Backwards Planning with Generative AI: Case Study Evidence from US K12 Teachers

Samantha Keppler
Stephen M. Ross School of Business, University of Michigan

Wichinpong Park Sinchaisri
Hass School of Business, University of California, Berkeley

Clare Snyder
Stephen M. Ross School of Business, University of Michigan

Abstract. *Problem definition:* Generative AI technologies can help workers do tasks faster and better. However, many workers must plan which tasks to do, in addition to doing them. A prime example are K12 teachers, who plan their teaching tasks (i.e., activities, quizzes, projects) working backwards from end-of-year goals defined by state standards (a process known as *backwards planning*). In this paper, we ask: How are teachers, who must both plan and do tasks, beginning to use generative AI? *Methodology/results:* We conduct a longitudinal case study of 24 US public school teachers (all new to AI), sampled to vary by subject area and grade level. During the 2023–2024 school year, we gather from these teachers 360 minutes of recorded observation of generative AI use for their own work (not predefined by us) involving over 200 inputted prompts (and associated responses), together with 29 in-depth interviews and 34 generative AI use surveys. Analyzing this data corpus, we find that over the year teachers separate into three groups: (1) those who seek generative AI input (i.e., thoughts or ideas about learning plans) and output (i.e., quizzes, worksheets), (2) those who only seek generative AI outputs, and (3) those not using generative AI. The teachers in the first group—but not the second group—report productivity gains in terms of workload and work quality. *Managerial implications:* Workers can use generative AI for known tasks (task-level) or for deciding which tasks to do (workflow-level). Task-level use may save time completing tasks and increase the number of tasks done, but workflow-level use can save time deliberating about which tasks to do and nudge people toward the "right" tasks. AI developers, particularly in the education sector, ought to design tools that go beyond task-level use and nudge people toward more effective and efficient workflows.

Key words: generative AI; productivity; workflow; education operations; algorithm aversion

1. Introduction

Generative artificial intelligence (e.g., ChatGPT) seems poised to fundamentally transform work processes. By automating and augmenting some tasks within a workflow, it could allow people to spend more time on "human" tasks or introduce new activities to the process. However, updating workflows in light of generative AI's capabilities presents a nontrivial challenge (Ghosh et al. 2023). Even without generative AI, people struggle to effectively allocate time between tasks (Kagan et al. 2018) and prioritize tasks (Ibanez et al. 2018). Planning work is difficult, and the choices made

affect worker productivity. As generative AI expands the decision space by creating more tasks and more ways to complete these tasks, planning productive workflows may become more challenging.

While it is understood that generative AI is likely to upend existing work processes, little is known about how generative AI affects workflows. Most of the emerging research focuses on the way people use generative AI at the task-level, meaning for specific, predefined tasks, and the performance outcomes (work quality and speed) that result from their behavior (Noy and Zhang 2023, Brynjolfsson et al. 2023, Dell'Acqua et al. 2023). These studies reveal that generative AI is already advanced enough to improve human work in many cases, although people may not always request or use its outputs. Equally important to task-level research, however, is workflow-level research that examines how people are using, and how they should be using, generative AI within the context of real, discretionary workflows. It is one thing to ask whether and how to use generative AI when it is provided for one mandatory task, but another thing entirely to ask how to use generative AI for a job comprising several optional tasks. For example, people may change the tasks they elect to do when they have access to generative AI.

Even as generative AI's impact on discretionary workflows is not yet well understood, the technology is often positioned as being particularly helpful for such workflows. Teacher work in the K12 education sector is a prime example. Teachers have substantial autonomy over their work (Glazer and Peurach 2015)—including which tasks they work on, the order of these tasks, and the time spent on them—which they typically plan backwards, starting from mandated learning outcomes set by governments (Wiggings and McTighe 2005). Business leaders, including Sam Altman, the CEO of OpenAI, list education among the top areas for productivity gain from generative AI in the near future (Gates 2024). Yet because teacher workflows are discretionary, what does it look like when teachers use generative AI to be more productive? The gap between the theoretical research on task-specific performance and the reality of discretionary, dynamic workflows could partially explain why so few teachers have been trained on and are currently using generative AI tools. For example, in the UK, only about 10% of teachers are currently using generative AI at least once per week, and three quarters of teachers said they needed more training about how to use generative AI tools effectively (Picton and Clark 2024). Because K12 teacher work is a prime example of a profession with discretionary workflows, and because there is so much speculation about the impact of generative AI on teacher work, our study asks: *How are teachers, who must both plan and do tasks, beginning to use generative AI?*

To explore this question, we conduct a longitudinal multiple case study (Yin 2016). We closely follow 24 K12 public school teachers from the upper Midwestern United States during the 2023–2024 school year. We select these teachers for variation by subject area, grade level, and experience (2–27 years). In fall 2023, we first interview these 24 teachers, present them with a generative AI tool

(ChatGPT Plus), and directly observe their early uses of it within their work processes. After this initial contact, we follow up with the same teachers through surveys during the winter and spring of 2024 to understand their evolving generative AI use. We conclude our data collection in June 2024 with follow-up interviews of a subset of the original teachers. The result is a rich longitudinal dataset of real teacher experiences with generative AI in the beginning of this transformational era that includes 360 minutes of recorded observation of teacher use of generative AI for their own work (not predefined by us) with over 200 inputted prompts (and associated responses), together with 29 in-depth interviews and 34 generative AI use surveys. We analyze these data qualitatively and quantitatively to understand evolving task- versus workflow-level use among these teachers and its reported impact on productivity.

We find that all 24 teachers had similarly minimal experience with generative AI at the beginning of 2023–2024, but that this diverged significantly over the course of the school year. In fall 2023, all teachers were either novice users or had never tried any generative AI tool. Every teacher received the same exposure to ChatGPT in our initial interview and observation: a standardized practice session with 12 required prompts, then an unstructured period during which they could generate their own prompts related to teaching tasks of their choice, under our observation. By spring 2024, the teachers separated into three distinct groups: (1) those who seek generative AI's input about how to plan their workflow (e.g., ideas about what tasks to do) and its outputs that help complete specific tasks within their workflow (e.g., quizzes, worksheets), (2) those who only seek generative AI outputs but *not* inputs, and (3) those not using generative AI. We find teachers in the first group—but not the second group—report feeling more productive with generative AI in terms of workload and work quality. The data suggests that productivity gains arise from the first group because generative AI's input nudges teachers toward new and better tasks they may not have done otherwise, resulting in a more efficient and effective workflow. Using generative AI only for outputs (like the second group) occurs when teachers already have a task-to-be-done in mind, and thus the only potential productivity gain is at the task-level. To feel more productive, it seems that teachers used generative AI to help them "get it right" rather than just "get it done."

Our findings contribute new evidence about task- *and* workflow-level use of generative AI to the emerging scholarship about generative AI's impact on human work (e.g., Brynjolfsson et al. 2023), as well as a new perspective to the operational issue of human behavior within modern technology-supported workflows (Ibanez et al. 2018). Although flexibility and autonomy are often a desirable feature of workflows in theory, workers often benefit from exogenously-imposed constraints in practice (Acar et al. 2019). For example, the time it takes for workers to decide how to exercise their workflow discretion can undermine the benefits of worker discretion in a workflow (Kc et al. 2020). Nudges or other guidance about structuring work processes may reduce workers' cognitive

efforts and lead them to more effective task structures when transition decisions are endogenous (Kagan et al. 2018). Within discretionary workflows, our findings suggest that when generative AI is used to plan work (asked to provide *input*) it operates as a nudge for individuals planning their work. Thus, the productivity potential of generative AI is not just at the task level.

Our data and analysis also lead to practical implications. To school and education leaders—and more generally to those managing professions that require both planning and doing work—our findings suggest using generative AI to plan workflows may be particularly important for improving worker productivity. Generative AI's ability to create outputs for a specific task may have limited productivity upside when not highly customized and calibrated, and when it requires the work to refine prompts several times or to make significant changes to those outputs. When users seek input from generative AI about how to plan their work, they are seeking confirmation that they are doing the right and best activities to achieve their desired objectives. To generative AI developers, particularly developers of specialized tools for education and similar settings, our findings suggest the potential value of training models for designing workflows, in addition to more standard material output creation.

2. Literature Review

Here we provide more background into what is currently known about productivity-enhancing technologies such as generative AI, particularly in the education context, and why workflows present an important lens through which to study generative AI.

2.1. Productivity-Enhancing Technologies

Although (generative) AI ostensibly enhances productivity, productivity growth has slowed in the last decade (Brynjolfsson et al. 2019); the relationship between productivity and new technologies, including generative AI, is complicated. New technologies might make people more productive over the long run, but at the cost of lower productivity in the short run as workers learn to integrate it into their workflow (Bhargava and Mishra 2014, Ramdas et al. 2018). New technologies are also unlikely to benefit all workers equally. Information technology (IT) appears to be an equalizer in some cases, boosting the productivity of marginalized workers the most (Ding et al. 2010).

AI algorithms are part of the latest wave of modern productivity-enhancing tools. Even before large language models like ChatGPT became widespread, conventional AI has been used for forecasting and classification tasks. However, beyond learning curves, *algorithm aversion* stands in the way of productivity gains from the technology (Dietvorst et al. 2015). Algorithm aversion is a bias against algorithmic outputs, even when these outputs are as or more accurate than human predictions. People often struggle to know when to use an algorithm's advice (Balakrishnan et al. 2024), or how to do so effectively (Bastani et al. 2024a). While it might seem that people could use

algorithms' advice to speed up their decision-making, deliberation around using the advice at all can add time to the decision, undermining productivity benefits (Snyder et al. 2024). The decision about how to use *generative* AI is even more complicated—because this technology is so broad, it entails not only a binary choice about whether to seek its outputs, but a choice about how to prompt for this output. Effective prompting is nontrivial. Doctors using large language models for a diagnostic task did not outperform doctors without this AI, although the large language model itself did, because doctors did not use the technology to its full potential (Goh et al. 2024).

This study (Goh et al. 2024) highlights a larger theme within human-generative AI interaction research about the way users prompt the technology for specific, experimenter-defined tasks and the resulting impact of this behavior on task outcomes. In another project contributing to this theme, Noy and Zhang (2023) study productivity in a writing task. The authors find that participants assigned to use ChatGPT were on average more efficient. Chen and Chan (2024) find that the productivity gains on ad copy tasks were higher when the nonexpert users asked generative AI to provide feedback than when they used AI as a "ghostwriter." (Brynjolfsson et al. 2023) document productivity benefits from generative AI for real customer service tasks. Still, the nature of generative AI in this study is largely task-specific—it is designed to provide real-time suggestions about how agents should respond to customers, in contrast with more open-ended language models like ChatGPT, which users must think about how and for what to prompt.

It remains to be seen exactly how workers who must plan tasks, such as teachers, integrate generative AI into their workflows, and the productivity effects of the tools in these contexts. Beyond substituting human effort for particular tasks, generative AI could provide more meta advice about workflow organization and planning, serving as co-intelligence to augment human thinking (Mollick 2024). As a coach or tutor, generative AI could improve productivity not only by helping people complete the same tasks more quickly or better, but also to be more strategic about the tasks they complete.

2.1.1. *Artificial Intelligence in Education*

Generative AI is particularly promising for the K12 education in part given the sector's resource constraints. Teachers must commonly find workarounds to supplement classroom resources, including from nonprofit organizations (Keppler 2023) and crowdfunding donations (Keppler et al. 2022); these additional resources improve productivity as measured by student performance (Keppler et al. 2022). In theory, generative AI might improve teacher productivity by reducing the amount of resources (including time or money) required for teachers to achieve their objectives.

However, in practice, the productivity benefits from generative AI are less clear. AI has existed as a tool for educators and students since before the current era of ChatGPT and generative AI,

yet it has not necessarily transformed productivity in the K12 education sector (Ng et al. 2023). Teachers have largely been slow to adopt these technologies. It remains to be seen exactly how educators will use newer, generative AI tools. While there is emerging research in this area (e.g., Lo 2023), much of this work has focused on teaching strategies for student users. In fact, generative AI might reduce productivity as measured by student performance. Bastani et al. (2024b) show that when students' access to generative AI is taken away, these students perform worse than if they never had access to the tool at all, unless the generative AI is prompted to safeguard learning. In particular, generative AI might harm learning when its outputs are incorrect, or when students use the tool to replace their own efforts. Prompting generative AI to take a tutor role prevents such harms and promotes meaningful interactions with the technology (Bastani et al. 2024b).

2.2. Behavioral Nudges for Better Workflows

The design of effective workflows represents a critical operational challenge that predates today's generative AI era (Krishnan and Ulrich 2001); human behavior in discretionary workflows amplifies the complexity of this challenge. For instance, workers with task discretion prioritize easier tasks (Ibanez et al. 2018), particularly as workload increases, at the expense of throughput (Kc et al. 2020). Certain exercises (i.e., making specific implementation plans) might promote better outcomes for a single task or goal, but the same exercises in more complicated workflows have the opposite effect, as they underscore the difficulty of the undertaking (Dalton and Spiller 2012).

Nudges and framing can combat problematic human behaviors (e.g., Dalton and Spiller 2012). In a lab experiment, designers of a new product struggle to transition from one stage of their work process (ideation) to the next (execution), but nudges towards making this transition improve performance—and exogenously-imposed transitions improve performance even more (Kagan et al. 2018). In the field, at a dental clinic, imposing exogenous deadlines, even when these deadlines are not tied to a reward, improves customer response rates (Altmann et al. 2022).

In the education context, established best practices help teachers manage their flexible workflows more effectively. In particular, *backwards planning* is the current standard best practice for teacher work processes in US K12 education. Backwards planning is the process of devising teaching plans that are well-aligned to learning objectives and goals, such as the Common Core State Standards (Common Core State Standards 2010), in order to avoid aimless activities and "coverage" of topics (Wiggings and McTighe 2005). In other words, backwards planning starts with high-level objectives like "I want my students to understand X." and then moves to figuring out the necessary activities to achieve that objective and the assessments that will evidence the achievement (or not) of that objective. Backwards planning is also a common and useful project management strategy outside of education, for example in government and other business contexts (Wiese et al. 2016).

3. Methodology

3.1. Sample

In fall 2023, we recruited 24 US public school teachers from the Midwestern US. As shown in Table 1, we selected the sample to include a variety of teachers along theoretically important dimensions: different grade levels (elementary, middle, high), subject areas (math, science, ELA, social studies, foreign language, elementary education), and years of experience (from 2 to 27). Random samples are not appropriate for case study research; theoretical sampling like ours is a best practice (Glaser and Strauss 1967, Eisenhardt 1989). We identified our sample through, first, reaching out to five participants (indicated by * in Table 1) through pre-existing relationships with the authors of this paper (personally known public school teachers). We then recruited the remaining participants (19/24) through emails sent to 83 K12 public school teachers in southeast Michigan based on subject area and grade level sampling goals. All teachers who responded to our recruitment email were included in our study. Consistent with the logic of the multiple case study methodology, our theoretical sample is designed to maximize the breadth and richness of information gathered across "polar types" meaning teachers of different grades, subjects, and experience levels (Eisenhardt 1989). Our sample is not representative and cannot be used to infer general behavior or patterns of the wider population. Evidence from our small sample is valuable to the operations management field for idea generation, novel theorizing, and innovation (Fisher 2007).

3.2. Choice of Generative AI Tool

We focus primarily on teacher use of ChatGPT. ChatGPT is one of many productivity-enhancing generative AI tools, but it is "by far the most widely recognised" one (Fletcher and Nielsen 2024). This was particularly true when our study began, in fall 2023. None of the teachers in our sample worked at a school or in a district that formally adopted a generative AI tool during the 2023–2024 school year. While teachers were highly experienced with virtual learning platforms like Google Classroom and Kahn Academy, these tools belong to an older wave of AI. Of the new wave of generative AI, ChatGPT was most well-known (Diliberti et al. 2024). For these reasons, we opted to conduct our initial data collection during fall 2023 on teachers using ChatGPT (specifically a paid ChatGPT Plus account). After talking with the sampled teachers, we confirmed that ChatGPT indeed was the most familiar form of generative AI. All of the teachers in our sample had heard about ChatGPT, even if they had never used it. Teacher 6, for example, shared "I've never used it" but that "I hear about it everywhere."

Still, we used open-ended interview questions to ask teachers about their generative AI use more generally. Further, as education-tailored tools like SchoolAI and MagicSchool emerged over the school year, we adapted the language in our surveys to ask teachers about the other generative AI tools they use, in addition to or instead of ChatGPT.

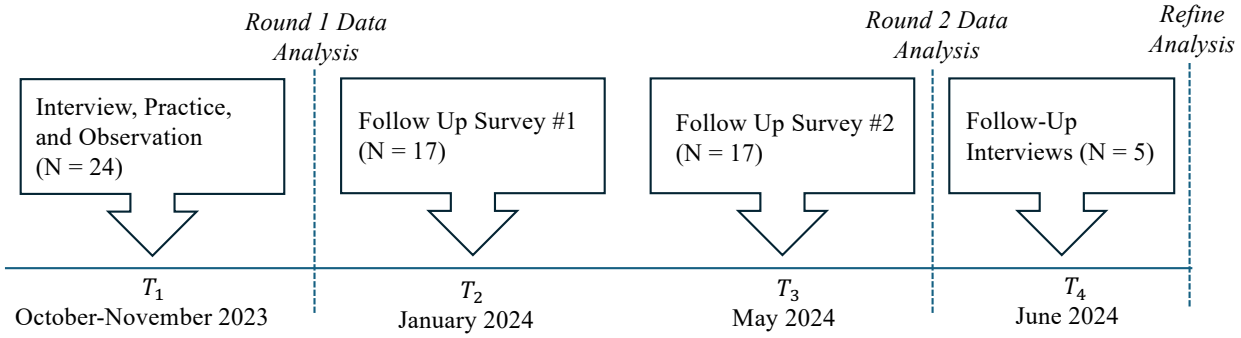
Table 1 Teacher Participants

ID	State	Grade Level	Subject Area	Experience Teaching (yrs)	T_1	T_2	T_3	T_4
T1*	Ohio	Middle School	General	12	Yes	No	No	No
T2	Michigan	High School	ELA	22	Yes	Yes	Yes	Yes
T3*	Pennsylvania	High School	Math	7	Yes	Yes	Yes	No
T4	Michigan	High School	ELA	21	Yes	Yes	Yes	Yes
T5*	Pennsylvania	High School	Foreign Language	12	Yes	Yes	Yes	No
T6*	Pennsylvania	Elementary School	General	25	Yes	Yes	Yes	No
T7	Michigan	Elementary School	General	6	Yes	Yes	No	No
T8	Michigan	Elementary School	General	12	Yes	Yes	Yes	No
T9	Michigan	Elementary School	General	3	Yes	No	Yes	No
T10	Michigan	High School	Science	25	Yes	Yes	Yes	No
T11*	Pennsylvania	Middle School	ELA	17	Yes	Yes	Yes	Yes
T12	Michigan	High School	Math	22	Yes	Yes	Yes	No
T13	Michigan	Middle School	Science	2	Yes	No	No	No
T14	Michigan	High School	Foreign Language	11	Yes	Yes	No	No
T15	Michigan	High School	Foreign Language	8	Yes	Yes	Yes	No
T16	Michigan	High School	Foreign Language	14	Yes	No	No	No
T17	Michigan	High School	Science	8	Yes	No	No	No
T18	Michigan	High School	Math	12	Yes	No	No	No
T19	Michigan	Middle School	Social Studies	27	Yes	Yes	Yes	No
T20	Michigan	High School	ELA	20	Yes	Yes	Yes	No
T21	Michigan	High School	Social Studies	9	Yes	Yes	Yes	Yes
T22	Michigan	High School	Math	2	Yes	Yes	Yes	Yes
T23	Michigan	High School	Math	27	Yes	No	Yes	No
T24	Michigan	High School	Science	3	Yes	Yes	Yes	No
Total Number of Teachers Per Round of Data Collection:					**24**	**17**	**17**	**5**

3.3. Data Collection

We collect multiple types of data at multiple points in time. There are four points of data collection—T_1, T_2, T_3, T_4—as shown in Figure 1. Below, we discuss each data collection effort in detail. We contacted all 24 teachers in T_1, T_2, and T_3, and only a subset of the original set of teachers in T_4. Which teachers opted to participate at which time is indicated in Table 1.

Figure 1 2023–2024 School Year: Data Collection and Analysis Timeline



3.3.1. *Initial Exposure, Observation, and Interview (T_1)*

We collected multiple types of data from all 24 teachers during a 1-hour one-on-one Zoom session conducted in October or November 2023 (T_1). With a shared semi-structured protocol (Appendix A) one member of the research team did 13 of the interviews and another member did 11 interviews, which is a best practice in qualitative data collection that among other things can enhance validity when common patterns are found from different interviewers. All teachers earned $50 for participating. With each teacher's consent, the audio and video of each Zoom call was recorded and the audio transcribed using an automated online transcription service (Sonix). The hour-long session involved four parts.

Background Interview (15 minutes): Each session began with a background interview where teachers were asked about their workflows for material creation (colloquially described as "planning" or "prep"). We asked about how they spend their days or weeks preparing for teaching, and the types of tasks involved. We also asked about whether they do their planning alone, or with other teachers. We did ask about their experience with ChatGPT or other generative AI tools at this time, and all had very limited to no experience using the tool for their work. Only two teachers (Teachers 13 and 22) mentioned receiving any training on generative AI before our initial session. Thus, all teachers in our sample should be understood as new generative AI users at T_1.

Structured Exposure to ChatGPT (10 minutes): Next, we asked teachers to log into our ChatGPT Plus (GPT-4)[1] account on their own computers and share their screen. After this, each teacher engaged in a structured practice session with ChatGPT where they entered our pre-designed 12 prompts (Table 2) one-by-one in our prescribed order (we sent each prompt by chat individually, after which the teacher copied it into their ChatGPT window). Our pre-designed exposure prompts are consistent with similar prompts used in contemporary studies of ChatGPT (Chen et al. 2023, Noy and Zhang 2023). The rationale of this standardized exposure is to make sure all teachers in our sample have a shared foundational understanding of what ChatGPT can and cannot do.

Observation of Teachers' ChatGPT Use (15 minutes): The third part of the interview was the observation period. We directed teachers to use ChatGPT for 15-minutes for their actual teaching

[1] Our data collection effort overlapped with OpenAI's first DevDay, which took place November 6, 2023. At that time, ChatGPT Plus was upgraded and became an integrated multi-tool that could interpret more input formats and create more output formats. For example, after the update, ChatGPT Plus could search the internet via Bing and input information from attached documents. It could also output textual data in tabular format and images. This update was rolled out to users (without specific advance notice) following the announcement, and our account updated on November 7, 2023 between 11:15am and 2:15pm. As it happened, four teacher sessions (T14, T15, T16, and T17) were conducted on November 7. The first two (T14 and T15) took place in the morning and were unaffected. The later two (T16 and T17) represented our first exposure to ChatGPT Plus' altered capabilities. The dashed line in Table 1 indicates the separation between the before-versus-after interview and observation sessions. In response to the update, we slightly modified three of our practice prompts to give subsequent teachers (T18 through T24) practice with its newest features. Table 2 marks the updated prompts with '. For example, we updated prompt 11 from "Design a simple workout plan for beginners" to "Design a simple workout plan for beginners and present it in table form" to evidence that ChatGPT Plus could now output text organized in a table.

Table 2 Required ChatGPT Prompts During the Structured Practice

No.	Prompt
1.	What is GPT-4?
2.	Is 17077 a prime number? Think step by step and then answer.
3.	What are today's top news headlines?
4.	What notable events happened on February 30, 2020?
5.	What notable events happened on February 29, 2020?
6.	Explain the economic impacts of the COVID-19 pandemic.
7.	Help me write an introductory paragraph for an essay on this topic.
8.	Rewrite the paragraph using simpler language.
9.	Summarize 'Pride and Prejudice' in one paragraph.
9'.	Summarize this text in one paragraph. (upload PDF - Chapter 43 of *Pride and Prejudice)*
10.	Please give the same summary as a rhyme.
11.	Design a simple workout plan for beginners.
11'.	Design a simple workout plan for beginners and present it in table form.
12.	Design a simple workout plan for beginners with limited free time.
12'.	Give a diagram of the proper form for one of these exercises.

work. We suggested to each teacher to "Pick one of the things you mentioned earlier for which you might use ChatGPT to help. Work as if you are trying to create the "finished product" in 15 minutes. You are welcome to use other technology in addition to ChatGPT such as Google Docs, Word, Excel, a web browser, etc. It's okay if you are unable to finish, just work like you'd typically work. Remember, the final product may be included in a publication as an example of how teachers use ChatGPT, so please try your best." We then set a timer for the teacher, and remained silent as they worked (some teachers narrated as they worked and in that case we engaged minimally). This observational design is an established approach in education research to examine how teachers plan, create, and modify materials on their own, without ChatGPT (Silver 2022).

Debrief Interview (10 minutes): At the end of the interview, we debriefed the teachers' experience with ChatGPT. We asked specific questions about the intention behind their prompts and their evaluation of the ChatGPT output during the observation. When time allowed, we asked teachers to share their thoughts about ChatGPT and how it might affect their work more generally.

3.3.2. *Generative AI Use Survey (T_2)*

The second wave of data collection was done in January 2024 (T_2) through a survey sent to the same 24 teachers that participated at T_1. The survey included both open and closed (Likert-style) questions about the *frequency* and *mode of* ChatGPT use in the months since T_1. We also asked about other generative AI tools the teachers were using at that time. The full survey is provided in Appendix B. We asked specifically about how teachers were using ChatGPT or another generative AI tool to *make*, *find*, *jumpstart*, and/or *iterate* on their teaching plans. We derived this four-type categorization from an analysis of the evidence from our unstructured observation period at T_1 (See 3.4.1 for details). This means that we conducted one round of analysis of our data before all data

collection was complete. The iteration between data collection and data analysis is a common and in fact recommended best practice for case study research because it allows iterative refinement of emerging ideas (Yin 2016). The objective of the T_2 survey was to assess whether and how the *potential* uses of generative AI identified during our initial T_1 exposure and observation period materialized into real use *in practice*. Statistical analysis and extrapolation to a wider population was not an objective. For completion of the survey, teachers were compensated $10. 17 of the 24 (71%) of the original teachers responded to the January survey (See Table 1).

3.3.3. *Generative AI Use Survey (T_3)*

The third wave of data collection was done in May 2024 (T_3) through another survey sent to the same 24 teachers that participated at T_1. The survey contained the same questions as the January 2024 survey, but also included additional Likert-style questions about how generative AI is impacting their own learning, stress, number of tasks completed, hours working per week, and work quality. We asked separately about these impacts of generative AI for their teaching, preparation, grading, and emailing. The additional May 2024 survey questions are provided in Appendix C. These new questions aim to identify who among our sample reported productivity gains (or losses) from generative AI at that time. For completion of the slightly-longer survey, teachers were compensated $15. 17 of the 24 (71%) of the original teachers responded to the May survey (See Table 1).

3.3.4. *End-of-Year Interviews (T_4)*

We conducted five teacher interviews in June 2024 (T_4). These five teachers indicated in their May 2024 survey (T_3) that they have more they would like to share about their generative AI use. For example, at one teachers' school (Teacher 4), there was a training done on generative AI and a lot of discussion about student use of generative AI at the end of the school year. Another teacher (Teacher 21) wanted to talk to us more about their conscious decision to *not* use any generative AI. Among the five teachers interviewed at T_4, one reported productivity improvements in the May 2024 survey, two were strong non-users, and two were in the middle. We customized our interview questions based on these differences, as shown in Appendix D. These final interviews provided us an opportunity to identify why there might be different in-practice use of generative AI despite consistent initial exposure at T_1.

3.4. Data Analysis

During the 2023–2024 school year, we gathered in total from teachers 360 minutes of recorded observation of teacher use of generative AI for their own work (not predefined by us) involving over 200 inputted prompts (and associated responses), together with 29 in-depth interviews and 34 generative AI use surveys. On this rich data corpus, we conducted both qualitative and quantitative analyses. We describe these analysis in turn.

Table 3 Prompt Coding Scheme

Code	Description
Make for me	Requests for fully-developed content (e.g., problems, quizzes, essays, poems, images) within well-defined parameters; user engages ChatGPT as a task executor
Find for me	Informational requests seeking pre-existing, factual information (e.g., existing facts, quotes, resources, or examples); user engages ChatGPT as a search engine
Jumpstart for me	Requests to initiate the development of often-lengthy and complex materials like activities, projects, lessons, or unit plans; user engages ChatGPT as a catalyst
Iterate with me	Requests for advice, or to understand/refine/re-think concepts or teaching approaches; user engages ChatGPT as a sounding board like they would a teaching colleague; additionally these prompts are often identifiable by words like "explain," "discuss," or "describe"

3.4.1. *Round 1 Data Analysis (Between T_1 and T_2)*

Our first round of data analysis focused narrowly on coding the prompts teachers wrote and submitted during the unstructured observation period in T_1. The objective was to define a use typology grounded in teachers' own prompting choices. We recorded all prompts written and sent by each teacher during the 15-minute unstructured observation period in a large spreadsheet. We did this by watching the recorded videos back, and copying the prompts verbatim into the spreadsheet. The 24 teachers inputted 201 prompts in total (8.4 per teacher, on average). We then iteratively derived a coding scheme of different generative AI use cases, first through open coding prompts and then refining descriptions of similar prompts to match. That is, we would take pairs of prompts and discuss their similarities and differences, and their appropriate categorization. In the end, we settled on a four-category prompt coding scheme: (1) *make for me* (55% of prompts), (2) *find for me* (15% of prompts), (3) *jumpstart for me* (10.5% of prompts), and (4) *iterate with me* (15.5% of prompts). We used a fifth category, *show me what you can do*, for non-teaching-relevant requests testing ChatGPT's capabilities (4% of prompts). Table 3 gives descriptions of each code. Two members of the research team independently coded all 201 prompts. Even prompt was assigned only one code. The inter-rater reliability for the coding was 91%. Through discussion among the coders, different coding categorizations were resolved and in the end, all authors were in agreement. More details on this coding can be found in BLINDED.

This coding scheme was used to create the survey questions used at T_2 and T_3. That is, we asked teachers specifically how often they used generative AI to *find for me*, *make for me*, *jumpstart for me*, and *iterate with me in practice* (See Appendix B for exact questions).

3.4.2. *Round 2 Data Analysis (After T_3)*

After T_3, we conducted comparative case study data analysis to understand differences among teachers. In this round of analysis, we analyze data at the teacher level and treat each teacher

as one case of exploring whether and how to use generative AI. We follow the standard analysis steps of multiple case study research: (i) within-case analysis, (ii) across-case analysis, and (iii) theory-case analysis (Eisenhardt 1989).

Within-case analysis: The objective of within-case analysis is to develop an understanding of each of the 24 teachers' generative AI trajectories. We constructed a longitudinal database that compiled the data across multiple time points (T_1, T_2, and T_3). This allowed us to connect, for example, each teacher's prompts inputted during the observation at T_1 to their ongoing use at T_2 and T_3. In particular, the outcomes at T_3 related to their productivity, stress, and work quality we connected to their generative AI use originally and over time. The database also included each teacher's qualitative descriptions of their work and generative AI use. We also add numerical columns for each teacher's attributes such as years of experience, grade level, subject area, and whether they worked independently or as part of a teaching group.

Across-case analysis: The objective of across-case analysis is to compare pairs or subsets of cases with one another with a goal of first identifying and then refining emerging patterns. We separate the teachers by their productivity outcomes in the May 2024 survey (T_3). We code a teacher as reporting improved productivity if they indicate that as a result of generative AI they are (i) doing more tasks in less time, (ii) doing the same amount of tasks in less time, (iii) doing fewer tasks in less time, *or* (iv) doing more tasks in the same amount of time in one or more of the following areas of their work: teaching, preparation, grading, and emailing. Six teachers met this criteria. Note that although it was possible for teachers to report productivity gains even if they do not use generative AI themselves, this was not the case for these six teachers; all used generative AI at least once per month in May 2024. All six teachers also reported greater work *quality* as a result of using generative AI. The teachers who only reported increased *quality* and did not indicate improved productivity were not included in this group of six "more productive with generative AI" teachers. Among the other 11 teachers (recall, we only have T_3 survey responses from 17 of the 24 original teachers), five were non-users (they reported never using generative AI for their work) and six used generative AI but did not report productivity gains. This formed three distinct groups of teachers for our comparative, across-case analysis: improved-productivity teacher users, no-change teacher users, and non-users. The teacher grouping is reported in Table 4.

The next step was to compare the three groups of teachers to understand the different outcomes. It was curious to us that teachers who were very similar in fall 2024 when we began our study—novice users, little knowledge of generative AI, curious about the technology—and who all got a similar intervention from us in terms of the standard prompting and practice period could evolve over the same time period into distinct groups. We made conjectures about why, and tested them through a "logic of replication" (Eisenhardt 1989) with our teacher cases where we verified if a

Table 4 Teacher Group Categorization at T_3 (May 2024)

ID	Productivity at T_3	Work Quality at T_3	User or Non-User at T_3	**Categorization at T_3**
T2	Decreased	No Change	Non-User	Non-User
T3	Increased	Increased	User	Improved Productivity
T4	No Change	Increased	User	No Change to Productivity
T5	Increased	Increased	User	Improved Productivity
T6	No Change	No Change	Non-User	Non-User
T8	Increased	Increased	User	Improved Productivity
T9	Increased	Increased	User	Improved Productivity
T10	Increased	Increased	User	Improved Productivity
T11	Increased	Increased	User	Improved Productivity
T12	No Change	No Change	User	No Change to Productivity
T19	No Change	Increased	User	No Change to Productivity
T20	No Change	No Change	Non-User	Non-User
T21	No Change	No Change	Non-User	Non-User
T22	Decreased	No Change	Non-User	Non-User
T23	No Change	Increased	User	No Change to Productivity
T24	No Change	Increased	User	No Change to Productivity

hypothesized relationship consistently held true for all teachers in our sample. For example, one conjecture was that teachers' experience in the classroom could explain this divergence. However, in the *improved productivity* group the experience level ranges from 3 to 25 years, in the *no change* group ranges from 3 to 29 years, and in the *non-user* group ranges from 2 to 22 years. While we cannot conclude if the hypothesized pattern might exist more generally, it does not seem to explain the productivity differences we observe.

We compared teachers by *how* they used of generative AI, starting with the observation period at T_1. Perhaps teachers who reports productivity improvements and those that do not approach generative AI differently. This was not the case at T_1. All groups—improved-productivity users, no-change users, and non-users—had inputted a similar variety of prompts at T_1 as defined by the coding in Round 1 (*make*, *jumpstart*, *find*, and *iterate*). The average number of different prompts entered by teachers in each of the groups was 2.5, 2.3, and 2.6, respectively. A clustering and correlation analysis reveals significant differences based on teaching experience but no differences across subject areas. Teachers with greater experience displayed a more streamlined approach to generative AI. They engaged in fewer prompts per session but initiated more separate interactions over time, indicating that experienced teachers segmented their tasks more effectively. By contrast, less-experienced teachers tended to submit higher numbers of prompts per session, often experimenting to explore AI outputs. Experienced teachers also demonstrated a preference for iterative use of generative AI, using it as a planning partner to refine ideas rather than as a direct content generator. Contrary to expectations, subject area (e.g., STEM versus non-STEM) did not significantly affect usage patterns. Teachers across both STEM and humanities domains engaged with generative AI in similar ways, reinforcing the versatility of generative AI tools for supporting content creation and workflow design across disciplines. Overall, though there were some differences

Table 5 Change in Teacher Generative AI Use from T_2 to T_3 (January to May 2024)

Frequency of Use Mode *(Scale 1 to 5, 1 = Never, 5 = Weekly)*	**Improved Productivity (N = 6)**	**No Change (N = 6)**
Make For Me	Increase (Avg. + 0.2)	Increase (Avg. + 1.0)
Find For Me	Increase (Avg. + 0.2)	Decrease (Avg. - 0.2)
Jumpstart For Me	Increase (Avg. + 1)	Increase (Avg. 0.4)
Iterate With Me	Increase (Avg. + 1.2)	Decrease (Avg. -0.2)
Overall*	Increase (Avg. + 0.6)	Increase (Avg. + 0.4)

**Overall change in frequency of use is a separate question asking teachers their overall use frequency, and is not derived as a total of the change in usage of each type (make for me, find for me, jumpstart for me, and iterate with me).*

by years of experience, it seemed that early on in our controlled observation setting, all teachers were exploring how to generative AI.

We looked therefore closely at their reported use in January and May to see how it evolved in practice (T_2 and T_3). In doing so, we uncovered that there was a particular divergence between improved-productivity and no-change users in January and May. As shown in Table 5, both improved-productivity and no-change users similarly increased the frequency with which they used generative AI *overall* (by 0.6 and 0.4, respectively). However, that increase derived from different modes of use. No-change teacher users largely increased their reported use of generative AI to *make*, whereas improved-productivity users largely increased their reported use of generative AI to *iterate* and also *jumpstart* work. In fact, *iterate* evidenced the largest divergence between the two groups in terms of their evolving reported usage. This provided suggestive initial evidence of why some teachers were reporting productivity gains whereas others were not.

We went back to the *iterate with me* and *jumpstart for me* prompts teachers entered into ChatGPT during our observation period to identify what about these types of requests could stimulate productivity improvements. Specifically, we conducted textual analyses on the 200 prompts entered during our observation period at T_1 to further understand patterns that might explain productivity differences, such as the specificity of the prompt. Interestingly, we find that for task-level generative AI use, a highly specific prompt is important for enduring reliable outputs that reduced post-processing time. Yet, for workflow level use that involves task planning, teachers appeared to value variability, as diverse outputs allowed for a broader exploration of ideas. We report these textual analyses in more detail in Section 4.2. We also read through the open-ended examples from the surveys at T_2 and T_3 about how teachers were using generative AI to iterate, make, find, or jumpstart. Through these analyses, we came to understand that relying on generative AI to iterate or jumpstart was similar to asking generative AI for *input* into their teaching plans, whereas using generative AI to make or find (but mostly to make) largely involved creating *outputs*. For example, Teacher 1 prompted ChatGPT, "Can you explain how to add a negative number and a positive number" and then "What manipulatives besides a number line can I use?" As a result of the *input*

from ChatGPT, she decided to revise her original workflow: "maybe I make this a word bank and print it off on a half sheet, and they can keep that in their math notebook right there doing their practice problems." She explained that when it comes to deciding how to achieve a learning goal when something has not been working is "a lot of times where I get stuck." This approach is different *from a workflow perspective* than prompting ChatGPT for specific material outputs, for example prompting "Make a word bank for students for word problems so they know what sign (positive or negative) to use." Overall, the insight from across-case analysis was that teachers who reported productivity gains with generative AI (compared to those who did not) seemed to be using generative AI more at the workflow-level for input on their work plans, in addition to the task-level for specific materials outputs.

Theory-case analysis: Theory-case analysis is the final analysis step in case study research: integrating emerging insights with the existing knowledge and theory. With the insight that the improved-productivity teacher users in our sample seemed to be using generative AI for input into workflow plans as well as for specific outputs, we turned to the literature on workflow planning and design (Ibanez et al. 2018). Comparing our own evidence to the literature, we identified an important feature teacher workflows: they stem backwards from *state-mandated learning objectives*. (See Appendix E for examples of state learning objectives; there is no direction on *how* or *what* to teach, just what students need to be able to do in the end.) This pointed us to a related literature on backwards planning (Wiggings and McTighe 2005, Wiese et al. 2016). We adopted a backwards planning perspective to help situate our evidence within the wider scholarly knowledge, and generate knowledge about the role of generative AI in teaching work.

3.4.3. *Additional Refinement (T_4)*

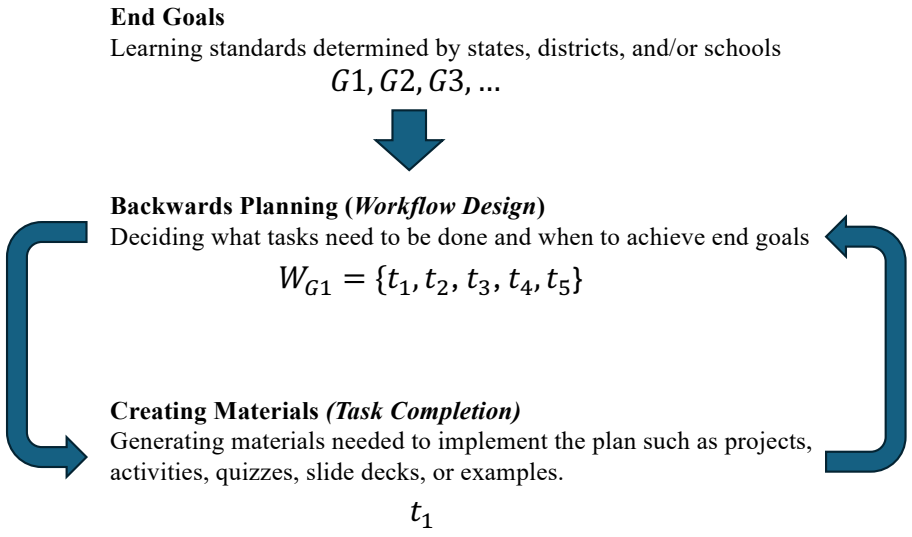
With the insight from our data collection that there was heterogeneity among users of generative AI, and that some teachers were non-users of generative AI, we leveraged a small set of five additional follow-up interviews to refine our understanding of *why*: improved productivity user (T11), no change user (T4), two always non-users (T2 and T21), and one who initially used generative AI in January but then became a non-user by May (T22). We integrated the evidence from these interviews into our teacher database. The interviews were particularly instructive about non-use of generative AI, and learning why a subset of teachers in our sample were choosing that option at this time. From interviews with the two teachers using generative AI, we identified differences mostly consistent with our emerging theory. The improved-productivity teacher explained how she was using generative AI for both input and outputs, while the no-change teacher user spoke at length about the increased stress and frustration about student use of generative AI since January 2024 (T4 is a high school ELA teacher, a subject and grade level particularly affected by unauthorized student use). This helped us refine our emerging ideas to incorporate contextual factors that may amplify or hamper each teachers' motivation and ability to use generative AI for their work.

4. Findings

4.1. Using Generative AI at the Task versus Workflow Levels

Backwards planning is a process illustrated in Figure 2. It begins with end goals $G1, G2, G3, ...$, such as the learning standards in Appendix E. With these goals in mind, a teacher works backwards by first planning the set of tasks necessary to accomplish that goal. This set of tasks is a *workflow* i.e., $W_{G1} = \{t_1, t_2, t_3, t_4, t_5\}$. Once the workflow is planned, teachers then move to complete a specific task, t_i, within the workflow. For teaching, tasks often (but not always) involve material creation, such as activities, lessons, discussions, quizzes, or slide decks. There is a return arrow between creating materials (task-level execution) and backwards planning (workflow planning) because teachers regularly make revisions to their original plans. As tasks are completed, teachers gather information about progress toward goals. This can happen, for example, when teachers grade a quiz. At that point, teachers may adapt their original backwards plan by adding more tasks (i.e., $W'_{G1} = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7\}$ and/or modifying original tasks (i.e., $W'_{G1} = \{t_1, t_2, t_3, t'_4, t'_5\}$. The iterative nature of planning-then-doing in K12 teacher work means that teachers are frequently (re-)planning workflows. As Teacher 6 put it, *"I'm constantly on my feet, revising mentally or having to take a look at the upcoming common assessments."* Thus, the cycle shown in Figure 2 is short and tends to occur weekly, or even every few days, in K12 teaching work.

Figure 2 Backwards Planning Workflow



In our initial interviews with teachers at T_1, we learned 92% of the 24 teachers in our sample manage *backwards planning*, 96% are involved in *material creation*, and 100% are involved in *modifying workflows*. The few teachers who are not involved with original planning and creation get materials from other teachers, and modify them as needed. 42% of the 24 teachers in our sample

provide their plans and created materials to other teachers. Overall, this evidence indicates the teachers we study have a high degree of ownership and involvement in their task planning.

Common Use: Task-Level AI Use For Creating *Outputs*. All teachers in our study that report using generative AI do so to create outputs necessary for completing their work tasks. When teachers use generative AI in this way, they already have the workflow in mind (i.e., $W_{G1} = \{t_1, t_2, t_3, t_4, t_5\}$) and ask for generative AI support on a particular task (i.e., t_1). Consider the prompting shown in Table 6 from our initial observation period where teachers asked generative AI for task-level support. Teacher 3, for example, requests generative AI to make an in-out table he can use in an in-class activity. Teacher 5 requests multiple choice questions. In the observation period, 80% (19/24) teachers requested generative AI support specifically for making something.

Table 6 Examples of Generative AI Prompts For Outputs

Teacher ID	Output Prompting Examples from T_1 Observation Period
3 (HS Math)	• Can you give me a puzzle where I have to find the next thing in a visual pattern. • Give me an in-out table where the input is not a number but the output is a number. • Make it a little more complicated. • Make the input not words. • Make it less complicated.
5 (HS Spanish)	• Make 10 multiple choice questions for chapter 7 for the story *Mi Proprio Auto*. • Make 10 multiple choice questions using the subjunctive for the story *Mi Proprio Auto* with an answer key.
6 (Elementary)	• Create a 5 question quiz for 3rd grade on the topic of forces and motion.
12 (HS Math)	• Write a calculus question that would make it clear that a student understands how to find an absolute maximum
24 (HS Science)	• Make a lab for 9th grade students using the following criteria: [attached a long list of criteria copied from course notes about a lab launching marshmallows] • Give more guidance on the data collection and organizing section in day 3. • Make slides for the teacher to present this all to the class. • Add diagrams to the presentation. • Instead of diagrams, add drawings of students doing the lab to the slides. • Give a diagram of a student participating in this lab. • Give a diagram of an adult participating in this lab. • What about this is not aligned with the content policy for images? • Give a diagram of a person demonstrating this lab.

Using generative AI to help create outputs was also evident in the follow-up surveys. *Make for me* support was the most frequently reported way teachers were using generative AI in practice in both the January and May 2024 surveys, on average. One teacher we interviewed in June (T11) explained why generative AI's ability to support teachers for particular tasks within workflows is helpful: *"The way we usually teach, we introduce the skill on day one. And then we do practice throughout the week. So with [the curriculum] only providing us one general worksheet, it's not very helpful. So I had to use [ChatGPT] quite frequently, like, 'oh, create a worksheet on this for fifth*

grade'. And then if I didn't like it, I just ask it in a different way." On the May 2024 survey, teacher (T15) shared, *"I have [ChatGPT] make vocab lists from readings or to generate questions to intro topics. I can use it to make images. I also ask it to make questions from chapters or readings."*

Using generative AI to create outputs within backwards-planned workflows is similar in process to using generative AI for task completion in other workflows, though its implications may be different. Much of the existing research on generative AI has compared worker productivity and quality on particular tasks with and without generative AI (i.e., Dell'Acqua et al. 2023, Brynjolfsson et al. 2023, Chen et al. 2023) and found a significant improvement in workers using generative AI. While all the teachers in our study who use generative AI use it in a similar way to participants in other studies, only some of the teachers we study report productivity gains. No-change teacher users are using generative AI to *make* something at almost the same frequency as improved-productivity teacher users (3.5 versus 3.7 on a scale of 1 to 5, respectively). The similarity in the way teachers are using generative AI for this purpose, but different productivity outcomes, suggests that the productivity outcomes may be explained by a different use case.

Different Use: Workflow-Level AI Use For *Input* into Work Plans. A subset of teachers in our sample sought generative AI for *input* into their work plans, in addition to asking for *outputs*. Table 7 gives examples of this from the observation period at T_1. As the examples show, teachers sometimes shift from planning to material creation (such as when Teacher 18 asked for help creating a problem set). The important difference compared to examples in Table 6 is the initiation of the conversation with learning standards and objectives in mind, rather than a particular task.

During the observation period, 54% (13/24) of teachers sought input from generative AI in at least one instance. Yet, in January 2024, only nine teachers reported using it in this way. For example, one teacher (T19) shared she used ChatGPT to help her figure out *"how to teach dividing fractions."* Yet, another teacher shared that while she used it for input during the observation period, *"I don't use ChatGPT in this way"* in practice (T8). By May 2024, there is even more divergence. For example, only eight teachers reported using generative AI to *iterate*, with four teachers reporting that they never use generative AI to iterate on their work plans and four teachers reporting they nearly always do. One of the frequent users of generative for this purpose, Teacher 3, explained how he *"typically starts a prompt with 'create a high level plan …' so it will give me an outline of steps required for a project and things to consider at a high level."* Overall, while asking generative AI for outputs became common among all teacher users over the 2023–2024 school year, asking generative AI for input became frequent for some teachers but never happened for others.

The teachers who report using generative AI for both input and outputs are the same group of teachers who report productivity gains. In May 2024, improved-productivity teachers report using generative AI to iterate on teaching plans and ideas nearly twice as frequently as no-change

Table 7 Examples of Generative AI Prompts For Input

Teacher ID	Input Prompting Examples from T_1 Observation Period
1 (MS Special Ed)	• Can you explain how to add a negative number and a positive number? • Create real world math problems within 100 that uses this concept. • Add in multi-step word problems. • Real word examples using numbers within 20. • What manipulatives besides a number line can I use? • My student doesn't understand how a positive and a negative number added can still be negative. Can you help me explain?
4 (HS ELA)	• Describe standards by which a "great American novel" is determined. • Describe novels contemporary to *Adventures of Huckleberry Finn* that reflect similar social and cultural issues. • Describe novels contemporary to *The Great Gatsby* that reflect similar social and cultural issues. • Describe essays, pamphlets, and books contemporary to *The Great Gatsby* that could inform a student's understanding of the Jazz Age and/or class stratification in the United States in the early 20th century. • Discuss how Thorstein Veblen's *Theory of the American Leisure Class* illuminates social mores shown in *The Great Gatsby*.
18 (HS Math)	• Construct a unit plan for teaching the binomial theorem at the Math HL level (use the IB syllabus). List in a table form including key concepts. • Can you expand on week 3? What are some good examples? • You say 'Example: Using the binomial theorem in problems involving physics, like calculating the potential energy in a spring system.' Can you give an example of such a problem • Can you come up with an application where n is far greater than 2? • Can you write a problem set with 4 questions related to the binomial theorem. One should be a word problem. • Can you write a question that involves both trig identities and the binomial theorem used in conjunction? • Can you rewrite that question in a 'show that' form? • Produce a solution key to your problem

teachers (average of 3.2 versus 1.7, on a 1 to 5 scale), even as they report more similar overall use (average of 3.8 versus 3.3, respectively). Taken together with the results that using generative AI for material creation alone does not seem to produce productivity gains, our findings stimulate two hypotheses. The first is that generative AI's input—but not its output—generates productivity improvements for teachers. The second is perceptual: that teachers do not *feel* more productive when using generative AI to create outputs. Either way, there are implications of our findings for understanding how to integrate generative AI into backwards, goal-oriented workflows.

4.2. Additional Evidence of Workflow versus Task Level Use of Generative AI

Through additional survey analyses (N = 34) and textual analyses of observed teacher-inputted generative AI prompts (N = 200), we consistently unearth a productivity-relevant difference between workflow- and task-level generative AI use.

4.2.1. *Relationship between Teachers' Goals and Productivity Gains*

We analyze the role of goal clarity during the generative AI observation period at T_1 and its relationship with reported productivity outcomes in T_3. To assess goal clarity, we asked teachers:

"During the open-ended material creation stage, did you have clear goals or visions for how you would use ChatGPT?" We find that teachers who use ChatGPT more frequently also had clear goals for their class plans. 80% (4 out of 5 teachers) reported productivity increases and had clear class goals. A positive correlation seems to further support the hypothesis that backwards planning, facilitated by AI, leads to better workload and work quality outcomes for those who combine task planning with generative AI task support.

The results show goal clarity strongly correlates with productivity gains: 72% of teachers with reported clear backwards-planning goals achieved measurable improvements, versus 23% of those with ambiguous or unclear objectives. Teachers with clear goals often began their interactions by seeking input into their planning workflows. For instance, a high school math teacher (T18) prompted ChatGPT with: *"Construct a unit plan for teaching the binomial theorem at the Math HL level using the IB syllabus."* This approach enabled him to generate structured plans and iterate on specific components, such as requesting example problems that integrated trigonometric identities. In contrast, teachers without clear goals typically relied on generative AI for task execution, such as requesting pre-made worksheets or quizzes. While this use was perceived as helpful, it did not result in self-reported productivity gains.

These findings further reinforce a critical distinction: generative AI appears to be more effective when used to support backwards planning and iterative workflow design. Teachers who used AI for planning input—not just task output—reported time savings and improved work quality, whereas teachers who relied primarily on AI-generated outputs experienced limited productivity benefits.

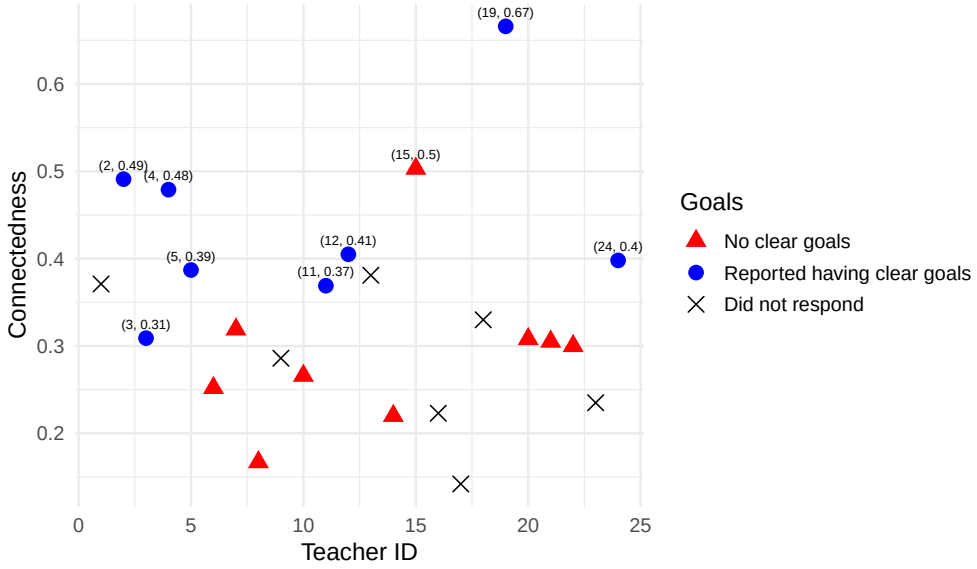
4.2.2. *Connectedness of Prompts*

Another source of evidence that workflow-level use is particularly important for eventual productivity improvement is found by measuring how connected subsequent prompts are within each conversation for each teacher during the observation period. If a teacher had a clear goal in backwards planning, then it follows that the prompts in their conversation with generative AI would exhibit a higher degree of semantic connectedness. Conversely, a lack of clear goals would result in less coherence, with prompts appearing fragmented or unrelated. To operationalize this, we analyze the semantic similarity between prompts using a state-of-the-art embedding approach.

To quantify connectedness, we leverage *SBERT* (Sentence Bidirectional Encoder Representations from Transformers) (Reimers and Gurevych 2019), specifically the pre-trained `all-MiniLM-L6-v2` model from the `SentenceTransformer` library. SBERT generates high-dimensional embeddings for each prompt, capturing their semantic meaning. For each unique teacher ID and conversation ID, we compute pairwise cosine similarity scores between all prompts within the conversation, allowing us to assess the degree of coherence and connectedness in a teacher's line of inquiry. A score of 1 indicates high connectedness and a score of 0 suggests that the prompts are unrelated.

Figure 3 illustrates the connectedness scores for each teacher's series of prompts and whether they reported having clear goals during the study. The analysis highlights that productive use of generative AI often coincides with greater connectedness, reflecting clear instructional goals and iterative refinement of ideas. The results show substantial variation in prompt connectedness among teachers and evidence of how goal clarity influences generative AI use. Teachers who later reported that they had clear and specific goals during the study (that is, Teachers 2, 4, 5, 11, 12, 19, and 24) exhibited high connectedness scores, while those stating the lack of clear goals (except T15) ended up generating prompts with low-level connectedness. Note that Teachers 2, 15, and 19 entered four or fewer prompts, which limited the ability to reliably assess connectedness. The teachers' reported lack of productivity gains suggests that these brief isolated interactions were unlikely to reflect structured or goal-driven planning.

Figure 3 Connectedness scores of prompts within conversations for each teacher and their self-reported goals



Interestingly, among teachers who reported productivity gains–Teachers 3, 4, 8, 12, and 24–there remains variability in connectedness scores. While teachers with high similarity scores demonstrated structured, goal-aligned prompting strategies, those with lower scores relied on volume rather than coherence to extract outputs. This suggests high prompt volume alone might be insufficient to support increased productivity without an overarching planning framework to guide interactions.

4.2.3. *Quality of Prompts: Similarity, Complexity, and Readability*

Finally, we investigate how teachers responded to the outputs generated by ChatGPT and improved their prompts over time within the observation period. To assess this, we proxy prompt quality by

analyzing the consistency of outputs produced by ChatGPT in response to repeated simulations of the same prompts. Consistency provides insight into how well teachers crafted their prompts to obtain reproducible outputs. Importantly, while high consistency (low variance) may reflect clarity and precision, low consistency (high variance) could be more desirable for creative tasks, where diverse outputs are often preferred. This analysis helps explain why certain teachers may have reported varying levels of productivity improvement later on.

To operationalize this, we simulate each prompt sequence used by the teachers and generate eight additional outputs for each prompt, supplementing the outputs obtained during the study. These simulations are conducted using ChatGPT-4 and ChatGPT-4 Turbo–the same models experienced by teachers–to ensure fidelity. To avoid model biases or memory effects, ChatGPT's state is reset after each simulation set. For each set of simulated outputs, we apply SBERT (Reimers and Gurevych 2019) to generate semantic embeddings and calculate pairwise cosine similarity between all outputs for the same prompt. A similarity score close to 1 indicates highly consistent outputs, while a score closer to 0 reflects significant variability and divergence.

Additionally, we compute the variance of output embeddings to quantify variability more explicitly. For each set of output embeddings, the variance is calculated across each dimension. We then average the variance across all dimensions to obtain a single measure of output variability for each prompt. A higher variance score reflects more divergent outputs, while a lower variance score indicates greater consistency.

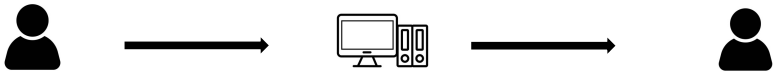
The results show that, on average, each prompt generates outputs with 78.09% similarity, indicating a moderate degree of consistency across responses. However, we observe substantial heterogeneity across prompts, with similarity scores ranging from 30.11% to 94.75%. This variation highlights that certain prompts yield highly consistent outputs—often reflecting precise, specific requests—while others generate diverse responses, suggesting prompt ambiguity or open-endedness.

Interestingly, and consistent with our theory about workflow- versus task-level use, teachers who later reported productivity gains tended to generate prompts that achieved a balance between consistency and variability. For task-focused goals, high consistency ensured reliable outputs that reduced post-processing time. For creative or strategic planning tasks, teachers appeared to value variability, as diverse outputs allowed for a broader exploration of ideas. This finding underscores that prompt quality is not defined solely by precision but by alignment with the teacher's intended goal—whether it be generating predictable results or encouraging creativity.

4.3. Summary of Findings: How Teachers Seek Generative AI's Inputs and Outputs

Figure 4 shows the steps taken by teachers in our study when trying to use generative AI in the two different ways: for at the workflow-level for input versus at the task-level for specific outputs.

Figure 4 The Way Teachers Seek Generative AI Support



Seeking Input (Workflow-Level Use of GenAI)	Through reflection, the teacher identifies a potential that there may be a better way---that they do not know about---to achieve the target learning goals.	Teachers prompt generative AI to iterate on and explore new ideas about how to teach and get students to understand target learning goals.	Teachers either: 1. Feel more confident that what they are doing is an effective way forward. 2. Have ideas for a new workflow plan that they can be more confident about, which they then create.
Seeking Outputs (Task-Level Use of GenAI)	The teacher identifies materials that they need but do not have even a draft of, and so they would like help creating it.	Teachers prompt generative AI to make the materials they need.	Teachers either: 1. Have a draft of the needed materials, which they then modify and improve upon. 2. Have a useable, final version of the materials they need.

Teachers seek input from generative AI when they identify a potential that there is a more effective way to reach the target learning goals. Teacher 1 explained, *"I know a lot of my students really struggle with word problems, and a lot of times like those are hard to create on your own to make sure they align with the actual objective, or you don't use the same example over and over to be like with fractions...it's nice to have other ideas."* Teachers seeking input from generative AI on their plans use it to come up with alternative ideas, not to make an entire plan for their course without incorporating their own expertise and input. As we shared earlier, Teacher 1, for example, came up with the idea to make a word bank for her students. A consequence of generative AI's input is that teachers can feel more confident. Teacher 13 explained, *"I think it saves me some time. But for the most part, I think it saves me a stress instead. It's not like this is like I'm all done. However, once I have a clear outline, then I feel like I'm able to move more efficiently and also feel more confident about it."*

Teachers using generative AI to create outputs are asking generative AI to make materials for them. Yet, in our data, teachers explain that they almost always still do a lot of material creation work themselves modifying or improving upon what generative AI creates. For example, Teacher 6 asked ChatGPT to make a forces-in-motion quiz during our observation period, and she explained, *"I would absolutely need to insert illustrations for each of the questions, since a lot of the students at the age that I have are very visual with regards to their learning."* A high school calculus teacher also mentioned the gap between what he can use and what ChatGPT creates: *"that's the issue with math - I'm going to spend most of my time trying to rewrite this notation...it's a bummer."* The follow-up work when using ChatGPT for content creation is necessary in part because the tool is

not specialized for teachers. New generative AI specifically for educators (i.e., MagicSchool) may be able to get closer to the final product teachers desire. Even with the limitations of ChatGPT, in a few examples in our study, teachers reported getting usable materials immediately, *"I think this is ready to go. This two pages is a perfect assignment for my kids. It's probably what I'm going to give them on tomorrow."* (T17).

4.4. Inhibiting Factors for Using Generative AI

While most of our analysis focuses on generative AI users, and the differences among them, about one third of the teachers in our sample were non-users. Analyzing their observations, interviews, and surveys, we identify three inhibiting factors.

The Features of Generative AI Technology: Generative AI, and ChatGPT in particular, is an evolving technology (García-Peñalvo and Vázquez-Ingelmo 2023). Notably, all six of the improved-productivity users in May 2024 were initially exposed to ChatGPT *before* its upgrade in November 2023. The upgrade added image generation and online searches to ChatGPT Plus, among other things, therefore enhancing generative AI's ability to create outputs compared to before the upgrade. It is possible, therefore, that the teachers exposed before the upgrade discerned an initial potential of generative AI for both providing input and creating outputs, and this initial perception was carried throughout the 2023–2024 school year. Meanwhile, after the upgrade directed teachers' attention primarily to the output creation capabilities. This of course is not possible to test with our data, but it is important to note that no teachers in our study who were exposed to ChatGPT after the upgrade reported productivity gains by the end of the year. Thus, the complex and evolving features of generative AI tools may be inhibiting use by some teachers.

Competing Demands on Time and Energy: Another factor is the extent to which teachers faced competing demands on their time and energy. Integrating generative AI requires adapting routines and the way work gets done (Leonardi 2011). Even if this is productive in the long-run, there is a short-term cost. For some teachers, the cost of adapting their routines and behaviors to integrate generative AI was too high, at least for the current school year. For example, two non-users (Teachers 20 and 22) were both fearful of district layoffs that were announced in January. This was part of the reason Teacher 22 stopped using generative AI: *"I was one of the potential layoffs. And I was recalled. And just all this emotional stuff. Learning new stuff was not on my docket."* Regarding ChatGPT she said, *"I will figure that out one day, you know. Then next month, or in the summer when there's more time."*

Another issue affecting teachers' time and energy was increasing non-sanctioned *student* use of generative AI over the 2023–2024 school year. In June 2024, Teacher 4 (high school English) explained, *"We just were alarmed at the incredible pace with which it seemed to take over content*

within this last school year. There was use of it last year [2022-2023]. But it was a far smaller number of students. And the students who did use it last year seemed to use it on more or less one off occasions in order to basically, you know, shortcut work. Whereas this year there are kids who were basically using it as a substitute for written language. Or their own written language, I guess I should say." More than that, he shared, "*we were just shocked at a how little interest the [school and district] leadership seemed to have and in really having that conversation [about student use].*"

A number of teachers in our sample—Teachers 2, 22, and 24— reported spending *more* time on grading as a result of growing generative AI use in their classes. Overall, teachers reported in our survey an *increase* in stress related to student use of generative AI over the 2023–2024 school year. Teacher 4 was one of the teachers who reported a higher stress level. He was using generative AI, but did not report productivity gains in part because he was alarmed by student use. He concluded with this statement, "*This feels to me bigger than cell phones in schools. People freak out about cell phones in school. There's all kinds of research that shows you know how detrimental they can be. And so there's slow progress, progress to like trying to limit students access to social media in school. [Generative AI] feels much bigger. This feels much potentially more disruptive to like the academic model that we've used, especially."* Overall, non-users and no-change teachers are significantly impacted by other environmental events related and unrelated to generative AI.

Ethical Aversions: Non-users, and some users, expressed varying levels and types of aversion to generative AI. In May 2024, Teacher 2 (a non-user) summarized her stance about teacher use: *"it feels like cheating."* She went on: *"I mean, there's all these lawsuits, right? New York Times and all these other publications that are suing because the [AI companies] are profiting off of uncompensated writing by professional authors."* Another non-user, Teacher 21, similarly emphasized concerns with the way generative AI is designed and trained. In the follow-up interview in May 2024 he explained that *"I think ChatGPT kind of takes people's work... broadly, kind of scraping the internet, to be trained on. And then it seems like large companies like OpenAI, broadly profit off said work."* As a result, *"I think [ChatGPT, and similar genAI] is predicated on a pretty predatory and destructive model just off the base, and that's completely separate from my concerns as a teacher. Conceptually, it's bad."* Teacher 21 questioned what this model means for the future of generative AI. *" ...the current internet model is really based off of advertising, right? But it seems like a lot of generative AI ... you aren't even really going to these websites anymore where the original material is sourced from. So under our current sort of financial model, how are those websites to make money? And if they don't make money, how are they to survive? And if they don't survive, what will ChatGPT feed off of in the future to continue? It seems like a snake that eats itself with very little regulation. And it's already seemingly running low on materials to consume."* Even teachers who were using generative AI, such as Teacher 12, expressed these types of concerns: *"I use it to list sometimes (5*

main reasons for revolutionary war) but I question sources." In sum, ethical aversion also seemed to affect whether and how much teachers were using generative AI.

5. Discussion

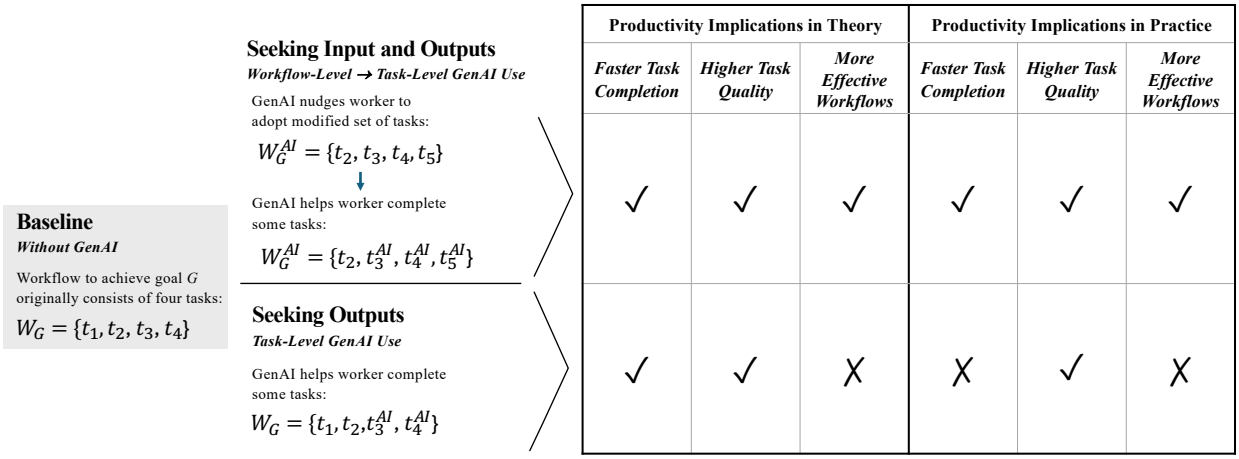
The adoption of new technology is commonly characterized by an initial period of struggle and perhaps even reduced productivity due to the organizational change in workflows and routines required to realize its productivity improvements (McAfee 2002, Brynjolfsson et al. 2019). For generative AI, scholarly attention focused first on demonstrating its long-term potential impact on productivity by testing human performance on specific tasks with versus without generative AI support (e.g., Noy and Zhang 2023, Dell'Acqua et al. 2023, Chen and Chan 2024). Our study helps initiate the next phase of generative AI research that examines the technology's potential impact on human workflows, or the types and sequences of tasks they do as part of their work. Workflow-level research on generative AI is critical for understanding the realized productivity improvement of generative AI, which is likely different than its theoretical potential based on task performance alone.

Workflows vary significantly across different professions. At call centers, workers process calls as they come in and calls tend to involve a standardized set of tasks (Brynjolfsson et al. 2023). In hospitals, doctors and nurses may do fast-paced dynamic work with a high level of discretion, such as when making decisions about tests and interventions in labor and delivery (Freeman et al. 2017). Still others, like the teachers we study, are provided end objectives and plan their work tasks backwards to achieve those objectives in the time allowed (Wiggings and McTighe 2005). Generative AI might be integrated differently in these different workflows, and have different productivity implications. Generative AI tools also might be designed differently for these different workflows.

Figure 5 summarizes our emerging theory of generative AI in backwards planning workflows. The baseline (without generative AI) situation is that a worker has workflow W_G to achieve goal G that involves four tasks: $\{t_1, t_2, t_3, t_4\}$. What we identify is that workers use generative AI in one of two ways. The first way (shown on the top row) is to first seek input on the workflow, and adjust if desired, and then use generative AI to complete specific tasks within the workflow. In the example shown, generative AI nudges the worker to change their workflow to $W_G^{AI} = \{t_2, t_3, t_4, t_5\}$. That is, the worker no longer does task t_1 and instead adds to the end of the flow task t_5. The worker uses generative AI to help complete t_3, t_4, and t_5. In theory, with a sufficiently high-quality generative AI tool, this approach should speed task completion, improve task quality, and improve workflow effectiveness (i.e., it is better for the worker to do W_G^{AI} than W_G). We observe in our data that teachers who use generative AI in this way do feel more productive along these dimensions.

The second way workers might use generative AI (shown on the bottom row) is to use the technology only to help complete some tasks in their original workflow W_G. The theoretical improvement

Figure 5 Generative AI in Backwards Planning Workflows



	Productivity Implications in Theory			Productivity Implications in Practice		
	Faster Task Completion	*Higher Task Quality*	*More Effective Workflows*	*Faster Task Completion*	*Higher Task Quality*	*More Effective Workflows*
Seeking Input and Outputs	✓	✓	✓	✓	✓	✓
Seeking Outputs	✓	✓	✗	✗	✓	✗

in this case comes from faster task completion and higher task quality, but there is no impact on the workflow itself. The worker is doing the same set of tasks they planned at baseline without generative AI. Our data reveals that these theoretical productivity improvements are not necessarily evident; teachers do not report feeling more productive in this use case. We cannot say for sure why this is the case, but our evidence suggests that teachers are unwilling or unable to directly use generative AI outputs in most cases; it takes them as much time to re-prompt, review, and modify the outputs as it would for them to create original materials in the baseline context. Further, when workers use generative AI at the workflow level, they uncover tasks (like t_5) that they might not have been able to do otherwise. Consider, for example, Teacher 9's prompt to generative AI to *"Create a star wars de-codable story that allow first and second grade readers to practice their short o."* Teacher 9 was excited about the possibility of making personalized stories for students based on their interests, which is something she might not have considered doing otherwise.

Another reason why teachers may feel more productive using generative for both inputs and outputs, compared to only for outputs, might be that the generative AI tool we study (ChatGPT) is not customized to their work tasks or style. Custom-designed generative AI tools might be more successful at improving productivity of teachers using it only to create material outputs. At the same time, if desired material outputs are highly variable over time rather than repetitive, then making custom-designed AI for each output might not achieve the productivity gains. In other words, task repetitiveness within workflows might be important for realizing faster task completion times.

5.1. Limitations and Future Work

Our study is only a first step toward understanding how generative AI will and will not improve worker productivity in different workflow conditions. Multiple case study research is intended to

be generative rather than conclusive, and provide more questions than answers (Fisher 2007). Our study follows only a small group of teachers and examines a very early, non-customized generative AI tool. There is much we still do not know. For example, it is hard to disentangle whether the reported productivity improvement from teachers using generative AI at both the workflow and task levels is driven primarily by the use at the workflow level. It could be that teachers who use generative AI are more savvy in some way, and would be more productive even if they only used the technology for creating material outputs. Or, maybe they are more reflective in their workflows generally and thus it is the difference in their mindset about their work, rather than solely their technology adoption choices, that drives the productivity differences we observe.

Our findings motivate at least three new research directions at the intersection of workflow operations and generative AI. The first is to compare workflow- and task-level use of generative AI when workers are using different generative AI tools. For example, in K12 education, there are new tools that primarily offer teachers task-level use (i.e., MagicSchool) and there are tools that enable using AI for both task completion and workflow planning (i.e., YourWay). Such comparisons can illuminate the extent to which productivity improvements are driven by teacher versus technology characteristics. The second is to experiment with training workers (and their managers or organizational leaders) differently. Many workers have already adopted a mindset that generative AI will help them complete things they need to do, faster, thus leaving them disappointed when that does not happen. Our study illuminates a different framing of generative AI, and one that is an extension of Mollick's notion of "co-intelligence" (Mollick 2024): generative AI can help you plan smarter, evaluate decisions, and increase confidence in your workflows. The third new research direction is to conduct studies similar to ours in other professions with different kinds of workflows, perhaps even within the education sector. For example, college counseling involves many repetitive tasks like preparing personal statements and letters of recommendation for dozens and perhaps hundreds of students. In this workflow, the productivity impact of generative AI may look very different than what we observe here.

6. Conclusion

In conclusion, our study leverages an in-depth longitudinal case study of US public school teachers to identify new knowledge about the relationship between workflow operations and generative AI. Teachers must both plan and do tasks for their work. Specifically, they work backwards. They start with learning goals and then plan the set of tasks they (and their students) ought to do to achieve those goals. In this type of workflow, we find two ways teachers are using generative AI: (i) at the task-level to create *outputs* necessary for specific tasks and (ii) at the workflow-level for its *input* on their teaching plans. Teachers using generative AI in *both* ways are the ones that

report productivity gains from using generative AI. From this, we derive an emerging theory of generative AI use within backwards planning workflows that elevates the particular productivity implications for workflow-level use, so workers might "get it right" and not just "get it done." There is still much more to understand about how generative AI will impact work in general, and teacher work in particular. Our study stimulates new directions to explore related to generative AI's capabilities for providing input versus creating outputs, and more generally how generative AI use and productivity potential varies based on workflow structure.

References

Acar, O. A., Tarakci, M., and Van Knippenberg, D. (2019). Creativity and innovation under constraints: A cross-disciplinary integrative review. *Journal of management*, 45(1):96–121.

Altmann, S., Traxler, C., and Weinschenk, P. (2022). Deadlines and memory limitations. *Management Science*, 68(9):6733–6750.

Balakrishnan, M., Ferreira, K., and Tong, J. (2024). Human-algorithm collaboration with private information: Naıve advice weighting behavior and mitigation. *Available at SSRN*.

Bastani, H., Bastani, O., and Sinchaisri, W. P. (2024a). Improving human sequential decision-making with reinforcement learning.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2024b). Generative ai can harm learning. *Available at SSRN 4895486*.

Bhargava, H. K. and Mishra, A. N. (2014). Electronic medical records and physician productivity: Evidence from panel data analysis. *Management Science*, 60(10):2543–2562.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative ai at work. Technical report, National Bureau of Economic Research.

Brynjolfsson, E., Rock, D., and Syverson, C. (2019). Artificial intelligence and the modern productivity paradox. *The economics of artificial intelligence: An agenda*, 23:23–57.

Chen, L., Zaharia, M., and Zou, J. (2023). How is chatgpt's behavior changing over time?

Chen, Z. and Chan, J. (2024). Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 70(12):9101–9117.

Common Core State Standards (2010). *National Governors Association: Washington, DC*.

Dalton, A. N. and Spiller, S. A. (2012). Too much of a good thing: The benefits of implementation intentions depend on the number of goals. *Journal of Consumer Research*, 39(3):600–614.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

Diliberti, M., Schwartz, H. L., Doan, S., Shapiro, A. K., Rainey, L., and Lake, R. J. (2024). *Using Artificial Intelligence Tools in K-12 Classrooms*. RAND.

Ding, W. W., Levin, S. G., Stephan, P. E., and Winkler, A. E. (2010). The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science*, 56(9):1439–1461.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of management review*, 14(4):532–550.

Fisher, M. (2007). Strengthening the empirical base of operations management. *Manufacturing & Service Operations Management*, 9(4):368–382.

Fletcher, R. and Nielsen, R. (2024). What does the public in six countries think of generative ai in news? *Reuters Institute for the Study of Journalism*.

Freeman, M., Savva, N., and Scholtes, S. (2017). Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*, 63(10):3147–3167.

García-Peñalvo, F. and Vázquez-Ingelmo, A. (2023). What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai.

Gates, B. (2024). Unconfuse Me with Bill Gates. https://www.gatesnotes.com/Unconfuse-Me-podcast-with-guest-Sam-Altman.

Ghosh, B., Wilson, H. J., and Castagnino, T. (2023). Genai will change how we design jobs. here's how. Technical report, Harvard Business Review.

Glaser, B. and Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. Routledge: London, UK.

Glazer, J. L. and Peurach, D. J. (2015). Occupational control in education: The logic and leverage of epistemic communities. *Harvard Educational Review*, 85(2):172–202.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., and Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.

Ibanez, M. R., Clark, J. R., Huckman, R. S., and Staats, B. R. (2018). Discretionary task ordering: Queue management in radiological services. *Management Science*, 64(9):4389–4407.

Kagan, E., Leider, S., and Lovejoy, W. S. (2018). Ideation–execution transition in product development: An experimental analysis. *Management Science*, 64(5):2238–2262.

Kc, D. S., Staats, B. R., Kouchaki, M., and Gino, F. (2020). Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science*, 66(10):4397–4416.

Keppler, S. M. (2023). Little's law and educational inequality: A comparative case study of teacher workaround productivity. *Management Science*.

Keppler, S. M., Li, J., and Wu, D. A. (2022). Crowdfunding the front lines: An empirical study of teacher-driven school improvement. *Management Science*, 68(12):8809–8828.

Krishnan, V. and Ulrich, K. T. (2001). Product development decisions: A review of the literature. *Management science*, 47(1):1–21.

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS quarterly*, pages 147–167.

Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410.

McAfee, A. (2002). The impact of enterprise information technology adoption on operational performance: An empirical investigation. *Production and operations management*, 11(1):33–53.

Mollick, E. (2024). *Co-Intelligence*. Random House UK.

Ng, D. T. K., Lee, M., Tan, R. J. Y., Hu, X., Downie, J. S., and Chu, S. K. W. (2023). A review of ai teaching and learning from 2000 to 2020. *Education and Information Technologies*, 28(7):8445–8501.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Picton, I. and Clark, C. (2024). Teachers' use of generative ai to support literacy in 2024. *National Literacy Trust*.

Ramdas, K., Saleh, K., Stern, S., and Liu, H. (2018). Variety and experience: Learning and forgetting in the use of surgical devices. *Management Science*, 64(6):2590–2608.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Silver, D. (2022). A theoretical framework for studying teachers' curriculum supplementation. *Review of Educational Research*, 92(3):455–489.

Snyder, C., Keppler, S., and Leider, S. (2024). Algorithm reliance, fast and slow. *Management Science*, Forthcoming.

Wiese, J., Buehler, R., and Griffin, D. (2016). Backward planning: Effects of planning direction on predictions of task completion time. *Judgment and Decision Making*, 11(2):147–167.

Wiggings, G. P. and McTighe, J. (2005). *Understanding by Design (2nd Edition)*. Association for Supervision and Curriculum Development (ASCD): Alexandria, VA.

Yin, R. K. (2016). *Case Study Research and Applications: Design and Methods*. SAGE Publications: Los Angeles, CA, 6th edition.

Appendix A: T_1 Semi-Structured Protocol

Part 1 (10-15min): Introduction

Tell me about the types of materials you create on your own for your class every week.

- Every day? Every month? At the start of the school year?
- How long does it take you?
- How do you feel about creating this stuff? (probe: Love? Hate? Scale of 1- 10?)
- Do you get materials from anyone else – TeachersPayTeachers? Another teacher?
- Have you ever used ChatGPT/generative AI to help you create any of these materials?
- Is there a school policy about ChatGPT use by teachers?

Part 2 (25 min): Observation

ChatGPT practice (10 min)

(log in to ChatGPT)

Please copy the following prompts (one at a time) into ChatGPT.

1. What is GPT-4?
2. Is 17077 a prime number? Think step by step and then answer.
3. What are today's top news headlines?
4. What notable events happened on February 30, 2020?
5. What notable events happened on February 29, 2020?
6. Explain the economic impacts of the COVID-19 pandemic.
7. Help me write an introductory paragraph for an essay on this topic.
8. Rewrite the paragraph using simpler language.
9. Summarize 'Pride and Prejudice' in one paragraph. [Update: Summarize this text in one paragraph. *(upload PDF - Chapter 43 of 'Pride and Prejudice')*]
10. Please give the same summary as a rhyme.
11. Design a simple workout plan for beginners. [Update: Design a simple workout plan for beginners and present it in table form.]
12. Design a simple workout plan for beginners with limited free time. [Update: Give a diagram of the proper form for one of these exercises.]

Open-Ended Observation (15 min)

Pick one of the things you mentioned earlier (in Part 1) for which you might use ChatGPT to help, and create whatever it is from scratch. Work as if you are trying to create the "finished product" in 15 minutes. You are welcome to use other technology in addition to ChatGPT such as Google Docs, Word, Excel, a web browser, etc. It's okay if you are unable to finish, just work like you'd typically work. Remember, the final product may be included in a publication as an example of how teachers use ChatGPT, so please try your

best. [give 3-minute warning when time is almost up]

Part 3 (10-15 min): Debrief

- Tell me about were you creating.
- Describe what you were thinking about before using ChatGPT.
- Describe what you were thinking while using ChatGPT.
- What was the quality of ChatGPT's output? (probe: Love? Hate? Scale of 1- 10?)
- (If not finished) Describe what else you would do to finish.
- Would you use ChatGPT in practice for something like this? How similar/different is simulation from reality?
- Based on your experience, how useful would ChatGPT be for you in practice?
- Any further reflections / thoughts / questions?

Appendix B: T_2 (January 2024) Survey Questions

1. What is your name? (First and Last)
2. At the time of our interview, how often did you use ChatGPT for your work?
    - Never
    - Occasionally (about once every few months)
    - Sometimes (about once a month)
    - Often (a few times per month)
    - Always (about weekly or more frequently)
3. During our interview, did you have clear goals or visions of what you would use ChatGPT for during the open-ended material creation stage? For example, were you preparing for materials that you'd soon have to make anyways?
    - No - I did not have clear goals of what to make with ChatGPT; just wanted to try
    - I had some ideas but there were no specific or concrete materials I was trying to make
    - Yes - I was trying to make / prepare materials that I could use in my upcoming class / the near future
    - I don't remember.
4. How would you rate the outputs generated by ChatGPT during our interview?
    - Likert scale from 1–5, where 1 indicates "Really bad/not useful" and 5 indicates "Really good/useful."
5. Did that (the quality of the outputs) match your expectation?
    - Likert scale from 1–5, where 1 indicates "Much worse than my expectation" and 5 indicates "Much better than my expectation."
6. Currently, how often do you use ChatGPT for your work?
    - Never
    - Occasionally (about once every few months)
    - Sometimes (about once a month)
    - Often (a few times times per month)
    - Always (about weekly or more frequently)

Display logic: if "Never" is selected for Question 6.

7. Please describe why you do not use ChatGPT for your work.
8. Are there any functions/features you wish it has that would encourage you to use ChatGPT?
9. Please rate how useful you think ChatGPT would be for each of these four common functions people use ChatGPT for:
    (a) to jumpstart your new project/task
    (b) to make or write things for you
    (c) to iterate and work through ideas

(d) to search for and find information

- For each function, a Likert scale from 1–5, where 1 indicates "Not useful" and 5 indicates "Very useful."

10. "Jumpstart for me": Please describe what you think of the use of ChatGPT to jumpstart your new project or task.
11. "Make for me": Please describe what you think of the use of ChatGPT to make or write things for you.
12. "Iterate for me": Please describe what you think of the use of ChatGPT to iterate and work through ideas for your project or task.
13. "Find for me": Please describe what you think of the use of ChatGPT to search for and find information.

Display logic: if "Never" is ***not*** *selected for Question 6.*

7. How often do you use ChatGPT to...
   (a) ....jumpstart your new project/task
   (b) ...make or write things for you
   (c) ...iterate and work through ideas
   (d) ...search for and find information
   - For each function, a Likert scale from 1–5, where 1 indicates "Never" and 5 indicates "Always"
8. If there are functions of ChatGPT that you use but missing here, please state them and describe how/how often you use ChatGPT for those functions.
9. Rank your favorite "functions" of ChatGPT (1 = most favorite/useful and 4 = least favorite/useful)
   (a) ....jumpstart your new project/task
   (b) ...make or write things for you
   (c) ...iterate and work through ideas
   (d) ...search for and find information
10. "Jumpstart for me": Please describe how you may have used or what you think of the use of ChatGPT to jumpstart your new project or task.
11. "Make for me": Please describe how you may have used or what you think of the use of ChatGPT to make or write things for you.
12. "Iterate for me": Please describe how you may have used or what you think of the use of ChatGPT to iterate and/or work through ideas.
13. "Find for me": Please describe how you may have used or what you think of the use of ChatGPT to search for and find information.

Final questions, for all responses.

14. Do you have any closing thoughts on your experience and/or views about ChatGPT since the interview? If so, please describe them here.
15. Please list any other AI tools that you use for your work.
16. *Questions regarding payment details.*

Appendix C: T_3 (May 2024) Additional Survey Questions

1. *(Scaled 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree)* To what extent do you agree with the following statements: Using generative AI has helped me learn how to better do my....
   (a) prepping.
   (b) teaching.
   (c) grading.
   (d) emailing.
2. *(Scaled 1 to 5, where 1 = Strongly Disagree and 5 = Strongly Agree)* To what extent do you agree with the following statements: Generative AI has increased my stress about...
   (a) prepping.
   (b) teaching.
   (c) grading.
   (d) emailing.
3. For each aspect of your job, what is currently true about the effect of generative AI on the number of things/tasks you have to do each week? *(Do more, Do about the same/no change, or Do less)*
   (a) Prepping.
   (b) Teaching.
   (c) Grading.
   (d) Emailing.
4. For each aspect of your job, what is currently true about the effect of generative AI on the total number of hours you spend working each week? *(Do more, Do about the same/no change, or Do less)*
   (a) Prepping.
   (b) Teaching.
   (c) Grading.
   (d) Emailing.
5. For each aspect of student work, what is currently true about the effect of generative AI on the quality of your students' work? *(Higher quality, About the same quality/no change, Lower quality, NA)*
   (a) Writing - formal
   (b) Writing - informal
   (c) Independent learning
   (d) Critical thinking
   (e) Problem solving
   (f) Classroom engagement
   (g) Scientific thinking
   (h) Quantitative/math skills

Appendix D: T_4 (June 2024) Follow-Up Survey Questions

Questions for Improved-productivity Teachers

1. You indicated that you are seeing some productivity boosts from using ChatGPT [verify they agree]. Can you about how you came to be able to use it in that way?
   (a) Did it happen right away?
   (b) Did it change over the course of the year?
   (c) What have you learned?
   (d) What advice would you give to district leaders knowing what you know now? 2-3 things.
2. How are you feeling about the next school year given this AI trend?
3. Anything else you want to share?

Questions for Minimal Users

1. You indicated that you are minimally using ChatGPT [verify they agree]. Can you talk about why you decreased/stopped/are minimally using it?
   (a) Is it that you think the technology is bad, or it's just hard to learn?
   (b) How did things change over the course of the year?
   (c) What have you learned?
   (d) What advice would you give to district leaders knowing what you know now? 2-3 things.
2. How are you feeling about the next school year given this AI trend?
3. Anything else you want to share?

Questions for Non-Users

1. You indicated that you are not using ChatGPT/generative AI [verify they agree]. Can you talk about why?
   (a) If/how has your perspective changed over the course of the year?
   (b) What have you learned?
   (c) What advice would you give to district leaders knowing what you know now? 2-3 things.
2. How are you feeling about the next school year given this AI trend?
3. Anything else you want to share?

Appendix E: Sample State Learning Standards

Table A1 Examples of 5th Grade Common Core State Standards in ELA and Mathematics

Standard Number	Standard
CCSS.ELA-LITERACY.L.5.1.A	Explain the function of conjunctions, prepositions, and interjections in general and their function in particular sentences.
CCSS.ELA-LITERACY.L.5.4.B	Use common, grade-appropriate Greek and Latin affixes and roots as clues to the meaning of a word (e.g., photograph, photosynthesis).
CCSS.ELA-LITERACY.L.5.3.B	Compare and contrast the varieties of English (e.g., dialects, registers) used in stories, dramas, or poems.
CCSS.MATH.CONTENT.5.NBT.A.1	Recognize that in a multi-digit number, a digit in one place represents 10 times as much as it represents in the place to its right and 1/10 of what it represents in the place to its left.
CCSS.MATH.CONTENT.5.MD.A.1	Convert among different-sized standard measurement units within a given measurement system (e.g., convert 5 cm to 0.05 m), and use these conversions in solving multi-step, real world problems.
CCSS.MATH.CONTENT.5.MD.C.5.C	Recognize volume as additive. Find volumes of solid figures composed of two non-overlapping right rectangular prisms by adding the volumes of the non-overlapping parts, applying this technique to solve real world problems.

Appendix F: Clustering and Correlation Analysis of Teachers' ChatGPT Use

Table A2 Teacher Clusters' Descriptions

Cluster	Description	Mean Years of Experience	Median Years of Experience	Teacher IDs
1	The majority of the actions are *Iterate With Me* (at least 42.9%), and the rest of the actions are *Make For Me*. There are no other actions taken by this cluster.	14.25000	16.5	1, 4, 15, 16
2	Their actions were split fairly equally among *Jumpstart For Me, Make For Me,* and *Find For Me*.	15.5	13	8, 13, 17, 21, 22, 23
3	The majority of the actions are *Make For Me* (at least 61.5%)	11.55	11	3, 5, 6, 9, 11, 12, 14, 18, 19, 20, 24
4	The majority of the actions are *Find For Me* (at least 53.3%) and they took comparatively little actions.	17.67	22	2, 7, 10

Table A3 Summary Statistics of Teachers' ChatGPT Use and Correlation with Experience

Variable	Mean	Median	SD	Min	Max	Q1	Q3	Correlation with Experience
Number of conversations	3.24	3	2.036	1	7	1	5	More years of experience, more conversations ($\rho = 0.3589, p = 0.085$)
Number of prompts	8.375	8	4.362	2	17	5	11	More years of experience, fewer total prompts ($\rho = -0.4631, p = 0.0227$)
Average number of prompts per conversation	3.559	2.3	2.835	1	11	1.482	4.438	More years of experience, fewer prompts per conversation on average ($p = 0.003581$)
Median number of prompts per conversation	3.396	2	2.978	1	11	1	3.396	More years of experiences, fewer median number of prompts per conversation ($p = 0.002$)