on four families of outcomes: baseline knowledge assessment, immediate learning outcomes from Session One, knowledge retention measures from Session Two, and AI usage patterns.

For the baseline knowledge assessment:

- *Baseline Knowledge*: The number of correct responses on the baseline assessment administered before the learning task to establish baseline knowledge of the assigned topic.

For immediate learning outcomes, we examine:

- *Written Assessment Quality*: We create standardized indices (mean 0, standard deviation 1) for each rating dimension (comprehension, analytical reasoning, organization, and evidence use). We will use the overall quality as well as the average of these indices. We also analyze each dimension separately as secondary outcomes. Specifically, for each rating dimension d and participant i, we calculate:

$$z_{id} = \frac{y_{id} - \mu_d}{\sigma_d},$$

where y_{id} is the raw rating on the 0-10 scale, μ_d is the mean, and σ_d is the standard deviation.

- *Computational Metrics*: We analyze text length (word count), lexical diversity (type-token ratio), syntactic complexity (average sentence length and clause density), and readability scores (Flesch-Kincaid grade level). We will also compute the distance between the provided reading and the student output, as well as the distance between the ChatGPT output and the student output.

- *Post-Learning Assessment Performance*: We measure the number of correct responses on the 5-item multiple choice test administered after the written assessment.

- *Learning Gain*: The difference between post-learning assessment performance and baseline knowledge assessment, measuring immediate knowledge acquisition.

- *Time Spent Writing*: We measure the number of minutes participants spend actively composing their written response during the learning phase, using embedded Qualtrics timers, periodic screen of the writing interface, and self-reported data. This measure allows us to assess how AI access influences time allocation between information gathering and active synthesis, as well as overall engagement with the writing task.

For knowledge retention, we examine:

- *Follow-up Essay Quality*: Using the same standardized indices as Session One

- *Factual Retention*: Score on the 10-item multiple choice test administered in Session Two

- *Retention Rate*: Measured as the ratio of Session Two performance to Session One post-learning assessment performance, indicating how much knowledge was retained over time

3.3 Primary Analysis

Our primary specification for estimating treatment effects is:

$$Y_i = \beta_0 + \beta_1 \mathrm{AI}_i + \beta_2 \mathrm{BaselineKnowledge}_i + \varepsilon_i, \tag{1}$$

where Y_i is an outcome measure for participant i (either from Session One or Session Two), AI_i is an indicator for assignment to the AI-allowed condition, and $\mathrm{BaselineKnowledge}_i$ is the score on the baseline knowledge assessment. Including baseline knowledge as a control improves precision by accounting for initial differences in topic familiarity. We estimate this specification separately for each outcome of interest. We cluster the standard errors at the student level.

Intention-to-Treat and Treatment-on-Treated Analysis. Our primary analysis estimates the average effect of being randomly assigned to the AI-allowed condition regardless of actual AI usage. Because not all participants assigned to the AI-allowed condition may choose to use generative AI tools during the learning task, this effect represents an intent-to-treat (ITT) estimate.

We will complement our ITT estimates with a treatment-on-treated (TOT) analysis that measures the effect of actual AI usage rather than just assignment. Specifically, we will instrument actual AI use with the randomly assigned treatment status (AI-allowed). The resulting two-stage least squares (2SLS) estimates will provide the local average treatment effect (LATE) for participants whose AI usage was influenced by their treatment assignment. We will implement this with the following specification:

$$\text{FirstStage:} \quad \mathrm{UsedAI}_i = \delta_0 + \delta_1 \mathrm{AI}_i + \delta_2 \mathrm{BaselineKnowledge}_i + \mu_i \tag{2}$$

$$\text{SecondStage:} \quad Y_i = \pi_0 + \pi_1 \widehat{\mathrm{UsedAI}}_i + \pi_2 \mathrm{BaselineKnowledge}_i + \epsilon_i, \tag{3}$$

where UsedAI_i is an indicator for whether participant i actually used generative AI during the learning task, and $\widehat{\mathrm{UsedAI}}_i$ represents the predicted values from the first-stage regression. The coefficient π_1 provides the estimated effect of AI usage on outcome Y_i for compliers—participants who use AI when allowed to and abstain when instructed not to.

Hypotheses and Testing. Our primary hypotheses examine the treatment effect β_1 from equation 1 for different outcomes:

1. *Immediate Learning* (Session One outcomes):
   - H1a: AI access improves written assessment quality ($\beta_1 > 0$)
   - H1b: AI access increases computational text metrics ($\beta_1 > 0$)
   - H1c: AI access reduces post-learning assessment performance ($\beta_1 < 0$)
2. *Knowledge Retention* (Session Two outcomes):
   - H2a: AI access reduces quality of the follow-up written task ($\beta_1 < 0$)
   - H2b: AI access reduces factual retention ($\beta_1 < 0$)

To account for multiple hypothesis testing, we:

1. Create standardized indices for families of related outcomes following Kling et al. (2007)
2. Report both unadjusted p-values and q-values that control the false discovery rate

3.4 Heterogeneity Analysis

We examine treatment effect heterogeneity by estimating:

$$Y_i = \theta_0 + \theta_1 \mathrm{AI}_i + \theta_2 X_i + \theta_3(\mathrm{AI}_i \times X_i) + \theta_4 \mathrm{BaselineKnowledge}_i + X_i'\lambda + \nu_i \tag{4}$$

where X_i represents participant characteristics. We focus on four key dimensions of heterogeneity. First, we examine variation by demographic characteristics, such as gender and cohort. Second, we analyze how treatment effects vary with participants' prior AI skill level, as proxied by self-reported AI experience and frequency of AI use in academic contexts. Third, we consider academic performance, using self-reported GPA and test scores as a proxy for general academic ability.

For participants in the AI-allowed condition, we analyze AI interaction patterns to explore treatment effect heterogeneity. We systematically categorize participant queries (based on ChatGPT prompt history) using computational text analysis methods such as word embeddings, creating taxonomies based on query intent (e.g., requests for explanations, factual inquiries, summarization requests, draft generation). This classification enables us to measure both the quantity and temporal distribution of different AI interaction types throughout the learning phase. We then examine correlations between these interaction patterns and our outcome measures to identify which AI usage strategies are most conducive to effective learning. Additionally, we investigate whether AI's impact varies with participants' baseline knowledge, testing whether the technology provides differential benefits for students with limited versus extensive prior familiarity with the assigned topic

3.5 Mechanisms

To better understand how AI access affects learning outcomes, we examine participants' time allocation during the learning task to see how AI access influences the learning process itself. We ask participants to report how they divided their time across different activities: reading the provided texts, searching for additional information, editing, etc.. This data allows us to test whether AI access changes how students engage with learning materials. For instance, students with AI access might spend less time reading and processing the original texts, potentially affecting their retention of the material.

3.6 Robustness

We will examine robustness to:

1. *Alternative methods of aggregating rater scores.* Our primary analysis uses mean ratings across graders, but this approach may be sensitive to outlier ratings or individual grader effects. We therefore test robustness to using median ratings, which are less sensitive to extreme scores. We also examine results after excluding ratings that deviate substantially from others (defined as more than two points from the mean of other graders' scores). Finally, we employ principal components analysis to create weighted composite scores that account for correlation patterns across rating dimensions.

2. *Different approaches to standardizing assessment scores.* While our main analysis standardizes scores within the full sample, we will verify robustness to alternative approaches such as calculating percentile ranks or using raw scores.

3. *Alternative specifications for learning gains.* Our primary measure of learning is the score on the post-learning assessment controlling for baseline assessment scores. We test robustness to using the difference between these two scores and percentage improvement.

4. *Various sample restrictions.* To ensure our results are not driven by unusual testing conditions, we examine sensitivity to excluding participants who complete Session Two online rather than in the laboratory (if any). We also test robustness to restricting the sample to participants who complete both sessions within a narrower window of days, reducing variation in the time between sessions.

5. *Alternative specifications.* Our main analysis controls only for baseline knowledge, but we examine sensitivity to including various additional control variables such as academic domain fixed effects and demographic characteristics. We also test different approaches to clustering standard errors, including two-way clustering by participant and academic domain.

3.7 Attrition

Despite implementing substantial retention incentives ($50 completion bonus and lottery tickets that can only be redeemed at Session Two), we anticipate some attrition between sessions. To address this, we employ several complementary approaches. We begin by comparing attrition rates across treatment conditions to assess whether AI access influences the likelihood of returning for Session Two. Building on this analysis, we examine whether attrition patterns correlate with Session One outcomes or participant characteristics, which would suggest selective attrition that could bias our treatment effect estimates.

To account for potential selective attrition, we implement two main empirical strategies. First, we conduct Lee bounds analysis to estimate worst- and best-case scenarios for our treatment effects, providing a range of potential estimates that account for arbitrary selection into Session Two completion. Second, we use inverse probability weighting, leveraging rich data from Session One to adjust for observable differences between participants who complete both sessions and those who drop out. This approach reweights our sample to account for differential attrition based on observable characteristics.

3.8 Power Analysis

To determine the appropriate sample size for our study, we conducted power calculations based on effect sizes observed in recent research on generative AI and productivity (Noy and Zhang, 2023), as well as the established literature on the generation effect in learning and memory (Bertsch et al., 2007; McCurdy et al., 2020).

Written Assessment Quality. For written assessment quality, we anticipate an effect size of approximately 0.4 standard deviations based on the impact of generative AI on comparable cognitive tasks (Noy and Zhang, 2023). Using a two-sided t-test with $\alpha = 0.05$ and power of 0.8, we require 100 participants per treatment group, or 200 total participants.

Knowledge Acquisition and Retention. For knowledge acquisition (measured by the difference between baseline and post-learning assessments), we draw on the generation effect literature, which shows that actively generating information versus passively reading it produces a moderate to large effect on memory performance. Meta-analyses of the generation effect (Bertsch et al., 2007) report an average effect size of approximately 0.40 standard deviations. In our study, we anticipate that access to generative AI tools may influence how students engage with learning materials—potentially shifting some students from active generation to more passive consumption of AI-generated content. Based on this literature, we expect an effect size of approximately 0.4 standard deviations for immediate learning outcomes, requiring 100 participants per group for 80% power.

For Session Two retention outcomes, the generation effect literature suggests that generation benefits may become more pronounced over time. Meta-analyses indicate that generation effects are larger when measured after longer retention intervals (0.64 standard deviations for delays exceeding one day) compared to immediate testing (Bertsch et al., 2007). With $\alpha = 0.05$ and power of 0.8, this requires 40 participants per group, or 80 total participants. Accounting for an anticipated 10% attrition rate between sessions, we would need to 44 initial participants per group (88 total).

Minimum Detectable Effects. With our target sample size of 200 participants completing both sessions (approximately 180 after accounting for attrition), the minimum detectable effect size at 80% power and $\alpha = 0.05$ is approximately 0.398 standard deviations for our main between-group comparisons.

Table 1 summarizes our power calculations for different outcomes and assumptions.

Table 1: Power Calculations for Primary Outcomes

Outcome	Effect Size	Required N	Source
Writing Quality	0.4 SD	200	Noy and Zhang (2023)
Knowledge Acquisition	0.4 SD	200	Bertsch et al. (2007)
Retention (1 week)	0.64 SD	80	Bertsch et al. (2007)

Required N for $\alpha = 0.05$ and power = 0.80. Planned N = 200 participants (180 after 10% attrition).

Appendix

A Learning Materials

A.1 Overview of Reading Materials

To facilitate a standardized learning experience across treatment conditions, we developed original instructional texts on each of the three assigned topics: Blockchain, Carbon Capture, and CRISPR (see Appendices A.2, A.3, and A.4 for the text of each document). These materials were intentionally designed to be of comparable complexity and structure, ensuring that all participants face similar cognitive demands regardless of their topic assignment.

Each text follows a parallel structure, beginning with a general overview of the technology, followed by its historical development, technical underpinnings, core applications, and limitations. The materials are information-rich but self-contained, enabling comprehension without prior topic knowledge. Texts range from 1,253 to 1,399 words in length. Based on an average silent reading speed of approximately 250 words per minute for college students (Carver, 1992; Brysbaert, 2019), the estimated reading time for each text is between 5.0 and 5.6 minutes.

The texts are closely matched across several readability metrics (Appendix Table A1). Flesch-Kincaid Grade Levels fall between 15.1 and 15.9, corresponding to late undergraduate reading difficulty. Measures of lexical and syntactic complexity, including sentence length and word length, also suggest a high but consistent level of challenge across topics.

Table A1: Readability Metrics for Topic Texts

Metric	**Blockchain**	**Carbon Capture**	**CRISPR**
Flesch-Kincaid Grade Level	15.90	15.76	15.11
SMOG Index	16.75	17.28	16.50
Automated Readability Index	17.75	16.26	15.86
Type-Token Ratio	0.51	0.49	0.51
Average Sentence Length	21.59	24.15	20.80
Average Word Length	6.43	6.09	6.07
Word Count	1,253	1,333	1,399
Estimated Reading Time (250 wpm)	5.0 min	5.3 min	5.6 min

A.2 Topic Text: Blockchain Technology: A Revolution in Digital Trust

Blockchain technology is a digital ledger shared across many computers that records information securely, making it very difficult to change or hack. Unlike regular databases managed by one central authority, blockchain spreads identical copies among all participants. Each "block" is linked to the one before it using cryptography, creating a chain of records that can't be easily altered. This design ensures everyone can see the same information in the exact order it was added.

Origins and Evolution

The basic ideas behind blockchain first appeared in 1991 when researchers Stuart Haber and W. Scott Stornetta proposed a way to timestamp digital documents using a chain of blocks secured by cryptography. Later, computer scientist Nick Szabo developed these ideas further with his bit gold proposal, creating early frameworks for digital currency without central control.

The big breakthrough came in 2008 when someone using the name Satoshi Nakamoto published a paper describing Bitcoin, a digital cash system that works without a central authority. In January 2009, the Bitcoin network launched, using the first working blockchain as its public transaction record. This innovation solved a major problem called "double-spending" that had prevented earlier digital currencies from working properly—ensuring that once you spend a bitcoin, you can't copy it and spend it again.

Around 2014, developers realized blockchain could do much more than just handle digital money, beginning what's called Blockchain 2.0. The creation of Ethereum in 2015 was a huge step forward because it introduced "smart contracts"—programs that automatically execute agreements when certain conditions are met. This expanded what blockchain could do, enabling complex automated transactions and creating the foundation for decentralized applications.

Technical Architecture

Blockchain works on three main principles: block structure, distributed consensus, and cryptographic security.

Each block contains a group of valid transactions, a timestamp, and a link to the previous block. When a block is filled, it gets sealed with a unique digital fingerprint called a hash. This hash becomes part of the next block's header, creating an unbreakable chain—

if someone tries to change even one transaction, all following blocks would immediately show the tampering.

Blockchain networks typically operate peer-to-peer, with each full node (computer) keeping a complete copy of the ledger. To add new transactions, the network must agree through specific methods. Bitcoin introduced Proof of Work (PoW), where special nodes called miners compete to solve difficult math problems. The first to find a solution gets to add the next block and earns a reward. Other nodes quickly check this solution and agree on the current state of the ledger.

Cryptography keeps blockchain secure. Users sign transactions with private keys, creating signatures that can be verified with public keys. This ensures only the rightful account holders can start transactions, while cryptographic hashing guarantees data stays intact. These combined methods allow blockchains to work without requiring trust between users—participants only need to trust the math behind the system.

Other consensus methods have developed to address PoW's limitations. Notably, Proof of Stake (PoS) chooses validators based on how much cryptocurrency they hold rather than computing power. Ethereum's switch from PoW to PoS in 2022 greatly improved its energy efficiency and transaction speed while maintaining security.

Applications Beyond Cryptocurrencies

While blockchain became famous through Bitcoin, it now has many uses in areas where trust, transparency, and security are important:

- **Finance and Banking:** Beyond digital currencies, blockchains enable faster international payments with fewer middlemen. Banks are exploring blockchain for transaction settlement, trade finance, and digital assets. Decentralized Finance (DeFi) platforms use smart contracts to offer banking services without traditional institutions.

- **Supply Chain and Logistics:** Blockchain improves supply chain visibility by recording each product's journey on an unchangeable ledger. This helps companies and consumers verify authenticity, reduces fraud, and speeds up identifying contamination sources in food safety problems.

- **Healthcare and Identity:** In healthcare, blockchain allows secure management of patient records while maintaining data integrity and patient control. Similarly, blockchain-based digital identity systems let individuals own and selectively share their credentials without central oversight, reducing identity theft and simplifying verification.

- **Government and Public Services:** Governments are using blockchains for tamper-resistant public records like land registries and birth certificates. Blockchain-based voting systems could potentially increase election transparency, though privacy and security challenges remain significant.

- **Digital Assets and Media:** Non-fungible tokens (NFTs) show how blockchain can authenticate digital art and collectibles. Content creators use blockchain-powered platforms to establish ownership and receive direct payment without corporate middlemen.

The common thread across these applications is blockchain's ability to provide a shared, tamper-resistant record that reduces disagreements and disputes. Smart contracts further extend functionality by automating transactions when preset conditions are met, enabling new forms of interaction across different sectors.

Limitations and Challenges

Despite its potential to transform many industries, blockchain technology faces several limitations that slow widespread adoption:

- **Scalability:** Traditional blockchain networks can't process many transactions quickly. Bitcoin handles only about 7–10 transactions per second—far below conventional payment networks like Visa. This limitation causes congestion and higher fees during busy periods. While solutions (layer-2 networks, sharding, roll-ups) continue to develop, the industry faces what's called the "blockchain trilemma"—the challenge of maximizing decentralization, security, and scalability all at once without compromise.

- **Energy Consumption:** Proof of Work consensus uses substantial electricity. Bitcoin mining operations consume approximately as much power as an entire small country each year. This environmental impact has prompted moves toward renewable energy sources and the development of energy-efficient alternatives like Proof of Stake, which Ethereum adopted in 2022, drastically reducing its power needs.

- **Latency and Finality:** Blockchain confirmation times can be too slow for real-time applications. Bitcoin requires about one hour (six block confirmations) for transaction finality—impractical for many uses requiring instant settlement. Newer protocols address this issue but face inherent limitations in achieving global consensus.

- **Security Vulnerabilities:** While major networks like Bitcoin maintain strong security through broad participation, smaller blockchains remain vulnerable to "51% attacks," where an entity controlling most network resources could potentially manipulate the ledger. Security therefore depends strongly on network size and diverse participation.

- **Governance Challenges:** Decentralized systems face complex governance issues. Without central authority, protocol changes require community agreement, sometimes leading to contentious splits when disagreements persist. On the other hand, permissioned blockchains gain efficiency but lose the trustless properties that make public blockchains revolutionary.

- **Regulatory Uncertainty:** Blockchain's permanence can conflict with privacy regulations like GDPR's "right to be forgotten." Additionally, questions about jurisdiction, smart contract enforcement, and compliance remain unresolved in many regions, hindering adoption in regulated industries.

These challenges reflect blockchain's fundamental trade-offs: decentralization often reduces performance; cryptographic security must balance against privacy concerns; and the absence of central control complicates coordination. While ongoing innovations address these limitations, careful implementation requires assessing whether blockchain truly offers advantages over traditional databases for specific applications.

Looking Forward

Blockchain has evolved from an obscure concept into a transformative technology, combining distributed systems, cryptography, and consensus mechanisms to enable digital trust without central authorities. Its journey from cryptocurrency infrastructure to multi-sector innovation platform shows remarkable versatility, though not without inherent limitations.

As research advances in scalability, energy efficiency, and novel applications, blockchain may eventually become as fundamental to digital infrastructure as databases are today—but with enhanced decentralization and trust characteristics. Understanding both its revolutionary potential and practical constraints will be essential for determining where blockchain can deliver genuine impact rather than technological redundancy.

In the dynamic landscape of emerging technologies, blockchain stands as a breakthrough in distributed trust—its history brief but significant, its future still unfolding through continuous innovation and practical implementation.

A.3 Topic Text: Carbon Capture Technologies: The Long Pursuit of Locking Away Carbon

Carbon capture technologies are designed to collect CO_2 from industrial smokestacks or directly from the air before it reaches the atmosphere. They represent one of our more straightforward approaches to fighting climate change. Yet despite their simple concept, these technologies remain difficult to implement on a large scale. This article looks at their history, how they work, and their limitations, as well as their potential role in our future efforts to reduce carbon emissions.

From Industrial Utility to Climate Solution

Carbon capture didn't start as a climate solution. Beginning in the 1920s, industries separated CO_2 from gas streams for practical reasons: to purify natural gas or help extract more oil from aging oil fields. The Terrell gas plant in Texas, which in 1972 began sending captured CO_2 to declining oil fields, is a good example of this early practical approach.

Only in the late 1970s did carbon capture begin to shift toward climate protection. Scientists like Cesare Marchetti suggested capturing power plant emissions and disposing of them—initially thinking about injecting them deep into the ocean, an idea later found to be environmentally harmful. Interest grew in the 1980s as concerns about global warming increased, leading to Norway's important Sleipner project in 1996, where Statoil (now Equinor) began injecting about one million tons of CO_2 each year into an underwater salt water reservoir, partly motivated by Norway's carbon tax.

The 2000s brought more investment, though progress was slow—about 70% of proposed carbon capture and storage (CCS) projects never made it past the planning stage. Notable successes included SaskPower's Boundary Dam in Canada (2014), the first commercial coal-fired power plant with CCS, and various facilities in China, Europe, Australia, and the Middle East.

Today's deployment remains modest. About 44 plants operate globally with CCS, together capturing only about 0.1% of global CO_2 emissions—a tiny fraction of the 35+

billion tons released annually. Nevertheless, many climate models predict that CCS could eventually contribute 14–19% of CO_2 reductions needed by 2050, especially in industries that are difficult to decarbonize like cement and steel manufacturing.

The Mechanics: Capture, Compress, Store

Carbon capture involves three main steps: separation, compression, and storage. First, CO_2 is isolated from exhaust gases or ambient air using chemical or physical processes. The extracted CO_2 is then compressed into a dense, liquid-like state and transported (usually through pipelines) to storage sites. Finally, it's injected deep underground into geological formations—often depleted oil reservoirs or salty water aquifers at least a kilometer below the surface—where impermeable rock prevents it from escaping. Over time, the CO_2 may dissolve into the salty water or react with minerals to form stable carbonates.

In some cases, rather than simply storing the CO_2, it's put to use—the "U" in CCUS (Carbon Capture, Utilization, and Storage). Enhanced oil recovery remains the main application, though CO_2 can also be used in beverage carbonation, synthetic fuel production, or building material manufacturing. Unless these uses permanently lock the carbon in stable compounds, however, it eventually returns to the atmosphere.

A well-equipped facility can capture approximately 90% of its CO_2 emissions—significant at the individual source level, though engineering and economic challenges make global-scale implementation difficult.

Four Approaches to Capture

Carbon capture technologies fall into four major categories, each with distinct applications and limitations:

- **Post-Combustion Capture:** The most common approach, particularly for updating existing facilities. After fuel is burned, exhaust gases (typically containing 5–15% CO_2) are passed through chemical solutions that selectively absorb CO_2. Later, the solution is heated to release pure CO_2 and regenerate the capture medium. This method's flexibility allows it to be applied to various sources—power plants, steel mills, cement factories—though the energy-intensive process and large equipment present challenges.

- **Pre-Combustion Capture:** In this approach, CO_2 is removed before combustion. Fuel is first converted into syngas (primarily carbon monoxide, hydrogen, and CO_2), then processed through a "shift" reaction producing more CO_2 and hydrogen. After the CO_2 is separated, the remaining hydrogen-rich fuel burns cleanly. This approach works well in certain industrial settings, particularly ammonia production and "blue hydrogen" manufacturing. While more efficient than post-combustion for CO_2-rich gas streams, it requires specialized equipment and processes, limiting retrofit potential.

- **Oxy-Fuel Combustion:** Rather than separating CO_2 from nitrogen-diluted exhaust, oxy-fuel combustion burns fuel in pure oxygen (often mixed with recycled CO_2 or water to moderate temperature). The resulting exhaust—primarily CO_2 and water vapor—yields nearly pure CO_2 after water condensation. This method simplifies separation but requires energy-intensive oxygen production. While tested in pilot projects like Germany's Schwarze Pumpe plant (2008), oxy-fuel remains at the demonstration stage for power generation.

- **Direct Air Capture (DAC):** Unlike the others, DAC extracts CO_2 directly from ambient air rather than from specific sources. Large industrial fans move air through filtration materials that chemically bind CO_2. Heat or electrical energy later releases the concentrated CO_2 for compression and storage. Companies like Climeworks (Switzerland) and Carbon Engineering (Canada) have developed commercial-scale facilities, though still small ones—Climeworks' Orca plant in Iceland captures about 4,000 tons annually.

DAC offers location flexibility and can address diffuse emissions or even achieve negative emissions. Its main challenge lies in the extreme dilution of atmospheric CO_2 (0.04% versus concentrated industrial sources), making it energy-intensive and expensive—currently hundreds to over $1,000 per ton, compared to tens of dollars for point-source capture.

Promise and Limitations

Carbon capture's potential contribution to climate mitigation remains limited by several factors:

- **Scale Versus Need:** Current capacity (approximately 50 million tons annually) represents a tiny fraction of emissions. Meaningful climate impact would require gigatons-scale deployment—at least 1,000+ million tons yearly—requiring massive infrastructure expansion and investment.

- **Technical Efficiency:** While 90% capture rates are theoretically achievable, practical results often fall short due to operational issues or cost constraints. Additionally, CCS doesn't address upstream emissions (like methane leaks) or eliminate all facility emissions.
- **Storage Capacity:** Geological storage capacity appears sufficient—likely hundreds of years' worth at current emission rates. Most industrial regions have accessible storage formations, and experience from projects like Sleipner indicates that properly selected sites can retain over 99% of injected CO_2 for thousands of years. However, comprehensive site surveys, monitoring protocols, and legal frameworks remain incomplete in many regions.
- **Economic Hurdles:** CCS imposes substantial costs. The energy penalty alone is significant—a coal-fired power plant with CCS requires 20–30% more fuel to generate the same electricity. Capital expenses for equipment and the absence of revenue from underground storage (except in enhanced oil recovery) further undermine economic viability without policy support. Current capture costs range from $60–100 per ton for industrial sources to several hundred dollars for direct air capture.
- **Risk Management:** While generally considered safe, CCS involves potential risks, including CO_2 leakage and induced seismic activity from pressurized injection. Other environmental considerations include increased water consumption (up to 50% more for coal plants with CCS) and waste products from degraded chemicals. Public acceptance of pipelines and storage sites presents another hurdle, with some communities opposing CCS as prolonging fossil fuel dependence.

Strategic Role

Carbon capture represents one tool in a comprehensive climate strategy, not a complete solution. For sectors that can be electrified or use alternative energy sources, those options typically offer lower costs and additional benefits. However, CCS offers critical emissions reduction potential for hard-to-decarbonize industries like cement and steel manufacturing, where process emissions are inherent to current production methods.

Additionally, bioenergy with CCS (BECCS) and direct air capture with storage (DACCS) could potentially achieve negative emissions later this century, offsetting remaining emissions from other sectors. Their deployment at scale will depend on technological advancement, policy support, and international coordination.

Carbon capture technologies have evolved from industrial processes to potential climate solutions, with proven effectiveness at facility scale. Yet their global impact remains small, limited by economic, technical, and infrastructural challenges. With continued innovation and supportive policies, carbon capture could significantly contribute to emission reduction strategies, particularly in sectors where alternatives remain elusive, while complementing the broader transition to renewable energy and sustainable practices.

A.4 Topic Text: CRISPR: Rewriting the Code of Life

Scientists around the world now use a powerful tool that can edit DNA with amazing precision. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) works like molecular scissors that can cut and change DNA accurately. When paired with enzymes like Cas9, this technology has changed genetics by offering a faster, cheaper, and more precise method than older techniques.

From Bacterial Defense to Scientific Breakthrough

CRISPR's story begins with simple scientific curiosity. In 1987, Japanese researchers found unusual repeated DNA sequences in *E. coli* bacteria but didn't know what they did. Throughout the 1990s, scientists including Francisco Mojica noticed similar patterns in many microbes. The key insight came in 2005: these repeats were part of a bacterial immune system that stored pieces of viral DNA to recognize future attacks.

The shift from microbial curiosity to biotechnology revolution sped up when Emmanuelle Charpentier discovered tracrRNA, an important part of the system, in *Streptococcus* bacteria. Her teamwork with Jennifer Doudna proved crucial. In 2012, they turned the bacterial mechanism into a programmable tool by combining CRISPR's RNA components into a single "guide RNA" that could direct Cas9 to cut DNA at specific sites. By 2013, researchers showed this tool could edit genes in living mammalian cells. This scientific breakthrough earned Charpentier and Doudna the 2020 Nobel Prize in Chemistry.

The Molecular Scissors: How CRISPR Works

At its core, CRISPR-Cas9 is a find-and-cut system guided by RNA. The technology uses two main parts: a guide RNA molecule designed to match a specific DNA target and the Cas9 enzyme that acts as molecular scissors.

The editing process follows three basic steps. First, researchers design a guide RNA that matches their target gene. This RNA works like a GPS once inside cells, leading Cas9 to the exact genomic location. When the guide RNA finds its matching sequence, Cas9 cuts both strands of the DNA at that spot. The cut triggers the cell's repair systems that scientists can control—either letting the cell repair the break imperfectly (usually disabling the gene) or providing a template DNA with a desired sequence for the cell to incorporate during repair.

This simple approach—design a guide RNA, find a target, make a cut—gives CRISPR great flexibility. If the DNA sequence is known, it can be targeted, making the technology useful in almost any organism.

Medicine's New Frontier

CRISPR's most promising uses are in medicine, where gene editing could fix genetic defects, modify cells to fight disease, or eliminate pathogens completely.

The first major clinical success came in late 2023 when the FDA approved the first CRISPR-based therapy for sickle cell disease. This treatment edits a patient's own blood stem cells to produce healthy red blood cells, effectively curing this painful condition. Other genetic disorders like cystic fibrosis, muscular dystrophy, and hemophilia are now being studied in clinical trials.

Cancer treatment is another important area. CRISPR allows scientists to engineer immune cells with better cancer-fighting abilities. The first U.S. clinical trial testing CRISPR-edited T cells in cancer patients began in 2019, with early results showing these modified immune cells could be safely put back into patients. Beyond direct treatments, CRISPR serves as a powerful research tool in cancer studies, allowing scientists to discover cancer weaknesses by systematically turning off genes in tumor cells.

For infectious diseases, CRISPR offers new approaches for both treatment and diagnosis. Researchers have tried using CRISPR to remove HIV genomes from infected cells, potentially curing HIV/AIDS. During the COVID-19 pandemic, scientists developed quick CRISPR-based diagnostic tests that detect viral RNA with high sensitivity, showing the technology's usefulness beyond editing.

Agricultural Revolution and Industrial Applications

Beyond medicine, CRISPR is changing agriculture by enabling precise genetic improvements that were previously impossible or impractical. Plant scientists use CRISPR to develop crops with better disease resistance, drought tolerance, improved nutrition, and higher yields—often by changing a plant's existing genes rather than adding foreign DNA.

A notable early success was a CRISPR-modified white button mushroom designed not to brown when cut, extending shelf life without introducing foreign genes. Similar advances have produced virus-resistant tomatoes, mildew-resistant wheat, and rice that can survive flooding. Importantly, some CRISPR-edited crops have faced fewer regulatory hurdles than traditional GMOs because they contain no foreign DNA.

In animal farming, the technology enables precise breeding with desired traits in just one generation. Researchers have created hornless dairy cattle (avoiding painful dehorning procedures) and pigs resistant to devastating diseases like PRRS. The ability to introduce beneficial traits without lengthy breeding programs represents a major change in animal agriculture.

Industrial biotechnology also benefits from CRISPR's precision. Microorganisms can be engineered to produce pharmaceuticals, biofuels, or specialty chemicals more efficiently. CRISPR-modified yeast improves bioethanol production while edited bacteria can make medical compounds or even break down pollutants.

Technical Hurdles and Safety Concerns

Despite its revolutionary potential, CRISPR technology faces significant challenges. Perhaps most concerning are off-target effects—cases where Cas9 cuts DNA at unintended sites with sequences similar to the target. Such unintended edits could potentially disrupt important genes, with theoretical risks including cancer. Researchers have made good progress through improved design of CRISPR components and development of more accurate Cas variants, but concerns remain.

Delivery presents another major challenge. While introducing CRISPR into lab cell cultures is straightforward, delivering the system to specific tissues within the human body is much harder. Current approaches use modified viruses or chemical delivery systems, each with limitations. Viral vectors can only carry payloads of certain sizes, while the body's immune system may attack both the delivery vehicle and the bacterial Cas9 protein itself. Different tissues present unique challenges—crossing the blood-brain barrier remains especially difficult.

Even successful delivery may result in incomplete editing, where some cells receive the intended modification while others don't. This unevenness creates particular concerns for embryo editing, where uneven genetic changes could spread unpredictably through tissues.

Intellectual property disputes have also complicated CRISPR's development. High-profile patent battles between research institutions have influenced how the technology is shared, licensed, and commercialized, affecting investment patterns and collaboration potential in the field.

The Ethical Frontier

CRISPR's ability to alter genetic code raises deep ethical questions that society has only begun to address. The most controversial divide exists between somatic editing (modifying non-reproductive cells, affecting only the treated individual) and germline editing (altering eggs, sperm, or embryos with changes that pass to future generations).

Most scientists and regulators believe that germline editing is currently unethical and unsafe, with such procedures illegal in many countries. The 2018 announcement by Chinese scientist He Jiankui that he had created the first CRISPR-edited babies—modified as embryos to resist HIV—generated worldwide condemnation. The scientific community criticized the experiment as premature and irresponsible, highlighting that society remains unprepared for heritable human genome editing.

Even for somatic therapies, concerns about safety and informed consent remain important. Patients receiving experimental gene editing must fully understand the uncertainties involved, while treatments must prove they will not cause cancer or harmful mutations over time.

Questions of fairness and access are also significant. Gene therapies typically cost enormous amounts, raising concerns that revolutionary CRISPR cures might only reach wealthy people or citizens of developed nations. Ensuring fair access to life-changing treatments represents a moral imperative, with special attention needed for diseases like sickle cell that disproportionately affect disadvantaged populations.

Environmental applications bring additional complexities. Using CRISPR to drive certain genes through wild populations—such as modifying mosquitoes to reduce malaria transmission—carries ecological risks that must be carefully evaluated before implementation.

The Future: Writing Life's Code

CRISPR technology has progressed from obscure bacterial mechanism to revolutionary genetic tool in a remarkably short time. Its applications continue to expand, with each technical improvement opening new possibilities across medicine, agriculture, and biotechnology.

As scientists refine methods like base editing (changing individual DNA letters without cutting both strands) and prime editing (performing precise "search and replace" operations on longer sequences), CRISPR's capabilities grow increasingly sophisticated. The boundary between treatment and enhancement, between correction and creation, grows ever thinner.

Society now faces unprecedented questions about how far we should go in rewriting the code of life. CRISPR provides extraordinary power to heal diseases, feed growing populations, and protect our environment—but using such power requires extraordinary responsibility. The molecular scissors that earned Doudna and Charpentier their Nobel Prize have given humanity the ability to reshape biology itself, a prospect both exciting and sobering as we navigate this new genetic world.

B Additional Details on Experimental Implementation

B.1 Writing Assessment Prompts

This appendix presents the analytical writing prompts developed for each of the three topics. For each topic, we designed two complementary prompts of comparable complexity, one for use in Session One and the other in Session Two. These prompts were carefully constructed to assess participants' analytical reasoning abilities.

Each prompt requires participants to analyze relationships between concepts, evaluate the relative importance of different factors, and support their analysis with specific examples from the reading materials. The prompts were designed to be sufficiently challenging to differentiate between levels of understanding while remaining accessible to participants with no prior knowledge of the topics beyond the provided reading materials.

Blockchain Technology Prompts

- **Prompt 1:** "Examine how blockchain has grown beyond cryptocurrency. Analyze which two application areas have the strongest potential for transformative impact and explain the specific factors that support your analysis. Draw on examples from the reading to develop your response."

- **Prompt 2:** "Examine key differences between blockchain technology and traditional databases. Analyze which specific features make blockchain unique, which types of applications benefit most from these differences, and why. Support your analysis with concrete evidence or examples."

Carbon Capture Prompts

- **Prompt 1:** "Analyze the three main barriers to scaling carbon capture technologies (technical limitations, economic costs, and policy challenges). Identify which barrier you believe is most critical to overcome first, how it impacts the other challenges, and what approaches might address it. Support your analysis with specific examples."
- **Prompt 2:** "Compare direct air capture with point-source carbon capture approaches. Analyze the specific advantages and limitations of each method, identify which contexts each approach is better suited for, and explain what factors most influence their effectiveness. Draw on evidence from the reading to support your analysis."

CRISPR Prompts

- **Prompt 1:** "Analyze how CRISPR differs from previous gene editing technologies in terms of precision, accessibility, and versatility. Identify which two differences have been most consequential for scientific advancement, explain why these particular differences matter, and provide specific examples that demonstrate their impact."
- **Prompt 2:** "Examine the three main technical challenges that limit CRISPR today (delivery methods, off-target effects, and ethical considerations). Analyze which challenge is most important to address first, how it relates to the others, and what approaches show promise for overcoming it. Support your analysis with specific examples."

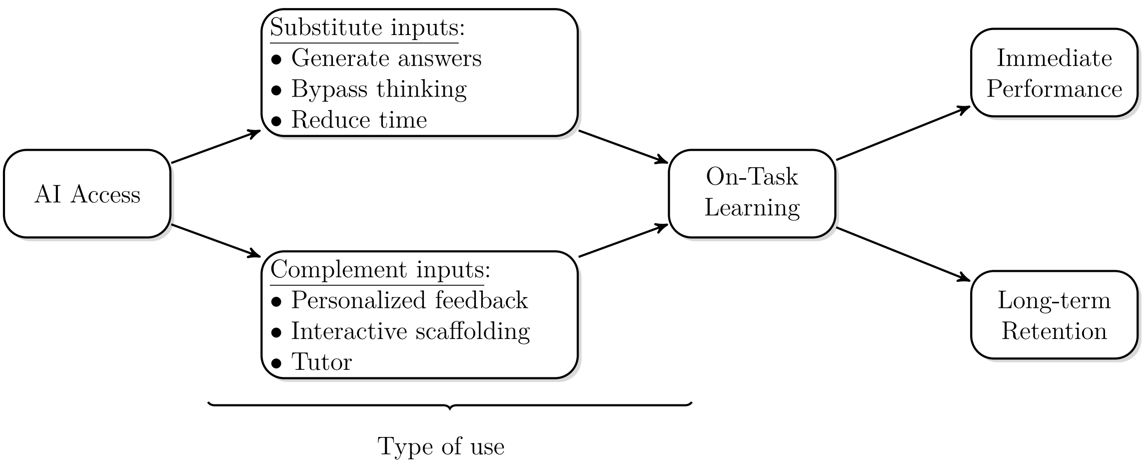
C Conceptual framework

We present a simple framework linking access to generative AI to learning outcomes. The framework builds on research showing that active generation of information significantly improves learning compared to passive consumption (Jacoby, 1978; Slamecka and Graf, 1978; Bertsch et al., 2007). Appendix Figure A1 illustrates these relationships.

The first component captures how access to generative AI influences learning through two potential usage patterns. Students may use AI as a substitute for their own cognitive effort, which can reduce active engagement and lead to passive consumption of AI-generated content. Alternatively, students may use AI as a complement to enhance their learning process through activities like requesting explanations, getting feedback on their work, or using AI to help refine their understanding of difficult concepts.

The second component connects these usage patterns to learning outcomes, which we measure through both immediate performance and longer-term knowledge retention. The generation effect in cognitive psychology suggests that learning outcomes improve when students actively engage with material (McCurdy et al., 2020). Therefore, the impact of AI on learning will depend on whether it promotes or impedes the cognitive processes that underlie durable skill acquisition.

Figure A1: Conceptual framework linking access to generative AI to learning outcomes



References

Bertsch, S., Pesta, B. J., Wiscott, R., and McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, 35:201–210.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R., et al. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain.* Longman New York.

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2):84–95.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of verbal learning and verbal behavior*, 17(6):649–667.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., and Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, 27:1139–1165.

Noy, S. and Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.

Rosner, Z. A., Elman, J. A., and Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex*, 49(7):1901–1909.

Slamecka, N. J. and Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 4(6):592.