

More Than What You Asked For: AI Answers Enriched with Task-Adjacent Knowledge Improve Users' Evaluation of What Those Answers Support

HITA KAMBHAMETTU, University of Pennsylvania, USA

PHILIPPE LABAN, Microsoft Research, USA

CHRISTIAN POELITZ, Microsoft Research, UK

ADVAIT SARKAR, Microsoft Research, UK

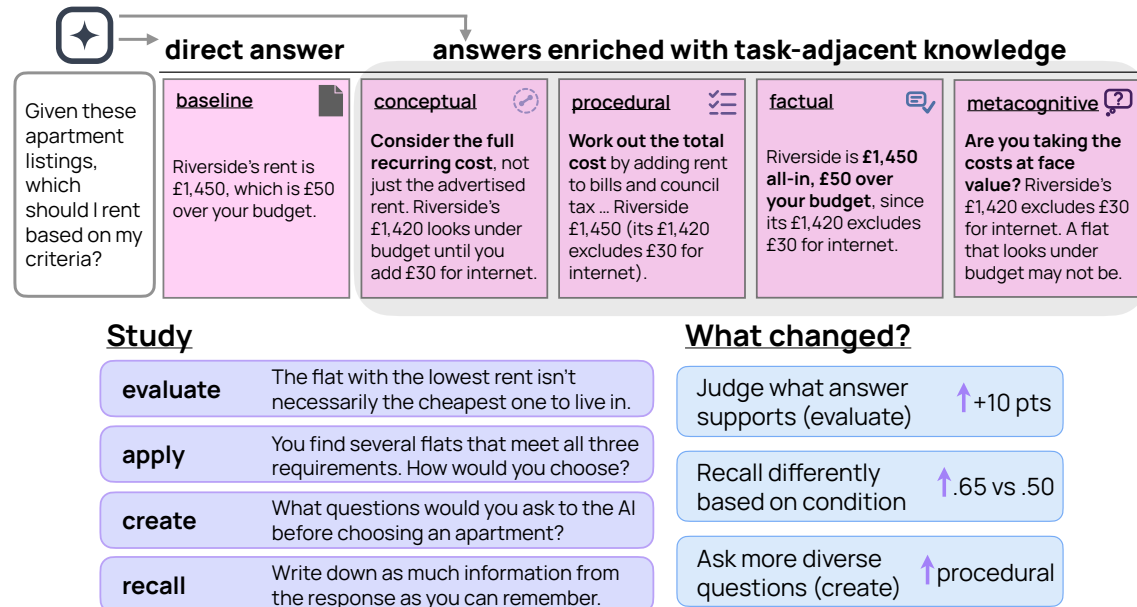


Fig. 1. An AI assistant can often complete a delegated task well enough. However, doing the task oneself might have surfaced more: the details, principles, and caveats picked up incidentally along the way. We add that *task-adjacent knowledge* back into the answer, in four forms from Anderson and Krathwohl's taxonomy [1]: factual, conceptual, procedural, and metacognitive. We ran a between-subjects study ($n = 347$) in which participants read AI answers to four everyday tasks in one of these five conditions. We find that answers enriched with task-adjacent knowledge helped readers judge what an answer did and did not support, what readers recalled carried traces of the form they had read, and procedural answers widened what they asked next.

People increasingly delegate tasks to AI assistants, which generate answers optimized to the query but include little of the breadth of knowledge that doing the task oneself might have surfaced. We ask whether incorporating this task-adjacent knowledge helps readers reason about the answers they receive. Drawing on Anderson and Krathwohl's taxonomy of learning objectives, we enrich answers with four kinds of task-adjacent knowledge: factual, conceptual, procedural, and metacognitive. In a between-subjects experiment ($N = 347$), participants read a baseline or knowledge-enriched AI answer and completed assessments of critical thinking about the answer. Enrichment increased the share of claims participants correctly evaluated as supported or unsupported by 10 percentage points. Participants' recall responses reflected the form of enrichment they read, and procedural knowledge enrichment increased the diversity of their follow-up questions. These findings highlight the potential for AI answers to support readers' critical thinking by including knowledge beyond what they asked for.

CCS Concepts: • Human-centered computing → Interactive systems and tools.

Additional Key Words and Phrases: sensemaking, overreliance, task-adjacent knowledge, AI-assisted task delegation

1 Introduction

Consider someone moving to a new city who has saved a few apartment listings and wants help choosing between them. They need an affordable place with a short commute and enough room to work from home. They ask an AI assistant which apartment they should rent. The assistant checks the options and recommends that they keep looking: none meets all their criteria.

However, if they had compared the listings themselves, they might also have learned something useful for their search. They might have noticed that rent listed as all bills included does not always include internet, which might encourage them to compare total monthly costs rather than rent alone. They might also have reconsidered their commute preferences: would they prefer a longer walk or a shorter bus ride? Even if they had reached the same conclusion as the assistant, they would better understand how to compare apartments and what to look for next.

AI assistants are increasingly capable of completing user-delegated tasks, but often leave out the surrounding knowledge that would help users reason about the broader task. In the example above, the assistant answers the immediate query but offers little guidance for comparing other apartments or deciding what to do next. We refer to such an answer, one that supplies what the query asked for and little else, as a direct answer. Such delegation is a form of cognitive offloading [57], and greater confidence in AI has been linked to less self-reported critical thinking [41]. AI assistants may suggest follow-up questions, but this can also offload the work of deciding what to ask without helping users understand why those questions matter. We call the knowledge these answers leave out *task-adjacent knowledge*: information that completing the task oneself would plausibly have surfaced, which is not required to answer the immediate query but is relevant to judging the answer, using it, and deciding what to ask next.

One might ask why this matters if users can simply delegate subsequent tasks to AI. We see two reasons. First, deciding whether to accept an AI answer or delegate the next task requires some understanding of and reasoning over the answer; a direct AI answer alone offers little basis for that judgment. Independent judgment is also worth protecting in its own right as more tasks become automatable [31]. Second, relying on AI without requisite critical thinking can have negative consequences: explanations can increase confidence without improving judgment [8, 66], AI assistance can leave learners worse off when withdrawn [2], and reliance on AI can narrow the diversity of ideas people collectively produce [18]. Understanding the task helps users ask better questions and check AI answers [72].

Prior work in HCI has explored ways to support reasoning and learning during AI-assisted work. Augmentations such as contextual explanations can help users inspect an answer and assess its basis [38]. Interactions such as cognitive forcing functions, questions, and provocations can encourage users to examine assumptions or consider alternatives [8, 23]. However, users still often use AI assistants without engaging with these additional supports [9]. We investigate whether enriching an AI answer itself with task-adjacent knowledge can recover some of the incidental learning benefits of completing the task oneself. We compare a baseline direct answer with enriched answers that retain its information and add task-adjacent knowledge. We ask:

RQ1 How does knowledge enrichment affect readers’ cognitive processes involved in thinking critically about an AI answer?

RQ2 How does knowledge enrichment affect readers’ subjective experiences of an AI answer?

We ran a between-subjects experiment ($n = 347$) in which participants read assistant-generated answers to *task-situated queries*: questions that arise while someone is working on a task and that a direct answer can satisfy. We

examined what participants remembered from the answers, how they used the information to evaluate claims and respond to new situations, and what they wanted to ask next. We also asked about their subjective experiences with the answers.

Participants were assigned to one of five conditions: a baseline assistant answer or an answer enriched with factual, conceptual, procedural, or metacognitive knowledge, following Anderson and Krathwohl’s revised Bloom’s taxonomy [1]. Figure 1 summarizes the design and what we found. Each of these answers retained the information in the baseline answer while adding task-adjacent knowledge; we refer to them throughout as *knowledge-enriched* answers. This additional knowledge varied in its form, but not in its content.

For RQ1, we measured how readers evaluate what an answer supports, use its information to answer further questions about the task (apply), write follow-up questions to the AI assistant (create), and recall what they read in the answer. All four knowledge-enriched conditions improved evaluation relative to the direct-answer baseline. Procedural enrichment produced a more semantically diverse set of follow-up questions than the baseline. Recall responses differed enough for a classifier to identify, at above-chance accuracy, which enriched condition’s answer participants had read. Enrichment did not significantly improve application quality. These findings show that adding knowledge helped readers judge what answers supported, while its presentation shaped what they asked next and recalled.

For RQ2, we used self-report scales to assess readers’ experiences, including trust, workload, and confidence. Enrichment increased reported workload, while differences in the other measures were not statistically significant.

Our paper makes three contributions:

- We develop a principled approach (Section 3) to enriching AI answers with task-adjacent knowledge, drawing on Anderson and Krathwohl’s taxonomy of learning objectives (Section 2.1).
- We report results from a between-subjects experiment ($n = 347$) examining how knowledge-enriched answers affect what readers remember (recall), how they judge (evaluate) and use (apply) the information, and what they ask next (create). All knowledge-enriched conditions improved evaluation, procedural enrichment led to a wider range of follow-up questions across readers, and participants recalled the same information in different ways depending on the condition they were in (Section 4).
- We outline how AI assistants can include knowledge beyond what directly answers the user’s query to help users reason with the answer more effectively (Section 5). Our findings complement interface augmentations and interaction designs that encourage critical thinking about AI outputs by showing that enriching the answer’s content can also support users’ judgment.

2 Background and Related Work

2.1 Frameworks for Critical Thinking

We study whether enriching an AI answer with task-adjacent knowledge affects readers’ critical thinking about that answer. We hypothesized that different ways of presenting this knowledge would affect different cognitive processes. Prior research supports this: instructional forms such as procedural directions, principles, and examples can support immediate performance, learning, and transfer differently [19]. This suggests that knowledge presentation is a consequential design choice, motivating us to examine how different forms of enrichment affect what readers remember, how they reason with an AI answer, and what they ask next.

Anderson and Krathwohl’s revision of Bloom’s taxonomy offers a framework for operationalizing this design choice. It delineates four knowledge types (factual, conceptual, procedural, and metacognitive) and six cognitive processes:

remembering, understanding, applying, analyzing, evaluating, and creating [1, 39]. The knowledge types provide concrete ways to design answer enrichments: factual knowledge emphasizes specific details; conceptual knowledge explains relationships and principles; procedural knowledge describes methods and how to use them; and metacognitive knowledge concerns strategies for monitoring and guiding one’s own thinking. We use these conditions to present the same task-adjacent content in different forms while preserving the information in the answer.

This taxonomy is one of several frameworks relevant to critical thinking. Other frameworks characterize reasoning skills and dispositions [20, 21], the teaching of reasoning across contexts [26], or standards for assessing reasoning [50]. Information-seeking and sensemaking frameworks explain how people encounter, seek, and organize information to develop their understanding [3, 52, 53]. We want to distinguish what we can vary in an answer from what readers subsequently do with it, and we use Anderson and Krathwohl’s taxonomy because of its organization of knowledge types and cognitive processes.

Searching-as-learning research provides precedents for using the taxonomy to design tasks with specific learning objectives [65] and align assessments with cognitive processes [33, 64]. We adapt this assessment approach to AI answers: participants recall what they read (remember), use the information to make task-related decisions (apply), judge whether the answer supports claims (evaluate), and formulate follow-up questions, our task-specific measure of create. We adopt Anderson and Krathwohl’s taxonomy to guide both the design of our knowledge-enriched answer conditions and the assessment of readers’ responses (Section 3).

2.2 Cognitive Offloading with AI

AI assistants increasingly complete user-delegated tasks by generating answers tailored to the user’s query, reducing the need for users to find and piece together relevant information themselves.

This can offer some benefits. Generative AI can support knowledge work such as drafting, idea development, and stylistic editing [7, 45], with experimental evidence showing faster completion and higher-quality output in professional writing tasks [48]. It can also support literature review activities, such as assessing the relevance of recommended papers and identifying connections between studies [43]. Productivity benefits extend to programming and customer support, including faster completion of programming tasks [51] and more support issues resolved per hour [6].

However, delegating this work can involve cognitive offloading: using actions and external resources to reduce internal cognitive demands [57]. Offloading can benefit learning and memory [60], but delegation also raises concerns about overreliance and fewer opportunities to develop skills through practice. In writing, for example, delegation may reduce opportunities for the deliberate practice involved in developing expertise [36]. In an educational setting, AI assistance without learning safeguards led to poorer subsequent unassisted performance than no AI access [2]. Reliance on external resources can also change what people remember: expecting information to remain available can reduce recall of its content while improving recall of where to find it [58], although this does not establish a general decline in memory ability.

These concerns echo what Potthast et al. call the “dilemma of the direct answer”: the tradeoff between the convenience of receiving an answer and the diligence of thoughtful, deliberative research [54]. We posit that a similar tradeoff arises in AI-assisted task completion. An answer tailored to the user’s request can elide work through which the user might have gained knowledge useful for checking the answer or continuing the task. We ask whether incorporating this task-adjacent knowledge into the answer supports readers’ critical thinking about the information provided. In our study, enrichment improved readers’ judgments of what an answer did and did not support, and the form of enrichment was reflected in what they recalled.

2.3 Learning with AI

Several interventions have been proposed to support learning and critical engagement during AI use.

One approach makes AI output easier to check. Sources, explanations, and alternative answers can help users inspect an answer’s basis and assess its correctness [17, 67]. Their effects, however, are inconsistent [49]. Explanations can increase confidence in an AI system without improving judgment [8, 66]. Their effectiveness appears to depend on the signals provided alongside the answer [38]. A related line of work scaffolds verification itself, helping users determine how much to rely on an output [5, 13].

A second approach introduces interactions that encourage users to engage with AI output before accepting it. Cognitive forcing functions can reduce overreliance in tasks involving explanations [8] and AI-generated plans [23]. Other approaches present AI output as questions [16, 56, 71], or require users to describe the next step before it is shown [34]. Related interventions encourage engagement through attention checks that promote systematic thinking [27], discussion that surfaces conflicting positions [42, 63], or proactive prompts [37, 55].

Metacognitive interventions also support users in directing AI-assisted workflows [62], including articulating goals and intent [15, 61] and decomposing tasks into steps [22, 35, 68–70]. Such interventions can require additional effort: in studies of cognitive forcing functions, interventions that most reduced overreliance were also perceived as more difficult and less usable [8, 12].

Our study complements these approaches by examining the knowledge included in the answer itself. We vary how task-adjacent knowledge is presented, asking whether different knowledge forms affect how readers reason with the answer. This treats answer content as a design choice that can support critical thinking within the conventional question-and-answer paradigm, without additional interface augmentations or interactions.

In summary, finding information oneself involves foraging, sensemaking, and opportunities for incidental learning [3, 52, 53]; delegating that work to AI can change what users learn, remember, and evaluate [2, 57, 58]. Existing interventions support inspection and reflection through output augmentations and additional interactions. We study whether answers can also support these activities by including knowledge that users might otherwise encounter while completing the task themselves. We enrich AI answers to queries arising within everyday tasks with task-adjacent knowledge, using Anderson and Krathwohl’s taxonomy to structure both what we add and what we measure: what readers remember, how they use the answer’s information, and what they ask next [1].

3 Methods

To answer our research questions about how knowledge enrichment affects readers’ critical thinking about an AI answer and their subjective experiences of it, we conducted an online experiment with participants recruited through Prolific. Participants read either a baseline AI answer or a knowledge-enriched answer to a query that was part of an everyday task delegated to an AI assistant.

We obtained ethics approval from our institutional review board. The study began with instructions explaining that they would read four everyday scenarios in which a person delegated a decision or task to an AI assistant. Participants were asked to read each answer carefully and respond to questions based on the answer they had read. Participants were not informed about the closed-book recall assessment so they would not read with the aim of memorizing the answers.

Participants then read either a baseline AI answer or the same answer enriched with factual, conceptual, procedural, or metacognitive knowledge. The enriched answers retained the core information in the baseline answer and included

the same additional knowledge in different forms. In total, we collected responses from 347 participants across five conditions and four everyday tasks. The study was between-subjects; each participant was assigned to a single condition and completed all their tasks in that condition.

To answer RQ1, we used mixed-effects regression models with dependent variables measuring how participants evaluated claims about the answer and applied its information to make new decisions (Table 1): their support ratings, whether their judgments matched the answer key, and whether their responses to application questions were coherent and referenced the answer. The main predictor was answer condition, allowing us to compare the baseline with knowledge-enriched answers collectively and by knowledge type. The models accounted for differences across participants, tasks, and questions. We also compared the semantic diversity of follow-up questions participants wrote and used a classifier to test whether participants’ recall responses predicted which answer condition they had been in. We also qualitatively analyzed written justifications, follow-up questions, and recall responses to highlight differences in how participants responded in different conditions.

To answer RQ2, we used mixed-effects regression models to examine differences in reported trust, workload, confidence, perceived ability to perform the task without AI, and preference for AI assistance. As in RQ1, the main predictor was answer condition, allowing us to compare the baseline with knowledge-enriched answers collectively and by knowledge type. The models accounted for differences across participants and tasks, and across questions for measures collected at the question level.

3.1 Survey Design

3.1.1 Study Tasks Prior research documents AI use for creating and revising content [41] and seeking practical guidance [10], along with interest in everyday planning and organization [29]. Drawing on these uses, we selected tasks according to four criteria: (1) they should be everyday tasks that participants could understand and assess without specialized knowledge; (2) they should reflect real uses of AI documented in prior research (i.e., have external validity); (3) they should involve tool use or analysis of task-specific materials, so that completing the task oneself would involve working through information that could support learning beyond the final output; and (4) they should be suitable for task-situated queries, so that they happen during the course of doing the task rather than as a request to learn about it.

All members of the author team reviewed, discussed, and refined candidate tasks until they agreed that each met these criteria. The study tasks comprised four tasks: choosing an apartment, analyzing marketing information, configuring a new scheduling tool, and adjusting a recipe. The full materials for each task, including the scenario, the delegated request, and the supporting materials, appear in Appendix A.

Each task included a short scenario, a request delegated to the AI assistant, and supporting materials (i.e. a list of apartments the user was choosing from). Participants completed all four tasks, reading answers from the same assigned condition throughout.

Task confidence. Prior work highlights the role of self-confidence and metacognitive monitoring in how people judge and rely on AI outputs [11, 46, 62]. We measured participants’ confidence in their evaluation judgments and application responses. After each evaluation item, participants answered, “How confident are you in your judgment?” After each application response, they answered, “How confident are you in the correctness of your answer?” Both used a six-point scale from “Not at all confident” to “Completely confident.”

Thinking dispositions. We measured two thinking dispositions at the beginning of the study to examine how participants’ general approaches to information related to their responses to AI answers. Need for Cognition captures the

tendency to engage in and enjoy effortful thinking [44]; we assessed it using a six-item scale with five-point responses. Actively Open-minded Thinking captures willingness to consider alternative views and reconsider beliefs in light of evidence [59]; we assessed it using a 13-item scale with six-point responses. Both included reverse-scored items.

Experience with and trust in generative AI. Prior research identifies trust as an important factor in users’ acceptance and use of AI [24]. At the beginning of the study, we collected participants’ general trust in generative AI alongside their self-reported familiarity, frequency of use at work, and tools used. These intake measures captured participants’ prior experiences and attitudes, and post-task trust questions measured their trust in the answers they had read.

Demographics. To characterize our sample, we collected participants’ age range, gender, country of residence, job title, and occupation category. Job titles captured participants’ specific professional roles, while occupation categories provided a broader description of the sample’s work backgrounds.

3.1.2 Study Measures

Answer conditions. We used Anderson and Krathwohl’s taxonomy to guide both answer design and the activities used to assess readers’ responses [1]. Its two dimensions, and how we use each, are summarized in Table 3. We designed five conditions: a baseline answer and four answers enriched with factual, conceptual, procedural, or metacognitive styles of knowledge presentation. The baseline provided a direct recommendation in response to the user’s request.

We developed these knowledge conditions iteratively, against the following criteria. First, every enriched condition had to include the same information, so no condition was advantaged in the study. Second, the kind of information that was included should be such that a user could have come across had they completed the task themselves. Third, each condition had to express that information faithfully in its assigned knowledge form.

We generated candidate answers in several ways, including prompting models to search the web so that information of a given kind surfaced as a user might encounter it, and simulating a user working through the task to see what information they would come across. A simpler approach worked best. Using GPT-5.6-Sol, which performed best in pilots, we first prompted the model to identify four pieces of additional information for each task that were absent from the baseline but relevant to understanding, evaluating, or using the answer. Each addition had to provide information a non-expert would not readily infer from the baseline, be something they could plausibly encounter while completing the task themselves, be expressible in all four knowledge forms, and be unnecessary for answering the request, so that the baseline answer remained a complete response to the query on its own.

We then prompted the model to express these additions in each knowledge form while preserving information in the baseline answer. This design was intended to vary how the added knowledge was presented while keeping the information available to readers consistent across enriched conditions. Comparisons with the baseline examined the effects of adding knowledge, while comparisons among enriched conditions examined the effects of its presentation.

Before study deployment, the author team revised and edited all twenty answers (four tasks across five conditions) until each met the aforementioned criteria. Enriched answers were longer by construction, averaging 272 words compared with 166 for the baseline.

Table 2 illustrates this approach using the apartment task. Each excerpt includes the same information: Riverside’s advertised monthly price excludes a separate internet charge. The conditions present this information in different knowledge forms. Appendix B gives the complete answers.

Table 1. The eleven dependent variables, grouped by the cognitive process or experience each targets. *Type* gives the measurement scale and *Level* the unit at which the measure was collected: once per statement (item), once per task, once per participant, or once per condition.

Target	Dependent variable	Type	Level
Remember	Condition recoverable from recall	balanced acc.	participant
Evaluate	Support rating	6-point	item
	Judgment matches answer key	binary	item
	Confidence in judgment	6-point	item
Apply	Response quality (LLM judge)	binary	item
	Confidence in answer	6-point	item
	Preference for continued assistant use	6-point	item
	Perceived ability without assistant	6-point	item
Create	Semantic diversity of questions	cosine dist.	condition
Experience	Trust (4 items)	5-point	task
	Workload (5 NASA-TLX items)	7-point	task

Table 2. Excerpts from a task in the study where a user is trying to decide which apartment to rent. Note how the same cost information appears in different knowledge forms. Ellipses indicate omitted text.

Knowledge enrichment type	Excerpt
Factual	“Riverside Apartments is £1,450 all-in—£50 over—with its £1,420 covering all bills except £30 for internet, a 15-minute walk, on a 12-month lease.”
Conceptual	“When calculating budget, consider the full recurring cost, not just the advertised rent. Riverside’s £1,420 looks under budget until you add its £30 internet charge—so ‘all bills included’ can quietly leave something out.”
Procedural	“First, work out each flat’s total cost by adding rent to bills and council tax. ... Doing this for all seven gives: ... Riverside £1,450 (its £1,420 does not include £30 for internet) ...”
Metacognitive	“First, are you taking the costs at face value? ‘All bills included’ can still leave something out. Riverside’s £1,420 excludes £30 for internet, making it £1,450. A flat that looks under budget may not be.”

Critical thinking and recall. Figure 2 summarizes the study procedure. Within each task, participants worked through three question sets in a fixed order. They wrote follow-up questions, made assertion judgments, then answered application questions, followed by end-of-task subjective ratings. These corresponded to the create, evaluate, and apply cognitive processes in the revised taxonomy; recall was assessed in a separate closed-book task after all four tasks were completed. The assistant’s answer was available in a collapsible panel throughout the tasks (except during the recall task), so that no participant had an advantage by copying or photographing it in advance. Across the four tasks, each participant wrote four sets of follow-up questions, 16 evaluation judgments, 16 application responses, and four recall responses.

We assessed four of the taxonomy’s six cognitive processes: remembering, applying, evaluating, and creating. These correspond to what readers retain from an answer to a delegated task, how they use it to make decisions it does not directly settle, how they judge what it supports, and what they ask next. We did not assess understanding or analyzing separately because they overlap with aspects of our existing assessments. Readers may demonstrate understanding through their recall and explanations of their judgments. Analyzing, which involves identifying parts of the material and the relationships among them, can contribute to deciding whether an answer supports a claim.

Follow-up questions (create). Participants were asked what additional questions they had for the AI assistant in order to make them more sure of completing the task. They were told to write as many questions as they could, one per line, in a single free-text box. Responses were split on participant-entered line breaks. We measured the semantic diversity of the resulting question pool (i.e. how widely the questions varied in meaning) and qualitatively analyzed what participants questions participants wrote.

Evaluation. Participants judged four statements per task, each prefaced by “Based on the agent’s answer, this statement is:” and rated on a fully labeled six-point scale from *Completely unsupported* to *Completely supported*. Each task included both supported and unsupported claims, according to an answer key prepared in advance. In the apartment task, for example, participants judged whether “the flat with the lowest rent isn’t necessarily the cheapest one to live in” (supported) and whether “the listing’s description of the workspace tells you whether your own desk and chair will fit” (unsupported). Each judgment was followed by a required free-text response to “Briefly, what’s the most important reason for your judgment?” and a six-point confidence rating. We measured whether the direction of each judgment agreed with the answer key and examined how participants used information from the answer in their written reasons.

We developed candidate statements using the baseline answer, then had the author team judge them independently. We discarded statements that authors interpreted differently, unanimously considered trivial, or could not assess from the baseline. Because enriched answers retained the baseline information, the same statements could be assessed in every condition. This allowed us to examine how added knowledge affected readers’ judgments while holding the assessment and the information needed to complete it constant.

Application. Participants answered four free-text questions per task posing decisions the answer did not directly settle. For example, in the task about renting an apartment, participants answered how they would choose among several flats that met all three of their stated requirements. Responses were scored by an LLM judge on whether the responses included a decision, referenced knowledge from the answer, and were internally consistent. We developed these questions by the same procedure used for the evaluation statements, so that each question could also be answered from the baseline answer alone.

Recall. After all four tasks were complete, participants were asked, for each task in turn, to “write down as much information from the agent’s response as you can remember,” one item per line. The original delegated request was shown as a reminder, but the answer was not available. Participants were not told this phase was coming, so they did not read the answers with the intent of memorizing them. We qualitatively analyzed the content of their recall, and trained a classifier on the recall text alone to test whether what participants remembered revealed which knowledge condition they had seen.

Subjective experience. We rated confidence after every evaluation and application response. Each application response was additionally followed by two six-point agreement items, “I would prefer to continue using an agent to complete this task” and “I would be able to complete this task without an AI agent.” At the end of each task, participants completed four answer-specific trust items on a five-point agreement scale (that the answer was reliable, dependable, and trustworthy, and the reverse-scored “I was suspicious of the agent’s answer”), followed by five adapted NASA-TLX items on seven-point scales on mental demand, time pressure, perceived performance, effort, and frustration. These let us examine whether enrichment changed readers’ experience of the answer alongside their performance.

Table 3. The two dimensions of Anderson and Krathwohl’s revised taxonomy [1]. We use the knowledge dimension to structure what our answer conditions add, and the cognitive process dimension to structure what we measure after reading.

Knowledge dimension — what is learned	
Factual	The discrete elements of a domain: terminology, specific details, and verifiable particulars
Conceptual	The relationships among those elements: classifications, principles, generalizations, and models that explain why something holds
Procedural	How to do something: methods, techniques, and the criteria for deciding when to use them
Metacognitive	Knowledge about cognition itself, including awareness of one’s own understanding — what one does and does not know, and what remains to be checked
Cognitive process dimension — what is done with it	
Remember	Retrieving relevant knowledge from memory
Understand	Constructing meaning from information, including interpreting, summarizing, and explaining
Apply	Carrying out or using a procedure in a given situation
Analyze	Breaking material into parts and determining how the parts relate to one another and to an overall structure
Evaluate	Making judgments based on criteria and standards
Create	Putting elements together to form a coherent whole, or reorganizing them into a new structure

3.2 Study Setup and Recruitment

We recruited participants through the Prolific platform, recruiting participants who had completed at least 100 prior tasks with an approval rating of 98% or higher. Because our tasks concern everyday delegation to an AI assistant, we did not screen out participants on the basis of prior GenAI use, but participants’ familiarity with such tools was recorded in the intake questionnaire and is reported in Table 4. Of 502 people who began the study, 347 completed it, and all of these participants met our inclusion criteria of all free-text responses being above 15 characters. Participants spent an average of 67.8 minutes ($SD = 24.5$, median = 64.1) on the study, and were attentive to the task for a median 99% of that time as measured by page visibility telemetry (i.e., the participant did not navigate away from the task page). The 347 participants (194 women, 153 men, 0 non-binary/other) came from different age groups and occupations (see Table 4). Participants were automatically randomized into conditions and the final sample was distributed across the five answer conditions as follows: 73 baseline, 72 factual, 58 conceptual, 71 procedural, and 73 metacognitive. Participants were compensated with GBP £12.50 for completing the study. All participants gave informed consent before beginning the study.

3.3 Analysis Procedure

Each participant completed all four tasks in one answer condition. The order of the four tasks was randomized per participant to avoid order and fatigue effects. Within each task, the three question sets were presented in a fixed order, so that asking follow-up questions would not be influenced by having already seen the evaluation claims. The closed-book recall tasks were after all four tasks were complete. We used quantitative and qualitative analyses to answer our research questions.

3.3.1 Dataset Overview Our 347 participants each completed four tasks, yielding 1,388 responses. These result in 1,388 sets of follow-up questions containing 3,208 individual questions (median 2 per response, maximum 12); 5,552 evaluation judgments, each with a six-point support rating, a required free-text reason, and a confidence rating; 5,552 application responses, each with a free-text answer, a confidence rating, and two reliance ratings; 1,388 end-of-task rating sets covering trust and workload; and 1,388 closed-book recall responses containing 3,588 recalled items.

Table 4. Participant demographics ($N = 347$).

Demographic	Categories	Participants
Gender	Woman	194 (55.9%)
	Man	153 (44.1%)
Age	18–24	38 (11.0%)
	25–34	139 (40.1%)
	35–44	93 (26.8%)
	45–55	46 (13.3%)
	55 or over	31 (8.9%)
Country of residency	United Kingdom	203 (58.5%)
	United States	144 (41.5%)
Occupation (top 5)	Computer and Mathematical	58 (16.7%)
	Office and Administrative Support	31 (8.9%)
	Business and Financial Operations	25 (7.2%)
	Educational Instruction and Library	23 (6.6%)
	Sales and Related	22 (6.3%)
GenAI familiarity	Expert	44 (12.7%)
	Proficient	223 (64.3%)
	Moderate	76 (21.9%)
	Minimal	4 (1.2%)
GenAI use at work	Very frequently	60 (17.3%)
	Often	145 (41.8%)
	Sometimes	80 (23.1%)
	Rarely	26 (7.5%)
	Never	36 (10.4%)
Trust in GenAI tools	A lot	45 (13.0%)
	Quite a bit	129 (37.2%)
	Moderate	141 (40.6%)
	Minimal	26 (7.5%)
	Not at all	6 (1.7%)

3.3.2 Quantitative Analysis We tested two hypotheses: first, that adding knowledge of any kind to an answer affects each dependent variable; second, that the form the added knowledge takes affects them. Each dependent variable was therefore modeled twice. In the first model, `enriched` is a binary predictor contrasting the four knowledge conditions pooled against the baseline. In the second, `condition` is a five-level factor, allowing the individual enriched conditions to be compared with the baseline and with one another. Baseline is the reference level throughout, so a positive coefficient means an enriched answer raised the outcome relative to a plain one.

We fit mixed-effects regression models, logistic for binary outcomes and linear for continuous ones. Because each participant contributed multiple responses to the same items across four tasks, we included random intercepts for participant, task, and item: $DV \sim \text{Enriched} + (1|\text{participant}) + (1|\text{task}) + (1|\text{item})$. Trust and workload were collected once per task rather than once per item, so their models include participant and task intercepts only. We report odds ratios with 95% confidence intervals as our effect size; for the binary outcomes an odds ratio above 1 indicates a higher probability of the outcome under enrichment, and for the rating scales it indicates a shift toward the upper end of the scale.

To account for multiple comparisons, we applied the Benjamini–Hochberg procedure [4] within families of related tests. The enrichment contrasts formed one family, testing whether adding knowledge of any kind affected each dependent variable. Pairwise comparisons among individual conditions formed a second family, with correction across all ten pairs. We also corrected across predictor terms when a model included multiple terms. We report adjusted p -values throughout and use them to determine statistical significance.

To test whether participant recall reflected the answer they had read, we embedded each recall response using frozen BAAI/bge-small-en-v1.5 sentence embeddings and trained class-weighted logistic-regression probes to predict the participant’s condition from the text alone. Probes were evaluated with five-fold stratified cross-validation grouped by participant, so that all four responses from each participant remained within either the training or the test folds. We report balanced accuracy against the chance level implied by the number of classes, with confidence intervals from a participant-level bootstrap and significance from a permutation test that shuffles condition labels across participants.

To measure the diversity of participants’ follow-up questions, we split each response on participant-entered line boundaries, embedded the resulting questions with the same model, and computed mean pairwise cosine distance. Because diversity is a property of a condition’s pool of questions rather than of any individual, and because conditions differed in size and participants in verbosity, we equalized both before comparing: each draw sampled the same number of participants from every condition and, within each task, the same number of questions from every condition, computing diversity per task and averaging across the four tasks. Inference again shuffled condition labels at the participant level.

3.3.3 Qualitative Analysis We open-coded [14] participants’ free-text responses in three places. Because the full corpus was too large to code exhaustively, we coded a sample of each under a fixed budget of roughly 5,000 words, setting each sample size by dividing that budget by the corpus’s mean response length, and drew each sample with a fixed random seed balanced across the twenty condition $\times$ task cells so that no condition or task was over-represented.

From 5,552 evaluation justifications (mean 15.6 words) we analyzed 320, stratifying further by whether the claim was supported and whether the participant’s judgment was key-consistent, four responses per stratum $\times$ condition $\times$ task cell and 314 participants in total; we coded whether a justification drew on information in the answer or rested on what the participant already believed, and, for those that drew on the answer, which kind of knowledge it invoked. From 3,208 follow-up questions to the assistant (mean 12.0 words) we analyzed 420, 21 per cell, coding what participants sought to find out. From 1,388 recall responses (mean 25.9 words) we analyzed 200, 10 per cell, coding the kinds of information each response contained.

The first author coded each sample. After an initial pass, the author team met to discuss the emerging codes and agreed upon a shared set; the first author then completed the coding. We report the codes and their distributions in Section 4, alongside the quantitative results for the same cognitive process, noting how frequently participants raised each theme.

4 Results

We report how knowledge-enriched answers affected each of the four cognitive processes we measured (i.e., Evaluate, Create, Recall, Apply), then how they affected participants’ subjective experience.

4.1 Measured Effects on Evaluate Process

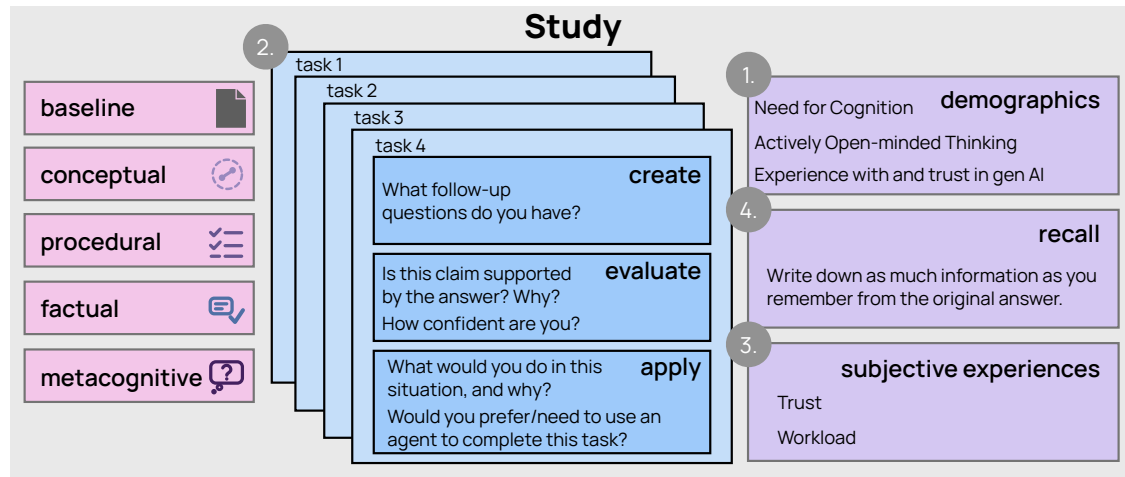


Fig. 2. **Study procedure.** ① Participants first completed a consent form, an intake questionnaire, and instructions, ② then worked through four everyday tasks with the assistant’s answer available (open book), ③ indicated their subjective experiences after each task, and ④ recalled each answer from memory without access to it (closed book). Conditions changed between participants: each participant saw the same one of five answer types — Baseline, Factual, Conceptual, Procedural, or Metacognitive, shown as swatches in that order — on all four tasks. Every task followed the same sequence, with the four question sets were always presented in this fixed order. The recall phase was not announced in advance, so participants could neither read in anticipation of it nor prepare for it.

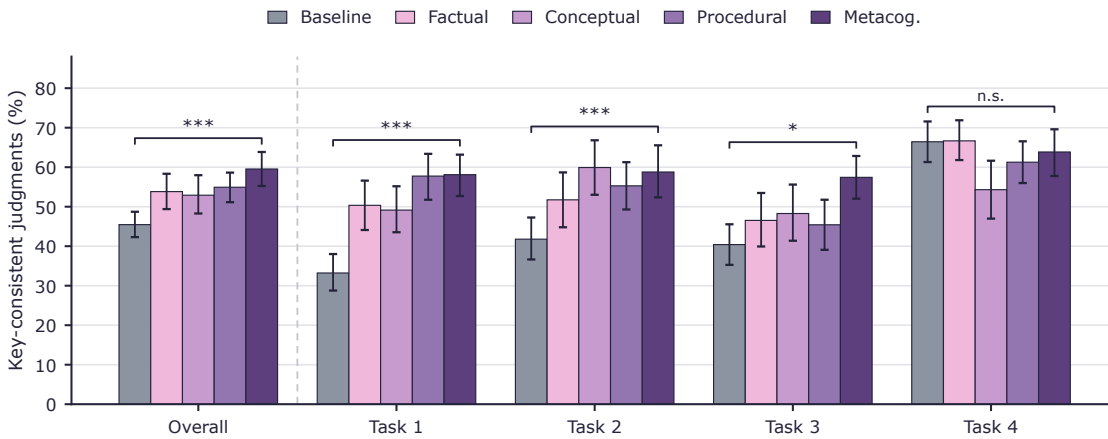


Fig. 3. Correct evaluations by condition, overall and by task. Bars are raw means over 5,568 judgments from 348 participants; whiskers are 95% percentile bootstrap intervals clustered on participants. Brackets compare the baseline with the four knowledge-enriched conditions pooled, marked with BY-adjusted significance (* $q < .05$, ** $q < .01$, *** $q < .001$; n.s. = not significant).

4.1.1 Knowledge enrichment improved key-consistent evaluation Figure 3 summarizes the results on Evaluate tasks, breaking down key-consistent judgments by study condition and task. Participants who read any of the four enriched-condition answers evaluated 55.5% of claims correctly, compared to 45.5% for those who read the baseline answer ($OR = 1.54$, 95% CI [1.31, 1.82], $p < .001$). Every enriched condition outperformed the baseline: metacognitive 59.6%,

procedural 54.9%, factual 53.8%, and conceptual 52.9%. Participants' ratings also moved further in the correct direction along the six-point scale, not merely across the supported/unsupported boundary ($OR = 1.64$, 95% CI [1.42, 1.89], $p < .001$).

The size of this improvement varied across tasks. Enrichment increased accuracy by 20.9 percentage points on task 1 (95% CI [15.3, 26.1]), 14.4 points on task 2 (95% CI [8.0, 21.0]), and 9.1 points on task 3 (95% CI [3.3, 15.2]). It produced no improvement on task 4, where participants reading the baseline answer already performed well. Across the four tasks, baseline accuracy correlated -0.96 with the size of the enrichment effect: enrichment helped where readers of the baseline answer could not separate competing readings of the evidence, and added little where they already could.

The gain came from better discrimination rather than from a greater willingness to reject claims. Sensitivity (d') rose from 0.01 in the baseline to 0.45 across the enriched conditions (difference 0.43, 95% CI [0.25, 0.62], $p < .001$). Baseline participants were therefore performing at chance, distinguishing supported from unsupported claims no better than a coin flip. Enrichment did not change how confident participants were in their evaluations.

When qualitatively analyzing participants' written justifications, two axes of variation emerged. The first is whether a justification referenced information in the answer at all, or whether it was solely based on a participant's prior beliefs. The second, among those that referenced the answer, is which kind of information they referenced.

4.1.2 Correct evaluations referenced the answer more often Justifications divided into those that referenced the answer and those that rested on what the participant already believed. Referencing the answer was more common in the enriched conditions than in the baseline: of the 64 sampled responses per condition, 28 referenced the answer in the baseline, against 33 in the factual, 36 in the conceptual, 36 in the procedural, and 39 in the metacognitive. It was also more common when the evaluation was correct than when it was incorrect, in 113 of 160 key-consistent judgments against 59 of 160 key-inconsistent ones. Whether the claim was itself supported or unsupported bore no relation to it (87 and 85 of 160 respectively).

The contrast is visible in two justifications of the same claim, "With daily limits set, your day could get too full", which the answer supported, in a task where they were setting up a new scheduling tool. Subscripts denote the condition a participant read (i.e., B for baseline, F for factual, etc.).

"The agent says daily limits apply separately to each Event Type, so multiple types could each hit their limit on the same day and make the day too full." — P_{F089} , correct

"The daily limits are there to stop your calendar getting overloaded." — P_{B231} , incorrect

P_{F089} points to a detail in the answer, that limits apply per event type, and reasons from it to the claim. P_{B231} instead states what a daily limit is generally for. Participants who evaluated correctly were more often reasoning over information in the answer than over what they already expected to be true.

4.1.3 Correct evaluations referenced factual and metacognitive information A participant in a condition could provide a justification that referenced any knowledge type (e.g., providing a metacognitive justification). Among justifications that referenced the answer, factual and metacognitive justifications were more common than conceptual or procedural ones. Among correct evaluations, factual details were most often referenced for supported claims and metacognitive details for unsupported ones.

Table 5 shows an example of each type of justification, alongside the claim being evaluated and the condition the participant read. The examples come from a task where participants were setting up a new scheduling tool and one where they were choosing a flat to rent.

Table 5. One correct justification referencing each kind of information, with the condition the participant read and the claim they were evaluating.

Justification Type	Study Condition	Claim	Justification
Factual	Conceptual	“With daily limits set, your day could get too full”	“The agent states that daily limits count per meeting type.” (P _C 013)
Conceptual	Baseline	“The flat with the lowest rent isn’t necessarily the cheapest one to live in”	“Information provided, but would need to make ‘trade off’ or adjust requirements slightly.” (P _B 081)
Procedural	Procedural	“The flat with the lowest rent isn’t necessarily the cheapest one to live in”	“the agent calculates the full monthly cost for every flat before making a recommendation, not just the advertised rent.” (P _P 097)
Metacognitive	Factual	“The listing’s description of the workspace tells you whether your own desk and chair will fit”	“No as no dimensions of a desk and chair were given.” (P _F 126)

This observation might follow from how one could effectively justify a correct and incorrect claim: the former means finding a factual detail in the answer that supports the claim, and the latter means noticing something the answer never claimed, which could look like a judgment about the limits of one’s own information. Participants might have referenced the kind of information needed to effectively justify their assertion judgments rather than the kind their answer was enriched with.

In sum, knowledge enrichment improved participants’ ability to evaluate what an answer supported. Correct evaluations were justified by referencing information in the answer rather than by prior expectation, and the information referenced was factual for supported claims and metacognitive for unsupported ones, unrelated to the participants’ own study condition.

4.2 Measured Effects on Create Process

Next, we turn to the “create” cognitive process, pertaining to the generation of novel ideas and commonly associated with the creative facet of critical thinking.

Semantic diversity of the content generated by participants is often associated with creativity by prior work. Methods typically associate a wider semantic diversity with a wider space of ideas explored. For example, Doshi and Hauser use this approach to show that writers given story ideas by a language model produced individually better but collectively more similar stories [18].

Applying this method, we asked participants to write follow-up questions they would ask about the task, and measured the collective semantic diversity of those questions as the mean pairwise cosine distance between their embeddings (Section 3.3.2).

4.2.1 Procedural answers increased diversity of follow-up questions The results of diversity analysis are summarized in Figure 4. We find that participants in the procedural condition produced more semantically varied questions than baseline participants (.4386 vs. .4212; difference = .0173, $p = .030$), with the difference holding in each of the four tasks. Participants in the factual, conceptual, and metacognitive conditions produced slightly more diverse questions in aggregate, but the difference was not statistically significant.

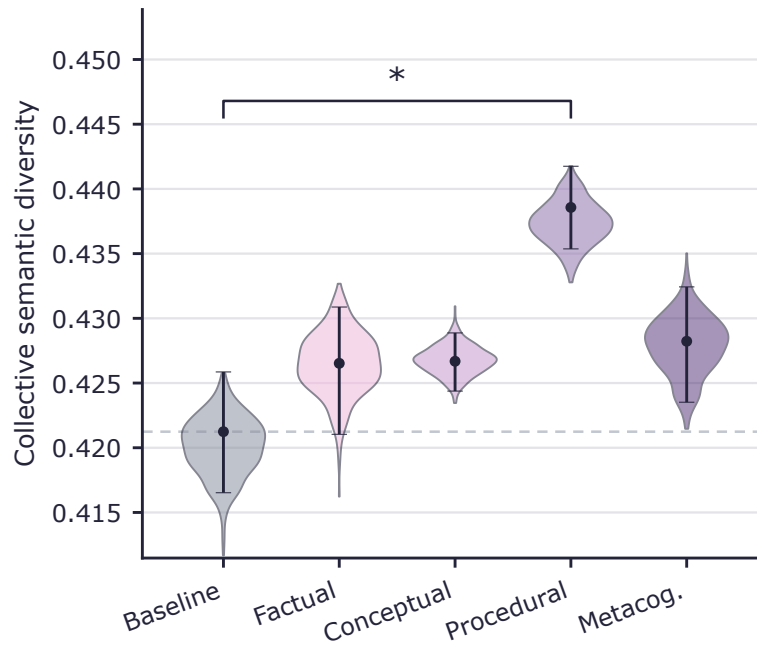


Fig. 4. Collective semantic diversity of participants' follow-up questions, pooled across tasks. Violins are the distribution of rarefaction draws, which equalize participants and questions across conditions within each task; the point and interval give the estimate and its 95% interval, and the dashed line marks the baseline. The bracket marks the confirmatory contrast of procedural against baseline ($\Delta = .017$, permutation $p = .003$, BH-adjusted $q = .030$).

We qualitatively coded a subsample of the participant-written questions to analyze the source of increased diversity.

In our subsample, we find four broad categories of questions, with examples drawn from across the four tasks: choosing a flat to rent, coming up with a marketing campaign for a tutoring center, setting up a new scheduling tool, and scaling a recipe for a larger group: (1) immediate clarifications about the answer ("What type of onions?", P_B095; "Are any bills included in the rent?", P_B066; "what exactly does 'Bills included' cover gas, electric, water, internet, council tax?", P_M149), (2) questions to clarify the answer's accuracy ("does one recipe really serve 4 people?", P_B059; "Is the Feb to April surge a genuine yearly pattern or a one-off spike?", P_F037), (3) questions about the participant's broader task ("What is my current teaching capacity?", P_C208; "What times do I want to offer?", P_P058; "do my working hours change from week to week?", P_B002), and (4) questions about applying the answer downstream ("How would I get more visibility?", P_M028; "how much would transport cost me if i reduced the rent requirement?", P_B050; "is the area safe", P_M224).

Participants in the procedural condition asked fewer immediate clarifications, and instead their questions were in reference to different specific parts of the answer rather than converging on one broad follow-up.

Consider an example from a task where participants were choosing a flat to rent. Asking what the rent covered, a participant in the baseline condition wrote "Are any bills included in the rent?" (P_B066), a participant in the procedural condition named particular bills, "Are utilities (water, gas, electricity) included in the rent?" (P_P146), and a participant in the metacognitive condition asked openly while enumerating what might be hidden, "what exactly does 'Bills included' "

cover gas, electric, water, internet, council tax?” (P_M149). One hypothesis for the observed phenomenon is that answers that leave an ambiguous step lead every reader to ask similar questions, converging on the same gap, whereas an answer that resolves that ambiguity step by step leaves readers free to ask about different things.

4.3 Measured Effects on Recall Process

We examine whether knowledge enrichment in an answer affects what a reader remembers of it.

A binary classifier trained on participants’ recall responses predicting whether the participant was in the baseline or enriched condition reached .650 balanced accuracy against a chance level of .500 (Figure 5), and one predicting which of the four enriched answers a participant had read reached .374 against .250. All ten pairwise comparisons exceeded chance, including the closest pair, factual against metacognitive, at .568 ($p = .002$; all others $p < .001$). This suggests that, because answers differed only in the kind of knowledge they added, the knowledge type left separable traces in what participants recalled. In short, each of the answer conditions with different knowledge forms led to a significant difference in what a reader took away from the answer.

To understand what the differences in recall consisted of, we qualitatively analyzed a sample of 200 recall responses (Section 3.3.3). Participants in every condition recalled many of the same details, but described them differently.

Consider how participants recorded a detail about buffer times in a task where they were setting up a new scheduling tool. Participants in the baseline and factual conditions recorded that buffers were needed: “Set up buffer times.” (R_B043), “Add a buffer around each meeting” (R_B026), “Buffer times” (R_F135). Participants in the other three conditions added further details about the setting: the scope it applies to in the conceptual condition (“Buffer will only work for new schedules preexisting schedules will not have the buffer in them”, R_C071), how it comes to be applied in the procedural condition (“Buffers are not automatic; they are booked with each event”, R_P020), or what it fails to cover in the metacognitive condition (“Buffer times only apply to new bookings”, R_M068).

The same pattern holds for a task where participants were coming up with a marketing campaign based on data from their tutoring center, in which one client had booked 64 lessons. Participants in the factual condition recorded the figure (“Some adults were booking 64 lessons for Business English”, R_F082), while the baseline recorded the recommendation it fed into (“Market to students coming up to exams. Just advertise French courses. Market Locally.”, R_B145). Participants in the other conditions were more likely to record what the figure meant: a general rule in the conceptual condition (“Focus on repeated patterns as they will bring more reliable profits”, R_C090), a sequence of steps in the procedural condition (“Calculate the amount of bookings by each referral type. Identify outliers. Identify the best rate of booking.”, R_P088), or a caution about inference in the metacognitive condition (“A single executive booking 64 lessons is an isolated outlier”, R_M068; “One customer isn’t a pattern.”, R_M137).

In a task where participants were scaling a recipe up to serve a larger group, those in the knowledge-enriched conditions also included more interpretive reasoning in their responses: “Add roughly 20 percent extra food because most guests are big eaters and may want seconds.” (R_F006), “Increased quantities because guests eat a lot” (R_M049), and “the amounts needed to be adjusted not just multiplied” (R_C001). Participants in the baseline typically did not: “ingredients list 1.2kg beef mince 400g vegetarian mince remaining ingredients” (R_B081) and “Ingredients provided with enough to provide 20 servings.” (R_B060).

4.4 Measured Effects on Apply Process

We next report results on the apply cognitive process: whether readers could use an answer’s information to make new decisions.

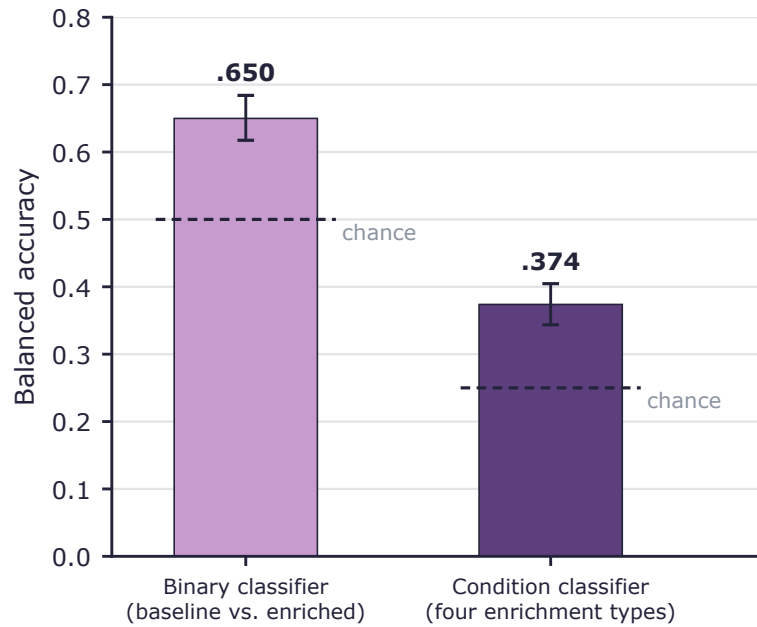


Fig. 5. The condition a participant read is recoverable from their recall text. Balanced accuracy of the two classifiers (binary and all-conditions) against their own respective levels of chance. Whiskers are 95% percentile bootstrap intervals.

Knowledge enrichment did not improve the quality of participants' application responses ($OR = .94$, 95% CI [.73, 1.20]), and no task showed a difference between conditions. Participants' confidence in their own answers was lower under enrichment, and their stated preference to continue using an assistant did not differ by condition, but neither effect was significant.

Perceived ability to complete the task without an AI assistant was also lower under enrichment, though not significantly so, and the pattern was concentrated on the rental task rather than holding generally. Three participant traits pointed in the same direction: (1) self-reported familiarity with generative AI, (2) actively open-minded thinking, and (3) need for cognition, but none of these effects were significant.

4.5 Measured Effects on Trust and Perceived Workload

Enriched answers carried more information than the baseline. We therefore measured whether gains in critical thinking came at a cost in perceived effort, and whether the added knowledge changed how far readers trusted what they were told.

Trust Knowledge enrichment did not change how much participants trusted the answers, which were rated similarly in every condition (baseline 4.12; enriched 4.03 to 4.08).

Prior trust in generative AI predicted post-task trust (+.196 points per standard deviation, 95% CI [.124, .267], $p < .001$); readers arrived with a disposition toward these tools and left with it largely intact, whichever answer

they had read. Higher Actively Open-minded Thinking was associated with lower reported workload ($-.200$, 95% CI $[-.320, -.080]$, $p = .031$).

Workload Participants reported higher workload under enrichment (difference $+.43$, 95% CI $[.15, .71]$, $p = .036$; baseline 3.29 , enriched 3.53 to 3.81). Every one of the five items rose. Three were individually significant: mental demand ($+.46$), frustration ($+.50$), and self-rated performance ($+.53$), where the scale runs from perfect to failure, so a rise means participants felt less successful. Effort ($+.37$) and temporal demand ($+.35$) were not affected.

5 Discussion

5.1 Implications for Designing AI Assistants

Adding task-adjacent knowledge to an answer improved readers' ability to judge what claims the answer did and did not support. Our qualitative analysis of participants' justifications suggests a possible explanation. Correct evaluations drew on information in the answer, with factual information supporting judgments of supported claims and metacognitive information supporting judgments of unsupported claims. Enrichment may make the grounds for these judgments more explicit, helping readers assess claims against the answer rather than rely on prior beliefs.

These findings suggest that task-adjacent knowledge could complement citation-based approaches to supporting judgment. Current AI systems display citations alongside generated answers [25, 47], allowing readers to inspect their sources [38, 67]. However, this requires readers to follow those citations. Citations are sometimes seldom used: one analysis found that users clicked a link within an AI-generated summary in only 1% of visits [9]. Including relevant knowledge directly in the answer may support evaluation even when readers do not consult external sources, although it does not replace independent verification.

Information foraging theory offers one way to interpret this complementarity. It describes how people weigh the cost of pursuing information against the cues that signal its value, or its information scent [52]. An AI-generated answer tailored to a query may appear to complete the task, giving readers little reason to continue searching. Following a citation may therefore feel like additional work after receiving an answer, whereas following a search result is part of obtaining one. From this perspective, citations in AI answers may need stronger cues about why further inspection is worthwhile. Task-adjacent knowledge could reduce the effort of accessing relevant context by placing it in the answer readers are already reviewing. Our findings motivate testing this possibility directly.

That said, this benefit came with higher reported workload. The increase may reflect both the greater length of enriched answers and the effort required to process the added knowledge. Users who delegate a task may primarily want it completed, making that effort an additional demand. Our results identify benefits to weigh against this cost: improved judgment of what an answer supports, and effects on what participants recalled. This tradeoff may be worthwhile when users need to carefully read an answer before making a consequential decision. Future work should examine when that benefit justifies the added effort and how much enrichment is needed to achieve it.

5.2 The Plurality of the AI-Generated Answer

Some effects of enrichment held across knowledge forms (i.e. evaluate, apply), while others were specific to a form (i.e. recall, create). These findings identify both the benefits and limits of knowledge enrichment. Furthermore, they suggest treating enrichment as a set of design choices whose effects should be assessed against the cognitive processes they are intended to support.

More broadly, our study suggests that the AI-generated answer itself can be treated as an interface designed to support critical thinking. This builds on prior work examining how generated text affects its readers. Reports of chatbot interactions have raised concerns that sycophantic agreement may reinforce users’ delusional beliefs [30]. In controlled settings, opinionated writing suggestions shifted participants’ stated views [32], explanations posed as questions rather than statements improved logical discernment [16], and AI-generated story ideas narrowed the collective diversity of writers’ outputs [18]. Our results extend this work by showing that the knowledge included in an answer can improve readers’ judgments of what it supports.

Our findings also suggest a broader target for designing and evaluating AI answers. Benchmarks such as Natural Questions and MMLU assess systems’ answers rather than what readers can do after receiving them [28, 40]. Our results motivate considering this downstream outcome as an additional criterion. It may be particularly relevant to concerns about AI-induced homogeneity, supported by our observation that participants who read procedurally enriched answers produced a collectively more semantically diverse set of follow-up questions than those who read the baseline. Alongside prior evidence that AI assistance can narrow the diversity of creative outputs [18], this suggests that answer design may influence whether users’ subsequent interactions with the AI assistant converge or diverge. While our study does not establish that enrichment reverses AI-induced homogeneity, it identifies a way to broaden what readers ask next. Collecting data on readers’ responses at a larger scale could help determine which additional knowledge, presented in which form, supports different goals and settings.

5.3 Limitations

Our study has several limitations. First, answers enriched with task-adjacent knowledge were longer than the baseline. Participants did not take longer to complete the study overall, but they spent more time with enriched answers open, so we cannot fully separate the effects of added knowledge from those of additional text and reading time. A length-matched design would help distinguish these effects and isolate differences between knowledge forms.

Second, the three question sets appeared in a fixed order: follow-up questions, evaluation, and application. This prevented later assessment items from priming earlier responses, but application was always attempted after participants had already engaged with the answer twice. Counterbalancing would help determine whether our null result for application depends on this ordering.

Third, our quantitative measures are proxies for the constructs they aim to capture. Semantic diversity measures variation in meaning across a condition’s questions, rather than the breadth of any individual’s information needs. The recall classifier shows that responses differed by condition, rather than that any condition produced better recall. Application responses were scored by a language model rather than human raters. Future work could validate these measures against human judgments or complement them with think-aloud protocols that capture reasoning as it unfolds.

Lastly, participants completed predefined tasks and may have felt less invested in outcomes that were not their own. Studies using requests from participants’ own work or observing use in natural settings could examine how readers engage with enriched answers when the outcomes matter to them.

5.4 Future Work

Beyond the directions outlined in this section, we see several opportunities for future work. First, further studies could examine whether the effects of task-adjacent knowledge differ between domain experts and novices. In our study,

enrichment showed no detectable benefit for the task on which participants already performed well in the baseline condition. This motivates examining whether prior knowledge moderates its effects.

Second, future work could identify what knowledge helps readers apply information to new decisions. Enrichment improved evaluation without significantly improving application, suggesting that what helps readers judge an answer may differ from what helps them act on it.

Third, deployment and longitudinal studies could examine these effects over time and in more realistic settings. Varying answer content in deployed AI systems and observing what users subsequently ask and do would allow us to test our hypotheses at scale.

Finally, we studied enrichment within a single conversation turn. Future work could examine how it functions in longer conversations, including whether assistants should vary the amount or form of task-adjacent knowledge as the conversation develops.

6 Conclusion

Delegating a task to AI can remove opportunities to learn from doing the task oneself. We studied whether adding task-adjacent knowledge to an AI answer could recover some of these learning opportunities. In a between-subjects experiment, 347 participants read AI answers to delegated tasks. Factual, conceptual, procedural, and metacognitive enrichment each improved readers' judgments of what an answer supported. Procedural enrichment broadened the collective range of follow-up questions, and recall reflected the knowledge form participants read. Our findings suggest that AI answers can support aspects of critical thinking, and point to the value of designing AI assistants to give answers that are sometimes more than what the user asked for.

7 AI Disclosure

The authors used generative AI tools to improve language and readability, draft or rewrite paragraphs throughout the manuscript. They were also used in the programming of the study UI, and of statistical scripts for data analysis. All outputs were reviewed and verified by the authors.

Acknowledgments

References

- [1] Lorin W. Anderson and David R. Krathwohl (Eds.). 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, NY.
- [2] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning. (2024). <https://doi.org/10.2139/ssrn.4895486>
- [3] Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Review* 13, 5 (1989), 407–424. <https://doi.org/10.1108/eb024320>
- [4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57, 1 (1995), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [5] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–23. <https://doi.org/10.1145/3706598.3714097>
- [6] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at Work. *The Quarterly Journal of Economics* 140, 2 (2025), 889–942. <https://doi.org/10.1093/qje/qjae044>
- [7] Oğuz 'Oz' Buruk. 2023. Academic Writing with GPT-3.5 (ChatGPT): Reflections on Practices, Efficacy and Transparency. In *26th International Academic Mindtrek Conference (Mindtrek '23)*. ACM, 144–153. <https://doi.org/10.1145/3616961.3616992>
- [8] Zana Bucinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21. <https://doi.org/10.1145/3449287>

- [9] Athena Chapekis and Anna Lieb. 2025. Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results. Pew Research Center. <https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/> Published 22 July 2025.
- [10] Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How People Use ChatGPT*. <https://doi.org/10.3386/w34255>
- [11] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32. <https://doi.org/10.1145/3610219>
- [12] Zichen Chen, Yunhao Luo, and Misha Sra. 2025. Engaging with AI: How Interface Design Shapes Human-AI Collaboration in High-Stakes Decision-Making. arXiv:2501.16627 [cs.HC] <https://arxiv.org/abs/2501.16627>
- [13] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–18. <https://doi.org/10.1145/3613904.3641904>
- [14] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology* 13, 1 (1990), 3–21. <https://doi.org/10.1007/bf00988593>
- [15] Adam J Coscia, Shunan Guo, Eunye Koh, and Alex Endert. 2025. OnGoal: Tracking and Visualizing Conversational Goals in Multi-Turn Dialogue with Large Language Models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. ACM, 1–18. <https://doi.org/10.1145/3746059.3747746>
- [16] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–13. <https://doi.org/10.1145/3544548.3580672>
- [17] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30. <https://doi.org/10.1145/3710946>
- [18] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (2024). <https://doi.org/10.1126/sciadv.adn5290>
- [19] Elsa Eiriksdottir and Richard Catrambone. 2011. Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning, and Transfer. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 6 (2011), 749–770. <https://doi.org/10.1177/0018720811419154>
- [20] Robert H. Ennis. 2011. The Nature of Critical Thinking: An Outline of Critical Thinking Dispositions and Abilities. https://education.illinois.edu/docs/default-source/faculty-documents/robert-ennis/thenatureofcriticalthinking_51711_000.pdf Unpublished manuscript, University of Illinois at Urbana-Champaign.
- [21] Peter A. Facione. 1990. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. The Delphi Report. Technical Report. California Academic Press, Millbrae, CA. ERIC Document ED315423.
- [22] K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S Weld, Amy X. Zhang, and Joseph Chee Chang. 2026. Cocoa: Co-Planning and Co-Execution with AI Agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, 1–23. <https://doi.org/10.1145/3772318.3791673>
- [23] Ahana Ghosh, Advait Sarkar, Siân Lindley, and Christian Poelitz. 2026. An Experimental Comparison of Cognitive Forcing Functions for Execution Plans in AI-Assisted Writing: Effects On Trust, Overreliance, and Perceived Critical Thinking. arXiv:2601.18033 [cs.HC] <https://arxiv.org/abs/2601.18033>
- [24] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- [25] Google. 2025. AI Features and Your Website. Google Search Central documentation. <https://developers.google.com/search/docs/appearance/ai-features> Last updated 10 December 2025.
- [26] Diane F. Halpern. 1998. Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist* 53, 4 (1998), 449–455. <https://doi.org/10.1037/0003-066x.53.4.449>
- [27] David J. Hauser and Norbert Schwarz. 2015. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48, 1 (2015), 400–407. <https://doi.org/10.3758/s13428-015-0578-z>
- [28] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*. arXiv:2009.03300 [cs.CY]
- [29] Saffron Huang, Shan Carter, Jake Eaton, Sarah Pollack, et al. 2026. What 81,000 People Want from AI. Anthropic. <https://www.anthropic.com/features/81k-interviews> Published 18 March 2026; qualitative study of 80,508 Claude users across 159 countries.
- [30] Alexandre Hudon and Emmanuel Stip. 2025. Delusional Experiences Emerging From AI Chatbot Interactions or “AI Psychosis”. *JMIR Mental Health* 12 (2025), e85799. <https://doi.org/10.2196/85799>
- [31] International Forum on Digital and Democracy. 2026. Cannes Declaration on the Sovereignty of Mind. <https://sovereigntyofmind.org/> World AI Cannes Festival, Cannes, France, 12 February 2026.
- [32] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–15. <https://doi.org/10.1145/3746059.3747746>

- 3544548.3581196
- [33] Rishita Kalyani and Ujwal Gadiraju. 2019. Understanding user search behavior across varying cognitive levels. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 123–132.
 - [34] Majeed Kazemitabaar, Oliver Huang, Sangho Suh, Austin Z Henley, and Tovi Grossman. 2025. Exploring the Design Space of Cognitive Engagement Techniques with AI-Generated Code for Enhanced Learning. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. ACM, 695–714. <https://doi.org/10.1145/3708359.3712104>
 - [35] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. ACM, 1–19. <https://doi.org/10.1145/3654777.3676345>
 - [36] Ronald T. Kellogg and Bascom A. Raulerson. 2007. Improving the writing skills of college students. *Psychonomic Bulletin & Review* 14, 2 (2007), 237–242. <https://doi.org/10.3758/bf03194058>
 - [37] Minyeong Kim, Jiwook Lee, Youngji Koh, Chanhee Lee, Uichin Lee, and Auk Kim. 2024. Interrupting for Microlearning: Understanding Perceptions and Interruptibility of Proactive Conversational Microlearning Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–21. <https://doi.org/10.1145/3613904.3642778>
 - [38] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–19. <https://doi.org/10.1145/3706598.3714020>
 - [39] David R. Krathwohl. 2002. A Revision of Bloom’s Taxonomy: An Overview. *Theory Into Practice* 41, 4 (2002), 212–218. https://doi.org/10.1207/s15430421tip4104_2
 - [40] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466. https://doi.org/10.1162/tacl_a_00276
 - [41] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–22. <https://doi.org/10.1145/3706598.3713778>
 - [42] Sunok Lee, Dasom Choi, Minha Lee, Jonghak Choi, and Sangsu Lee. 2023. Fostering Youth’s Critical Thinking Competency About AI through Exhibition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–22. <https://doi.org/10.1145/3544548.3581159>
 - [43] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
 - [44] Gabriel Lins de Holanda Coelho, Paul H. P. Hanel, and Lukas J. Wolf. 2020. The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment* 27, 8 (2020), 1870–1885. <https://doi.org/10.1177/1073191118793208>
 - [45] Tiago Lubiana, Rafael Lopes, Pedro Medeiros, Juan Carlo Silva, Andre Nicolau Aquime Goncalves, Vinicius Maracaja-Coutinho, and Helder I. Nakaya. 2023. Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLOS Computational Biology* 19, 8 (2023), e1011319. <https://doi.org/10.1371/journal.pcbi.1011319>
 - [46] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. “Are You Really Sure?” Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–20. <https://doi.org/10.1145/3613904.3642671>
 - [47] Microsoft. 2026. Transparency Note for Microsoft Copilot. Microsoft Support. <https://support.microsoft.com/en-us/microsoft-copilot/transparency-note-for-microsoft-copilot> Last updated 18 August 2026.
 - [48] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. <https://doi.org/10.1126/science.adh2586>
 - [49] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Technical Report MSR-TR-2022-12. Microsoft Research. <https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/>
 - [50] Richard Paul and Linda Elder. 2004. *The Miniature Guide to Critical Thinking: Concepts and Tools*. Foundation for Critical Thinking, Dillon Beach, CA.
 - [51] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirel. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv:2302.06590 [cs.SE] <https://arxiv.org/abs/2302.06590>
 - [52] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675. <https://doi.org/10.1037/0033-295x.106.4.643>
 - [53] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.
 - [54] Martin Potthast, Matthias Hagen, and Benno Stein. 2020. The dilemma of the direct answer. *ACM SIGIR Forum* 54, 1 (2020), 1–12. <https://doi.org/10.1145/3451964.3451978>

- [55] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI 2022)*. ACM, 1–10. <https://doi.org/10.1145/3543829.3543832>
- [56] Liam Richards Maldonado, Azza Abouzied, and Nancy W. Gleason. 2023. ReaderQuizzer: Augmenting Research Papers with Just-In-Time Learning Questions to Facilitate Deeper Understanding. In *Computer Supported Cooperative Work and Social Computing (CSCW '23)*. ACM, 391–394. <https://doi.org/10.1145/3584931.3607494>
- [57] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- [58] Betsy Sparrow, Jenny Liu, and Daniel M. Wegner. 2011. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science* 333, 6043 (2011), 776–778. <https://doi.org/10.1126/science.1207745>
- [59] Keith E. Stanovich and Maggie E. Toplak. 2023. Actively Open-Minded Thinking and Its Measurement. *Journal of Intelligence* 11, 2 (2023), 27. <https://doi.org/10.3390/jintelligence11020027>
- [60] Benjamin C. Storm and Sean M. Stone. 2015. Saving-Enhanced Memory: The Benefits of Saving on the Learning and Remembering of New Information. *Psychological Science* 26, 2 (2015), 182–188. <https://doi.org/10.1177/0956797614559285>
- [61] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–26. <https://doi.org/10.1145/3613904.3642400>
- [62] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–24. <https://doi.org/10.1145/3613904.3642902>
- [63] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–24. <https://doi.org/10.1145/3613904.3642513>
- [64] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2019. Anderson and Krathwohl’s Two-Dimensional Taxonomy Applied to Task Creation and Learning Assessment. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. ACM, 117–124. <https://doi.org/10.1145/3341981.3344226>
- [65] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2020. The Effects of Learning Objectives on Searchers’ Perceptions and Behaviors. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (ICTIR '20)*. ACM, 77–84. <https://doi.org/10.1145/3409256.3409815>
- [66] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- [67] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38. <https://doi.org/10.1145/3579605>
- [68] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22)*. ACM, 1–10. <https://doi.org/10.1145/3491101.3519729>
- [69] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, 1–22. <https://doi.org/10.1145/3491102.3517582>
- [70] Ryan Yen, Jiawen Stefanie Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2024. CoLadder: Manipulating Code Generation via Multi-Level Blocks. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. ACM, 1–20. <https://doi.org/10.1145/3654777.3676357>
- [71] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM, 1–17. <https://doi.org/10.1145/3586183.3606816>
- [72] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–21. <https://doi.org/10.1145/3544548.3581388>

A Study tasks

All four tasks used the same presentation format. Participants first read a short scenario written in the second person, then clicked to send the delegated question to the agent, and then read the agent’s answer. The material that had been attached to the question stayed available throughout in a collapsible panel. The answer itself collapsed once the participant started responding, so that every reopening was a deliberate, logged act.

The wording below is reproduced verbatim from the frozen study configuration. Instructions that were identical across all four tasks are given once, in Appendix A.5, rather than repeated four times.

A.1 Task 1 – Rental choice

Attached material flat_shortlist.csv – a shortlist of seven flats.

Scenario shown to the participant

You’re moving to a new city for a job and need to pick a flat to rent. Three things matter to you – staying within budget once bills are counted, a commute under 45 minutes, and room for a proper desk since you work from home twice a week. You’ve shortlisted several places, and you hand the shortlist to an AI agent to work out what you should do.

Delegated question sent to the agent

Attached are the flats I’ve shortlisted. My budget is £1,400 a month once bills and council tax are in, I need the commute under 45 minutes, and I need room for a proper desk because I work from home 2 days a week. Work out which flat I should rent.

Evaluate items Instructions. For each statement, decide whether the agent’s answer supports it or does not support it – then give the single most important reason and rate your confidence. Throughout, “you” and “my” refer to the person in the task setting – the one deciding which flat to rent.

- (ev1) Riverside is the strongest fallback option in the shortlist.
- (ev2) The listing’s description of the workspace tells you whether your own desk and chair will fit.
- (ev3) The listed commute time tells you whether the flat is within 45 minutes of work.
- (ev4) The flat with the lowest rent isn’t necessarily the cheapest one to live in.

Apply items Instructions. Answer each question below. After each answer, you’ll rate your confidence and choose whether you’d go with your own answer or ask an agent.

- (ap1) You see a new flat with a monthly rent of £1,380, with a 40-minute commute and space for a desk. Would you choose it, and why or why not?
- (ap2) You’re deciding between two otherwise similar flats. One has a 20-minute train commute with a change; the other is a 40-minute walk. Which would you choose and why?
- (ap3) The only in-budget flat with room for a desk is on a 12-month lease, but with a break clause allowing early exit on a month’s notice. Does that clause change whether you’d accept its 50-minute commute? Why?
- (ap4) You find several flats that meet all three requirements. How would you choose amongst them?

Create prompt Heading. What would you check first?

Prompt. Before you decide, what additional information would you seek or verify to make this rental decision more reliable? Write as many questions as you can. (Text-box placeholder: “One question per line.”)

Recall prompt (Part 2, closed book)

Write down as much information from the agent’s response as you can remember. Write one item per line.

The delegated question was redisplayed as a reminder of which task was being recalled. The agent’s answer was not available.

A.2 Task 2 — Tutoring-business targeting

Attached material tutoring_enquiries.csv — a log of 81 enquiries.

Scenario shown to the participant

You run a small business tutoring English and French, both online and in person. Over the past year you've logged every enquiry you've received from interested people. You have a small marketing budget and limited time, and want to start advertising to the group of people most likely to be valuable customers. You ask an AI agent to work out, from that log of enquiries, which group you should market to.

Delegated question sent to the agent

I run a business tutoring English and French both in person and online. Attached is a log of all of the people who have contacted me for my services. I have limited marketing budget and time (£100, 1 month). Work out which group of people I should market my company to.

Evaluate items Instructions. For each statement, decide whether the agent's answer supports it or does not support it — then give the single most important reason and rate your confidence. Throughout, “you” and “my” refer to the person in the task setting — the one deciding who to market the tutoring business to.

- (ev1) The profile of the client who has booked you the most lessons is the group you should market to.
- (ev2) Marketing to local groups through social media is the best way to reach paying students.
- (ev3) Backing GCSE French would guarantee the best return on your £100 marketing budget.
- (ev4) Running your one-month campaign right before GCSE exams would reach parents when demand is highest.

Apply items Instructions. Answer each question below. After each answer, you'll rate your confidence and choose whether you'd go with your own answer or ask an agent.

- (ap1) One new client booked 50 lessons of Italian — the most out of anyone all year. How would you change your marketing strategy, if at all, and why?
- (ap2) You're deciding between partnering with a social media influencer with 100k+ followers and putting the same money into a few school noticeboards seen by far fewer people. Which do you choose, and why?
- (ap3) There are four more senior executives who each booked 50 lessons of Business English. Would this change your marketing campaign? Why/why not?
- (ap4) You run your advertising campaign for GCSE French in the spring, and your bookings rise. Why do you think this is?

Create prompt Heading. What would you check first?

Prompt. Before you spend your marketing budget, what additional information would you seek or verify to make this targeting decision more reliable? Write as many questions as you can. (Text-box placeholder: “One question per line.”)

Recall prompt (Part 2, closed book)

Write down as much information from the agent's response as you can remember. Write one item per line.

The delegated question was redisplayed as a reminder of which task was being recalled. The agent's answer was not available.

A.3 Task 3 — Appointment scheduling

Attached material slotwise_docs.txt — the product documentation.

Scenario shown to the participant

You've signed up for a scheduling tool called Slotwise so clients can book meetings with you without the endless email back-and-forth. You've got its documentation open and want to set it up properly for how you actually work. You hand the docs to an AI agent and ask how to configure it.

Delegated question sent to the agent

I've just signed up for Slotwise so clients can book meetings with me instead of emailing back and forth. Attached are its docs. I take client calls most weekdays — a mix of short intro calls and longer review sessions — and I don't want to be booked back-to-back or at the last minute. Some of my clients are in other countries. Work out how I should set it up.

Evaluate items Instructions. For each statement, decide whether the agent's answer supports it or does not support it — then give the single most important reason and rate your confidence. Throughout, “you” and “my” refer to the person in the task setting — the one setting up their booking tool.

- (ev1) A buffer will keep space around meetings already in your calendar.
- (ev2) With daily limits set, your day could get too full.
- (ev3) Even with reminders switched on, a client could still not receive one.
- (ev4) Handpicking a few times to share would be a reliable way to handle calls that recur every week.

Apply items Instructions. Answer each question below. After each answer, you'll rate your confidence and choose whether you'd go with your own answer or ask an agent.

- (ap1) A few weeks in, you decide to add a new, longer type of session alongside your usual calls. What would you set up or change?
- (ap2) A client in a country five hours ahead of you is about to book a call. Will there be any issues with the time, and if so, how would you handle it?
- (ap3) Sometimes a new booking clashes with a meeting already in your calendar. How would you stop that from happening?
- (ap4) A week after setup, two calls still got booked with no gap between them. Why might that have happened?

Create prompt Heading. What would you check first?

Prompt. Before you decide, what additional information would you seek or verify to make this setup more reliable? Write as many questions as you can. (Text-box placeholder: “One question per line.”)

Recall prompt (Part 2, closed book)

Write down as much information from the agent's response as you can remember. Write one item per line.

The delegated question was redisplayed as a reminder of which task was being recalled. The agent's answer was not available.

A.4 Task 4 — Meal planning

Attached material lasagna_recipe.txt — a recipe written for four people.

Scenario shown to the participant

You’ve offered to host a lasagna dinner for 15 people, including yourself, but every recipe is written for 4 and it’s unclear how much of each ingredient you should actually buy. Two of your guests are vegetarian and most of them usually eat a lot. You hand the recipe to an AI agent and ask for a shopping list of what to actually buy.

Delegated question sent to the agent

I’m hosting a lasagna dinner for 15 people, including myself. Attached is the lasagna recipe I’m going to use. Two of my guests are vegetarian, and most of them usually eat a lot. Work out the shopping list — how much of each ingredient I should actually buy.

Evaluate items Instructions. For each statement, decide whether the agent’s answer supports it or does not support it — then give the single most important reason and rate your confidence. Throughout, “you” and “my” refer to the person in the task setting — the one planning the lasagna dinner.

- (ev1) Once you’ve bought everything, you can make the lasagnas by following the recipe as written.
- (ev2) The beef lasagna should be assembled before the vegetarian one, since it’s the bigger portion.
- (ev3) The shopping list gives you more of some ingredients than you’d actually use in the recipe.
- (ev4) Apart from the mince, both lasagnas share the same ingredients.

Apply items Instructions. Answer each question below. After each answer, you’ll rate your confidence and choose whether you’d go with your own answer or ask an agent.

- (ap1) Three more big eaters confirm the day before, one of them vegetarian. How would you change what you make and buy?
- (ap2) You are hosting another dinner, this time making pizza. There will be 5 vegan guests, 1 guest who does not eat gluten, and 9 guests with no dietary restrictions. How would you go about making the dish?
- (ap3) Your white sauce keeps getting lumpy as it heats. Why might this be happening, and what would you do to fix it?
- (ap4) Some guests eat a lot more than others, and you can’t know who’ll want seconds. How would you decide how much to make for each tray?

Create prompt Heading. What would you check first?

Prompt. Before you shop or begin cooking, what additional information would you seek or verify to make the plan more reliable? Write as many questions as you can. (Text-box placeholder: “One question per line.”)

Recall prompt (Part 2, closed book)

Write down as much information from the agent’s response as you can remember. Write one item per line.

The delegated question was redisplayed as a reminder of which task was being recalled. The agent’s answer was not available.

A.5 Instruments repeated in every task

Per-item scales inside Evaluate Each Evaluate statement was rated on a fully labelled six-point support scale, prefaced by “Based on the agent’s answer, this statement is:” with the points *Completely unsupported*, *Mostly unsupported*, *Somewhat unsupported*, *Somewhat supported*, *Mostly supported*, and *Completely supported*. Each rating was followed by a required free-text item, “Briefly, what’s the most important reason for your judgement?”, and by a six-point confidence rating, “How confident are you in your judgement?” (*Not at all confident* to *Completely confident*).

Per-item scales inside Apply Each Apply question required a free-text answer, followed by “How confident are you in the correctness of your answer?” on a six-point scale, and by two six-point agreement items: “I would prefer to continue using an agent to complete this task” and “I would be able to complete this task without an AI agent”.

End-of-task rating batteries Heading. A few final questions about the agent’s answer *Instructions*. Choose the response that best fits each item. Selecting an answer moves you to the next one.

Trust — 5-point scale (*Strongly disagree* to *Strongly agree*):

- The agent’s answer was reliable.
- I could depend on the agent’s answer for this task.
- The agent’s answer was trustworthy.
- I was suspicious of the agent’s answer. (reverse-scored)

Effort — 7-point scale (*Very low* to *Very high*):

- How mentally demanding was this task?
- How hurried or rushed was the pace of this task?
- How successful were you in doing what you were asked to do? (anchors: *Perfect* to *Failure*)
- How hard did you have to work to accomplish your level of performance?
- How stressed, irritated, or annoyed did you feel?

B Answers shown to participants

Every participant saw the same four tasks and, for each task, the single answer belonging to their assigned condition. The 20 answers (5 conditions $\times$ 4 tasks) reproduced below are the complete stimulus set, verbatim as participants saw them. Text set in bold was rendered in bold in the participant interface; no other formatting was rendered.

All answers were generated with GPT-5.6.

B.1 Task 1 — Rental choice

Table 6. Answers shown to participants for Task 1 (Rental choice), one per condition.

Condition	Answer shown to participants
Baseline	None of the flats on your shortlist meets all three of your requirements — an all-in cost of £1,400 or less, a commute under 45 minutes, and room for a proper desk — so I wouldn't rent any of them just yet. The cheaper, well-connected places (The Old Print Works and Kiln Yard) are studios on 6-month leases with no real room for a desk. The ones with room for a desk are either too far to reach (Willow Street and Maple Court are 50-55 minutes) or are over budget (Riverside, Central Quays, and Beckett House land between about £1,450 and £1,490). I'd keep looking — ideally a furnished one-bedroom or roomy studio with bills included and a commute under 45 minutes. If you had to choose from this list, the closest are Riverside (about £50 over budget but otherwise a match, on a 12-month lease) and Maple Court (within budget with room for a desk but a 50-minute commute, also on a 12-month lease) — and only if you're willing to bend that one requirement.
Factual	None of the seven flats meets all three requirements — £1,400 or less all-in, a commute under 45 minutes, and room for a proper desk — so don't rent any yet. Two are within your budget and commute requirements, but neither has room for a desk. The Old Print Works is £1,120 all-in, a 20-minute metro ride, on a 6-month lease. Kiln Yard is £1,200 all-in, 30 minutes by bus, also on a 6-month lease. Each listing's commute is a single stated time, not a rush-hour or door-to-door figure. Two are within your budget requirement and have room for a desk, but have a longer commute. Willow Street is £1,150 all-in, with a study nook, on a 6-month lease, but it is 55 minutes away by train and a short bus. Maple Court is £1,360 all-in — £1,080 rent plus about £150 for utilities and £130 for council tax — with an alcove that fits a desk, on a 12-month lease, but it is a 50-minute train ride. None of the listings give measurements for the workspace — just written descriptions like “a desk fits.” Three have room for a desk and are within your commute requirement, but go over budget. Riverside Apartments is £1,450 all-in — £50 over — with its £1,420 covering all bills except £30 for internet, a 15-minute walk, on a 12-month lease. Central Quays is £1,460 all-in — £60 over — with £1,300 rent plus about £160 for utilities and council tax, a 10-minute walk, on a 6-month lease. Beckett House is £1,490 all-in — £90 over — with £1,250 rent plus about £110 for utilities and £130 for council tax, 35 minutes by metro and a short walk, on a 6-month lease.
Conceptual	Think of your three requirements as constraints a flat has to satisfy all at once. On that basis, no flat satisfies all three. So, don't rent any yet. When making decisions like these, the useful move is to ask which constraint you would most willingly relax, because each choice opens a different set of options. If you increase the budget, there are three flats with a desk and a short commute: Riverside Apartments (£1,450, a 15-minute walk, 12-month lease, about £50 over), Central Quays (£1,460, a 10-minute walk, 6-month lease, about £60 over), and Beckett House (£1,490, 35 minutes by metro and a short walk, 6-month lease, about £90 over). When calculating budget, consider the full recurring cost, not just the advertised rent. Riverside's £1,420 looks under budget until you add its £30 internet charge — so “all bills included” can quietly leave something out. If you relax the commute, two flats fit your budget and have a desk: Willow Street (£1,150, 55 minutes by train and a short bus, 6-month lease) and Maple Court (£1,360, 50 minutes by train, 12-month lease). If you are ok without a desk, two affordable, well-connected studios open up: The Old Print Works (£1,120, 20-minute metro, 6-month lease) and Kiln Yard (£1,200, 30 minutes by bus, 6-month lease). The key ways to check that the constraints are satisfied are to understand whether these are the right constraints (what if you also need to consider the length of the lease, or the mode of transport you must use to commute?), and whether you have adequate information to confirm they have been satisfied (what if you don't really know whether a place has a proper desk, or just space for a desk you'd have to bring yourself?).

Condition	Answer shown to participants
Procedural	For each flat, pull the three things that map to your requirements: its monthly cost, its commute, and whether it fits a desk. Note the lease length as you go, too — it isn't one of your three requirements, but it sets how long you'd be committed (Riverside and Maple Court are 12 months, the rest 6). First, work out each flat's total cost by adding rent to bills and council tax. Beckett House, for example, is £1,250 in rent plus about £110 for utilities and £130 for council tax, which comes to £1,490 — £90 over your £1,400 limit. Doing this for all seven gives: The Old Print Works £1,120, Kiln Yard £1,200, Willow Street £1,150, Maple Court £1,360, Riverside £1,450 (its £1,420 does not include £30 for internet), Central Quays £1,460, and Beckett House £1,490. Then check each flat against all three requirements. None of the seven passes — so the answer is to keep looking. To see which come close, filter on two requirements at a time. If you filter for places that are within your budget and commute requirements, consider The Old Print Works (20-minute metro) and Kiln Yard (30-minute bus), though both have no desk. If you filter for places that are within your budget and have a desk, consider Willow Street (55 minutes by train and a short bus) and Maple Court (50-minute train), though both are over your commute limit. The door-to-door commute might be longer. If you filter for places that are within your commute limit and have a desk, consider Riverside (15-minute walk), Central Quays (10-minute walk), and Beckett House (35 minutes by metro and a short walk). Also note that you might want to get actual desk and chair measurements rather than trusting the listing.
Metacognitive	If these three requirements are complete and non-negotiable, none of these flats meets them. So, don't rent any yet. It's worth questioning a few assumptions built into both your requirements and the shortlist. First, are you taking the costs at face value? "All bills included" can still leave something out. Riverside's £1,420 excludes £30 for internet, making it £1,450. A flat that looks under budget may not be. Second, are you assuming that commute times are accurate? The modes of commute are different. A metro or train can be delayed, a bus caught in traffic, and a walk slowed by weather. Your real door-to-door time could be longer. Third, you've treated these places as equivalent because they're all "furnished," but that doesn't tell you whether a desk is included or something you'd supply yourself, and no listing gives measurements. You may also be overlooking the leases: Riverside and Maple Court are 12-month commitments, the others 6-month. Finally, which requirement matters most? Because you work from home two days a week, the desk may be your most important requirement. If you're willing to increase your budget, three flats have a desk and short commute: Riverside (£1,450, 15-minute walk, about £50 over budget), Central Quays (£1,460, 10-minute walk, about £60 over), and Beckett House (£1,490, 35 minutes by metro and a short walk, about £90 over). If you're ok with a longer commute, two flats are within budget and have a desk: Willow Street (£1,150, 55 minutes by train and a short bus) and Maple Court (£1,360, 50-minute train). And if you're ok without a desk, studios with no desk space are The Old Print Works (£1,120, 20-minute metro) and Kiln Yard (£1,200, 30 minutes by bus).

B.2 Task 2 — Tutoring-business targeting

Table 7. Answers shown to participants for Task 2 (Tutoring-business targeting), one per condition.

Condition	Answer shown to participants
Baseline	Use your £100 budget and the one month you've allotted to market to parents of Year 10 and Year 11 students. These are students approaching their GCSEs — a set of public academic qualifying exams taken across many subjects at the end of Year 11. Year 11 students will take them within months, and Year 10 students about a year later. According to your logs, these students are roughly 14 to 16 years old and need help with GCSE French, the French exam. In your log, this group has been your most dependable source of repeat work: of the eleven GCSE French enquiries you received, six became paying students, together booking about 110 lessons. No other group both signed up and kept coming back as reliably — younger children, university students, and working professionals mostly stopped early, asked for a lower rate, or booked only a session or two. So make GCSE French your single message, rather than advertising general English and French tuition. Keep the campaign small and local: name the year group, and lead with the exam result parents want. Send anyone who gets in touch to a short introductory call — that's where interest turns into booked lessons. And run it now, in the run-up to the practice and summer exams.
Factual	The clearest pattern in your log is parents of Year 10 and Year 11 students who want help with GCSE French — the national, standardized French exam taken around age 16. Years 10 and 11 are the two years leading up to GCSEs, so these students are roughly 14 to 16. There were eleven such enquiries, all from parents. All arrived between February 8 and April 8, in the run-up to the practice and summer exams. Six of those eleven became paying students. They booked 13, 15, 18, 19, 21, and 24 lessons — roughly 110 in total — at quoted rates of £30 to £38 an hour. All six came through online advertisements, school noticeboards, and parent referrals. The two who found you via social media didn't pay. The other three stalled over price or went quiet. The log doesn't show your actual revenue, your costs, or what advertising would cost. One senior executive booked about 64 lessons of Business English. There were two A-level French enquiries (A-level is the level taken around age 18, after GCSEs), but both became paying students, booking about 21 and 26 lessons.
Conceptual	When making data-driven decisions, a key principle is to follow a repeatable pattern, rather than being enticed by the extremes, like the most number of lessons booked. One senior executive in your log booked about 64 lessons of Business English. This is notable, but one person might be an outlier rather than a pattern. The more repeatable pattern is parents of Year 10 and Year 11 students wanting help with GCSE French. GCSEs are the public exams students sit across many subjects at the end of Year 11, around age 16. GCSE French is the French one. Years 10 and 11 are the two years leading up to them, so these students are roughly 14 to 16. This group made eleven enquiries, all from parents. Six of them became paying students, booking roughly 110 lessons between them. They booked about 13 to 24 lessons each at £30 to £38 an hour. Two enquiries for A-level French (A-level is the level taken around age 18, after GCSEs) became paying students. This looks great, but again is not a repeatable pattern. A useful way to read data like this is to dig into its context until you have a full enough picture to make a recommendation. Here, that means three things: how people find you, when they book, and what the log leaves out. How people find you matters: a channel is only worth the paying customers it brings. Your six paying GCSE French families came through online advertisements, school noticeboards, and parent referrals; the two who made enquiries through social media never paid. When they book matters too, because exam demand is seasonal. All eleven enquiries arrived between February 8 and April 8, before the practice and summer exams. The log also leaves out your revenue, your costs, and what advertising would cost, so you can't make assumptions about profit.

Condition	Answer shown to participants
Procedural	Go through the enquiries and sort them into groups by what services they wanted. You will see the following requested services: GCSE French (the national, standardized French exam taken around age 16), A-level French (the level taken around age 18, after GCSEs), primary-age French, university help, business or professional English, and so on. For each group, count three things: how many enquiries came in, how many turned into paying students, and how many lessons those students actually booked. One group stands out: parents of Year 10 and Year 11 students, students who are roughly ages 14 to 16, wanting help with the GCSE French exam. There were eleven such enquiries. Six became paying students, booking about 13 to 24 lessons each — roughly 110 in total — at £30 to £38 an hour. While one Business English executive booked about 64 lessons, that looks like an outlier. Both enquiries for A-level French — the level taken around age 18, after GCSEs — became paying students, but there were only two. Look at where the paying GCSE French families actually came from, which are online advertisements, school noticeboards, and parent referrals. The two who came through a local social media group did not pay. Also note when they came in — all eleven between February 8 and April 8 — and time your month to the run-up to the practice and summer exams. Aim your campaign at parents of students taking GCSE French: name the year group, lead with the exam result, and send responses to a short introductory call. As you run it, record each lead's source, your actual spend, the lessons booked, and the revenue — your current log doesn't capture this — so next time you decide on real profit, not lesson counts.
Metacognitive	Consider how much your log can actually tell you, versus what you'd be guessing. The one thing it shows well is which past enquiries turned into steady work. One group stands out: parents of Year 10 and Year 11 students wanting help with GCSE French. GCSEs are the public exams taken across many subjects at the end of Year 11. GCSE French is the French one. Years 10 and 11 are the two years leading up to them. These students are roughly 14 to 16. Of the eleven such enquiries, six became paying students. They booked 110 lessons between them at £30 to £38 an hour. Be more careful interpreting the case of the one executive who booked 64 lessons of Business English. One customer isn't a pattern, so your confidence should rest on the group that consistently booked lessons. The same caution applies to the two A-level French enquiries (A-level is the level taken around age 18, after GCSEs). Though both ended up paying for lessons, two is too few to consider as evidence to support the marketing campaign. Watch what you're assuming beyond the data, too. The log tells you that your paying customers usually came across your services through online advertisements, school noticeboards, and referrals. The two who came from social media did not end up paying for lessons. Your logs do not tell you which channel is cheapest to advertise on, because it records no ad spend. It also does not show revenue or your costs, so when it is noted that "lots of lessons" have been booked, this indicates demand, not proof of profit. And notice that all of your enquiries arrived between February 8 and April 8, in the run-up to the practice and summer exams. This might mean that a marketing campaign run in different months might have different results than this log suggests. So treat GCSE French parents as your best-supported option from your logs, rather than a definitive answer. Aim your marketing campaign at the parents, name the year group, and lead with the desired exam results your services may offer.

B.3 Task 3 — Appointment scheduling

Table 8. Answers shown to participants for Task 3 (Appointment scheduling), one per condition.

Condition	Answer shown to participants
Baseline	Use Slotwise’s Availability Rules — a recurring pattern of weekday hours you set once, which it then keeps offering automatically. Choose the hours you actually want calls in, for example Monday to Friday, 10:00 to 12:00 and 2:00 to 4:00, and connect the calendars you use (including your personal one) so it won’t offer times you’re already busy. Create an Event Type — a template for one kind of meeting — for each call you run, such as a 15-minute intro call and a 60-minute review. Give each a 15-minute buffer so calls aren’t booked back-to-back, a minimum notice of 24 hours, and a sensible daily limit. Finally, turn on time-zone auto-detect so overseas clients see their own local time, and switch on email reminders.
Factual	For regular weekly calls, use Slotwise’s Availability Rules. Availability Rules are a recurring pattern of open hours you set once that Slotwise keeps automatically. Conversely, Quick Slots are hand-picked one-off times that don’t refresh: once they’re booked or past, the link goes empty, so they suit single meetings only. Slotwise blocks out your busy times only on calendars you connect to it. An Event Type is a template for one kind of meeting, with its own length, booking link, and rules. You’d make one for each kind of call you plan to take, such as a 15-minute intro call and a 60-minute review. A buffer pads a meeting before and after so bookings can’t land back-to-back, and it applies only to bookings made after you turn it on. Minimum notice sets the lead time a booking needs. A daily limit caps how many of that meeting type can be booked in a day, and it’s counted for each Event Type separately, so two types can each fill to their own limit on the same day. Time-zone auto-detect, which shows overseas clients your times in their own local time, is off by default. Email reminders to both people are off by default too, and are set for each Event Type. Reschedule and cancel links are added to every booking automatically.
Conceptual	The main idea is to set it up in such a way that bookings can be made regularly, instead of arranging each one individually. That’s what Availability Rules are — a recurring pattern of open hours you set once and Slotwise keeps offering — and they fit weekly client calls, whereas Quick Slots (one-off times you pick) only make sense for a single meeting. For this to work reliably, it also needs one full picture of when you’re free, so connect every calendar you use, including your personal one; any you leave out, it can’t see — so it treats those hours as free and can book a client on top of a commitment already sitting there. When setting up these kinds of services, it’s important to remember to check what certain settings actually do. Because different kinds of call need different rules, give each one its own Event Type. Add a buffer so calls can’t be booked back-to-back. Buffers are only applied to bookings made from then on, not meetings already in your calendar. Add a daily limit so a single day can’t fill up completely. That limit counts per meeting type. A helpful note to remember when setting up recurring meetings is to check which features are enabled by default. Time-zone auto-detect, which shows overseas clients your times in their own local time, and email reminders both start off, so you turn them on yourself. Reschedule and cancel links, by contrast, come with every booking.

Condition	Answer shown to participants
Procedural	Open Slotwise and go to the Availability section. Create a new Availability Rule and set your recurring weekly hours. Select Monday to Friday and enter 10:00 to 12:00 and 2:00 to 4:00, then save. Leave the Quick Slots option alone; that's for sharing a few one-off times, not a regular schedule. Next, open the Calendars settings and click Connect for each calendar you use — your work one and your personal one. Make sure each shows as connected, because Slotwise can only block out busy times on calendars that are linked. Now go to Event Types and click New Event Type. Make one for a 15-minute intro call — give it a name and length and save — then click New Event Type again and make a 60-minute review the same way. Open the intro call's settings. Switch the Buffer toggle on and set it to 15 minutes before and after. Note that this only affects new bookings. Set Minimum notice to 24 hours. Set the Daily limit to how many of that call you'll take in a day. Daily limit is per Event Type, so if you want three calls total in a day, split that across the two types rather than putting three on each. Then open your account Settings and turn the Time-zone auto-detect toggle on, so overseas clients see your times in their own local time — it stays off until you do. Back in each Event Type, turn the Email reminders toggle on; it's off by default. Finally, copy your booking link and send yourself a test booking. Check the reminder email arrives and that the time shows correctly in another country's zone. Reschedule and cancel links are already on every booking — just set how much notice you want before someone can use them.
Metacognitive	Your choice of setup assumes that your calls are regular. If they are, Availability Rules, a recurring number of meeting hours you set once, are the right fit. Conversely, Quick Slots would let you set up one-off meetings that only happen once. You might expect the tool to know your whole schedule, but it only sees the calendars you connect to it. Connect every calendar you use, including your personal one, otherwise it may offer a time you're already busy. Some settings might need more manual configuration than you might think. A buffer changes only the bookings you make after you switch it on, not the meetings already in your calendar. And a daily limit applies to each kind of call you set up on its own, not to your day as a whole. So a limit of "three a day" across two call types isn't a cap of three; it's up to six. And you might expect some features to be enabled by default. However, time-zone auto-detect and email reminders both start switched off, so the thing you're relying on for overseas clients or no-shows may not actually be running. The best way to be sure is to test it: send yourself a booking and check that the reminder arrives and the time shows correctly for someone abroad.

B.4 Task 4 — Meal planning

Table 9. Answers shown to participants for Task 4 (Meal planning), one per condition.

Condition	Answer shown to participants
Baseline	For 15 people with big appetites, make one large beef lasagna and one smaller vegetarian one — about 20 servings in all, so there's plenty, with a whole small lasagna for the two vegetarians. Buy: - 1.6 kg beef mince - 400 g vegetarian mince - 5 medium onions - 2 bulbs garlic - 5 × 400 g tins chopped tomatoes - 1 × 200 g tube tomato purée - 1 small jar dried oregano - 1 bottle olive oil - 250 g butter - 500 g plain flour - 3 litres whole milk - 1 small jar ground nutmeg, optional - 1 kg dried lasagna sheets - 500 g mature cheddar - 150 g vegetarian Italian-style hard cheese - Salt and black pepper Use about four-fifths of these ingredients except the mince for the beef lasagna. Use the rest of the ingredients for the vegetarian one.
Factual	This shopping list makes five batches of the recipe: four beef and one vegetarian. That's about 20 servings for your 15 guests. Buy: - 1.6 kg beef mince - 400 g vegetarian mince - 5 medium onions - 2 bulbs garlic - 5 × 400 g tins chopped tomatoes - 1 × 200 g tube tomato purée - 1 small jar dried oregano - 1 bottle olive oil - 250 g butter - 500 g plain flour - 3 litres whole milk - 1 small jar ground nutmeg, optional - 1 kg dried lasagna sheets - 500 g mature cheddar - 150 g vegetarian Italian-style hard cheese - Salt and black pepper The five batches use about 200 g butter, 200 g flour, 2.5 litres milk, 400 g cheddar and 100 g hard cheese, so several of the packs above will have some left over. That 2.5 litres of white sauce is a large volume, which thickens slowly and lumps easily. The recipe's 35–40 minute bake is for a small serves-4 dish; a large, deep tray takes longer, and its top browns before the middle is hot. The beef tray takes about four-fifths of the ingredients apart from the mince, and the vegetarian tray the remaining fifth. Traditional Parmesan is set with calf rennet, an animal-derived ingredient, so it isn't vegetarian. And utensils that have touched beef can carry traces of meat.

Condition	Answer shown to participants
Conceptual	A few ideas will help you scale a recipe for a larger group. First, because you are cooking for people with different dietary restrictions, consider splitting the ingredients and items used to cook. Use about four-fifths of the ingredients, except the mince, for the beef lasagna. Use one-fifth for the vegetarian one. Also, make sure that the ingredients that you purchase for the vegetarian lasagna are indeed vegetarian, particularly the cheese, and that you are using separate pans to reduce cross-contamination. The list below estimates the package sizes available, but it might be more than you'll actually use. Buy: - 1.6 kg beef mince - 400 g vegetarian mince - 5 medium onions - 2 bulbs garlic - 5 × 400 g tins chopped tomatoes - 1 × 200 g tube tomato purée - 1 small jar dried oregano - 1 bottle olive oil - 250 g butter - 500 g plain flour - 3 litres whole milk - 1 small jar ground nutmeg, optional - 1 kg dried lasagna sheets - 500 g mature cheddar - 150 g vegetarian Italian-style hard cheese - Salt and black pepper Cooking for a large group from a recipe meant for far fewer means adjusting it, not just multiplying the ingredients. Watch the cook times especially. The white sauce may thicken slowly and lump easily, because there's so much more to heat evenly. And a large, deep tray holds more heat than the small dish the 35–40 minute bake assumes, so the middle takes longer to warm through while the top is already browning.
Procedural	The recipe serves 4 and you've got 15 hungry guests, so make five batches — about 20 servings. Make four batches into a large beef lasagna and one into a smaller vegetarian one, so the two vegetarians have their own. Buy: - 1.6 kg beef mince - 400 g vegetarian mince - 5 medium onions - 2 bulbs garlic - 5 × 400 g tins chopped tomatoes - 1 × 200 g tube tomato purée - 1 small jar dried oregano - 1 bottle olive oil - 250 g butter - 500 g plain flour - 3 litres whole milk - 1 small jar ground nutmeg, optional - 1 kg dried lasagna sheets - 500 g mature cheddar - 150 g vegetarian Italian-style hard cheese - Salt and black pepper When you cook, measure out the recipe amounts rather than using the whole packs. Five batches need about 200 g butter, 200 g flour, 2.5 litres milk, 400 g cheddar and 100 g hard cheese, and you keep the rest. Make the white sauce in a large pan, adding the milk gradually and stirring constantly, since that much thickens slowly and lumps easily. Grate the vegetarian cheese and assemble the vegetarian tray first, with clean utensils, before you touch the beef. When you bake, cover each large tray with foil so the top doesn't burn before the middle is hot, give it longer than the recipe's 35–40 minutes, and expect to bake the two trays in turn rather than both at once. Let each rest a few minutes before cutting.

Condition	Answer shown to participants
Metacognitive	Buy: - 1.6 kg beef mince - 400 g vegetarian mince - 5 medium onions - 2 bulbs garlic - 5 × 400 g tins chopped tomatoes - 1 × 200 g tube tomato purée - 1 small jar dried oregano - 1 bottle olive oil - 250 g butter - 500 g plain flour - 3 litres whole milk - 1 small jar ground nutmeg, optional - 1 kg dried lasagna sheets - 500 g mature cheddar - 150 g vegetarian Italian-style hard cheese - Salt and black pepper The list estimates quantities in typical package sizes, so it might be more than you'll actually use. You should plan on using about four-fifths of these ingredients for the beef lasagna, except the mince, and using the remaining one-fifth for the vegetarian one. The recipe is written to serve 4, and scaling it up to roughly 20 takes more than multiplying the amounts of ingredients. At this size you're likely working in much bigger, deeper trays and pans, so things might not cook the same way. The bake time is the clearest example: the 35–40 minutes is for a small serves-4 dish, so a large tray will likely take longer. Its top might brown before the middle is hot. Judge it done by whether the middle is hot. Cooking a large batch of white sauce might be challenging. It might thicken slowly and lump easily, so watch and adjust it as you go rather than following the recipe verbatim. And making the dish properly vegetarian takes more than just checking ingredients. Vegetarian mince covers the filling, but the cheese matters too — traditional Parmesan is set with animal rennet, so it often isn't vegetarian. Cross-contamination matters as well: utensils that have touched beef can carry traces of meat, so it's safest to make the vegetarian tray first, with clean equipment.