greCAPTCHA: Assessing Understanding as Evidence of Research Authorship Under Generative AI

Justin Payan[1,*], Bálint Gyevnár[2,*], Atoosa Kasirzadeh[3], Nihar B. Shah[1,4]

[1] Machine Learning Department, Carnegie Mellon University

[2] Institute for Complex Social Dynamics, Carnegie Mellon University

[3] Departments of Philosophy & Software and Societal Systems, Carnegie Mellon University

[4] Computer Science Department, Carnegie Mellon University

[*] Equal contribution

{jpayan,bgyevnar,akasirza,nihars}@andrew.cmu.edu

September 2026

Abstract

Conferences, journals, funders, schools, and universities are struggling with a surge of potentially AI-generated submissions from ostensibly human authors, who may not have exercised sufficient human oversight for their manuscripts. In turn, institutions evaluating submissions can no longer reliably credit expertise based solely on authors' names on submitted work. To address this problem, we propose greCAPTCHA, a proctored assessment approach that measures authors' understanding of research manuscripts via the construct of capacity to verify, which we define as the knowledge and reasoning required to critically assess the contents underlying one's contributions to a manuscript. greCAPTCHA generates questions assessing multiple levels of understanding and provides an evaluative report based on authors' responses. Using a prototype implementation, we conduct a user study and semi-structured interviews with 31 researchers to evaluate greCAPTCHA. Its automated scores predict which papers were or were not authored by study participants with an AUC of 0.90. Participants reported positive overall experiences with the system and remarked on the appropriate construct validity for author understanding, while also suggesting important changes to be made before deployment. Our results provide initial evidence that greCAPTCHA can assess manuscript-specific understanding under proctored conditions.

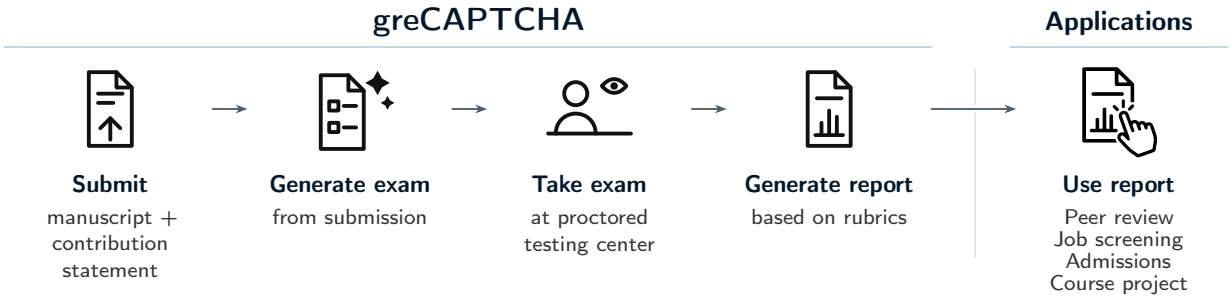

Figure 1. The process of assessing author understanding with greCAPTCHA. An author uploads a research manuscript and specifies their specific contributions, from which an exam is generated automatically. The author takes the exam in a proctored setting, either in a testing center (in a university, at a conference, or a dedicated center) or at home [Duo26; ETS26]. The results are then compiled into a report, which can be used for editorial decisions, applicant screening in academic hiring, student admissions assessment, grant proposal screening, or course project verification.

1 Introduction

While the definitions of research authorship are nuanced, they usually require an author to provide original content, substantial contributions, and assume accountability for a new piece of work [Ass26; Int26a]. Generative Artificial Intelligence (GenAI) makes it cheap to produce manuscripts that look polished without requiring the author to understand or verify their contents. This shift is exacerbating an existing gap between the quality of submitted manuscripts and the substantive knowledge of their authors. This gap is present wherever institutions primarily rely on submitted work to assess expertise or award credit, including in scientific publishing, academic hiring, grant evaluation, and education. While GenAI can assist in producing a manuscript, we think that for now responsibility for its claims must remain with the human authors.

The incentives and the expanding scale of academic submissions render this accountability gap particularly urgent. Acceptance at or rejection from prestigious venues can bring substantial career impacts, and submitting additional manuscripts is almost free. GenAI can further reduce the cost of creating these submissions, exacerbating the misalignment of incentives to prioritize quantity over quality. Meanwhile, rapid submission growth enabled by GenAI across research venues places increasing demands on human reviewers [arX26; ICL26; Tra26]. In education, similar concerns challenge the validity of written assessments and coursework projects that would serve as evidence of students' learning [CSK26]. These considerations motivate a pressing question:

> *How can institutions assess whether the people submitting a manuscript understand their contributions well enough to evaluate and take responsibility for them?*

1.1 Assessment for Research Authorship

There are various existing approaches to ensuring research integrity. Responses include restrictions on submission practices [Com26], sanctions for policy violations [Die26], and methods for detecting machine-generated content [Rao+25]. These interventions address different aspects of research integrity, but do not directly establish whether authors understand and can evaluate their contributions. In this paper, we propose a complementary approach based on authors' responsibility to verify the work for which they claim credit.

Drawing on the concept of *evaluative judgment*—the ability to judge the quality of work of oneself and others [Tai+18]—we propose *capacity to verify* as an assessment construct. We define this capacity as the knowledge and reasoning required to critically assess the claims, methods, and evidence underlying one's specific contributions to a manuscript. We operationalize it through two dimensions: *comprehension*, that is, understanding the relevant content and its basis; and *justification*, that is, explaining and evaluating the choices underlying the work. This capacity is a necessary prerequisite for verification, although demonstrating it does not establish that verification occurred or that the work is correct. This construct motivates our vision for a new kind of assessment:

greCAPTCHA. Authors demonstrate capacity to verify for their substantial contributions to a manuscript through an assessment tailored to the work and their contributions to it.

Borrowing its name from the Graduate Record Examinations (GRE) used in university admissions, we envision a greCAPTCHA as a proctored exam that asks authors to explain and critically evaluate relevant aspects of their manuscripts. Authors could complete proctored greCAPTCHAs in monitored testing centers at universities, at professional events such as conferences, or through certified at-home assessments [ETS26; Duo26]. Its name also alludes to CAPTCHAs or reCAPTCHAs used to distinguish human users from automated bots. However, unlike conventional CAPTCHAs, our objective is to distinguish authors who present AI-generated research as their own with no human contribution or oversight from those who contributed to or at least provided oversight of the research.

In greCAPTCHA, questions are tailored to authors' contributions, recognizing the division of expertise in collaborative work. Grounded in educational assessment [Daw20], our approach aims to elicit evidence of understanding while accommodating legitimate uses of GenAI, including language assistance [Gar+26], accessibility [Chh+26], and exploration [Asa+26]. Much like how students prepare for the GREs by learning, a

greCAPTCHA could encourage authors to study their work more carefully, thereby connecting accountability with a broader, positive goal: one of assurance of learning [Daw+24]. Although our experiment initially investigates scientific manuscripts, our proposal is relevant to educational and professional settings in which people submit work for which they are expected to take responsibility—such as academic hiring, student admissions, grant applications, and research proposals.

Whether greCAPTCHA offers valid indicators of actual human authorship remains an empirical question. Responses may reflect manuscript-specific understanding, but also domain expertise, question ambiguity, assessment conditions, or grading errors. An expert non-author may answer correctly, while a contributing author may struggle with questions that are outside their role. As such, an unsuccessful response does not by itself establish a lack of capacity to verify and should not be used categorically for determining a violation of submission policy. Instead, responsible development requires examining what a greCAPTCHA measures, the burdens it imposes on stakeholders, and how examinees' results could be interpreted and contested. As our initial investigation into its validity, we address two research questions:

RQ1. How does researchers' performance on greCAPTCHA differ between their own and unfamiliar papers, and how do these differences vary across question families and domain familiarity?

RQ2. How do researchers perceive the validity and acceptability of greCAPTCHAs, and what concerns do they identify about question quality, grading, deployment, and potential consequences?

1.2 Evaluating The First Prototype greCAPTCHA

We explore these research questions by running the first in-person greCAPTCHA with academics.

First, we design and implement a new prototype system which uses GenAI (GPT-5.6 Sol) to dynamically generate questions based on an author's manuscript and their stated contributions. In our prototype, authors read and respond to questions through a text interface, though an audio or video interface could be used in the future. Informed by the revised Bloom's taxonomy of educational objectives [Kra02], we design four question families which target factual, conceptual, and procedural knowledge through tasks that ask participants to identify errors, explain choices, demonstrate background understanding, and assess limitations. The corresponding question families are: (1) planted-error; (2) unstated rationale; (3) background knowledge; (4) failure mode.

We evaluate greCAPTCHA with a mixed-methods study asking 31 researchers each to complete a one-hour in-person session. In a counterbalanced within-subjects design, participants answered questions about one of their own papers and an unfamiliar paper selected by the research team. Unfamiliar papers were selected to be either within or outside participants' research fields, allowing us to examine the role of domain familiarity. Each assessment comprised eight questions, with two from each family, and a 15-minute response period. We then conducted semi-structured interviews to explore participants' experiences and their views on the validity, acceptability, and potential consequences of greCAPTCHAs. We report ROC AUC to quantify how well greCAPTCHA scores separate participants' own papers from unfamiliar papers across all possible score thresholds, without committing to a particular deployment cutoff.

Our main *quantitative* findings are:

- **High authorship separation:** Our prototype distinguishes performance on participants' own and unfamiliar papers with a high AUC of 0.90.
- **No statistically detectable field effect:** We find no statistically significant difference between scores on in-field and out-of-field unfamiliar papers overall.

Our main *qualitative* findings are:

- **Effective questions target less consequential knowledge:** Participants highlighted that deep questions about minor details are the best at providing evidence about capacity to verify in contrast to overall motivation or significance.

- **Assessment can promote learning and preparation:** Questions prompted reflection, a desire to learn, and even new insights for participants, some adding that it would be a valuable preparation aid for presentations. In some cases, disagreement with the system led participants to doubt their expertise—including the judgment needed to challenge the assessment.
- **Grading requires careful calibration:** Participants highlighted that grading and feedback requires careful design beyond the current prototype. The expected level of detail should be clear and the grader should be flexible if a participant gives an alternative but equally correct answer.
- **Participants support deployment but do not agree on role:** Participants largely supported deployment of greCAPTCHAs and offered insights on potential decision mechanisms, such as automated desk-rejection, score-based badges, and tiered screening. In addition, human oversight and accessible recourse mechanisms were broadly requested features.

These findings provide a promising initial empirical basis for further investigating greCAPTCHA assessments as a source of evidence about authors' understanding and accountability. The code to run our prototype implementation of greCAPTCHA as well as the anonymized responses of the 31 participants are available at `https://github.com/justinpayan/greCAPTCHA`.

2 Background and Related Work

We ground our research by first establishing what constitutes authorship in order to understand how changing practices of GenAI-enabled scientific writing impact the accountability of authors. We then review how academic institutions and venues are responding to these changes and how those responses can go awry. This information contextualizes our subsequent discussion of verification-relevant understanding and the automation of its measurement.

2.1 Authorship and Submission Policies

Scientific authorship has many definitions (see, for example [Ass26; Int26a]). While the exact details may differ among disciplines and venues, most criteria agree that authors are expected to provide accountable human judgment, not merely written content. Accountability is fundamental to greCAPTCHAs, not least because there are legitimate use cases for GenAI in scientific research. GenAI is often used for language assistance for non-native English speakers [Gar+26; Lia+23b; Liu+25; HZH25]. In addition, applications such as revisions and proofreading [Rat+25; Wan+26; Lia+23a], qualitative coding [Gao+24; Par+25; Qia+25], literature exploration [Asa+26; Rad+26; Ska+24; NM26], and accessibility [Aug+23; Chh+26] can be key reasons why GenAI should be allowed in scientific research.

While the use of GenAI for scientific research is not problematic by itself, its use does introduce a gap between authors and manuscripts which makes the use of text as evidence for authorship no longer admissible. This realization is partly what leads numerous venues to re-evaluate their submission policies. We organize these responses into four overlapping strands in Table 1: authorship and accountability, disclosure and verification, research evidence, and review integrity. For greCAPTCHAs, the central question is how these responsibilities can be assessed: what evidence supports judgments about compliance, and what can that evidence establish? Examining these strands helps us position manuscript-specific assessments within the broader effort to maintain research integrity.

The first strand, **authorship and accountability**, assigns responsibility for submitted work to human authors and establishes boundaries on the delegation of intellectual work to GenAI [Tho23; Int25; Int23; Com23; Int26b]. Authorship declarations, in which authors check a box or sign their name to indicate they accept the responsibilities of authorship, provide a record of an author accepting these responsibilities but not evidence of the understanding needed to fulfill them. Investigating suspected violations requires editors to separate legitimate assistance from inappropriate delegation, often without direct access to the research process. greCAPTCHAs could complement declarations by eliciting evidence of authors' understanding of

Table 1. Four overlapping strands of policies governing GenAI use in research submission and review. Summaries describe common rules, but requirements may vary across venues, publishers, and professional bodies.

Category	Policy summary	Example venues
Authorship & accountability	AI systems cannot be authors. Human authors remain responsible for all content. Scientific reasoning, interpretation, and judgment may not be delegated to GenAI.	Science [Tho23]; ICLR [Int25]; ICML [Int23]; COPE [Com23]; ICMJE [Int26b]
Disclosure & verification	Substantive GenAI use must be identified, though routine language editing may be exempt. Authors must verify generated claims, references, code, and analyses. Nondisclosure or unverified generation can lead to rejection or post-publication action.	ACL [BOR23]; NeurIPS [Con25]; CVPR [IEE25a]; IEEE [IEE24]; BMJ [BMJ26]; JAMA [FPB26]; USENIX [USE26]
Research evidence	GenAI may not fabricate or materially alter primary research images, data, or results. GenAI use in methods and for some explanatory graphics may be permitted when disclosed, validated, accurately captioned, and reproducible.	Nature [Nat26]; Elsevier [Els26]; Sage [Sag26]; Taylor & Francis [Tay26]; ACS [Ame24]; AIP [AIP26]
Review integrity	Reviewers may not upload confidential submissions to unapproved GenAI systems or delegate evaluation to GenAI. Limited language or background assistance may be allowed. Hidden prompts to manipulate AI-assisted review are misconduct.	ACL Rolling Review [ACL25]; ICML [Int26c]; ICCV [IEE25b]; CVPR [IEE26]; PLOS [Pub26]; IOP [IOP26a; IOP26b]

their stated contributions. Such evidence would inform the assessment of their capacity to take responsibility, without establishing who performed the original work.

The second strand, **disclosure and verification**, requires authors to report relevant GenAI use and check generated content, including claims, references, code, and analyses [BOR23; Con25; IEE25a; IEE24; BMJ26; FPB26; USE26]. Disclosures are used here to report tool use, and examining a manuscript can also reveal errors that warrant further investigation. However, neither establishes whether an author possesses the knowledge needed to evaluate the content. This distinction directly motivates our construct of capacity to verify because manuscript-specific questions may elicit evidence of the comprehension and reasoning needed for verification.

The third strand, **research evidence**, governs GenAI use in producing or modifying research materials and seeks to preserve the authenticity and traceability of scientific findings [Nat26; Els26; Sag26; Tay26; Ame24; AIP26]. Raw data, metadata, code, and provenance records can support investigations, though maintaining and inspecting them imposes documentation and inspection costs on authors and evaluators. This strand clarifies an important boundary of greCAPTCHA. Questions about methodological choices and limitations assess understanding, but cannot authenticate data or establish reproducibility. A knowledgeable author can still fabricate evidence. Assessments of understanding would therefore complement, rather than replace scrutiny of the underlying research materials.

The fourth strand, **review integrity**, protects confidentiality and human judgment in evaluation and prohibits manipulation of AI-assisted reviewing [ACL25; Int26c; IEE25b; IEE26; Pub26; IOP26a; IOP26b]. Although our study examines authors rather than reviewers, these policies also bear on the design of greCAPTCHAs: an assessment that uses GenAI to generate questions and evaluate responses must itself support trustworthy evaluation. In particular, generated questions and grades require scrutiny, manuscript confidentiality requires protection, and affected authors need meaningful ways to contest errors. Automating an assessment does not remove institutional responsibility for its quality or consequences.

Together, these strands distinguish requirements for responsible authorship from the evidence available to

Table 2. The four question families used in the greCAPTCHA evaluation, including response formats, descriptions, and example questions.

Question family	Response format	Description	Example
Planted-error detection	Multiple choice	Identify which version of a specific, verifiable claim accurately reflects the manuscript.	*"In validating the three-item ethical-concern scales, which construct had a Cronbach's α of 0.82? A) Autonomy; B) Privacy; C) Fairness; D) Transparency."*
Unstated rationale	Free-response	Explain why a methodological or design choice is appropriate, where the reason for that choice is not explicitly stated in the paper.	*"Table 1 places both sensing-not-in-use scenarios before all four sensing-in-use scenarios, while randomising only within those blocks. What supports this fixed ordering rather than randomising all six scenarios, and what interpretive cost does that choice create?"*
Background knowledge	Free-response	Define an important concept that the manuscript presupposes but does not define, and explain why that concept matters for understanding the reported work.	*"In your own words, explain the Mann–Whitney rank-sum procedure, including what it tests and the observation-level assumption it ordinarily makes. Then explain why this procedure matters for the condition comparisons reported."*
Failure mode	Free-response	Name a realistic, manuscript-specific condition under which the proposed method or central finding would degrade, and explain the mechanism producing that degradation.	*"Name one realistic learning condition for which the reported preference for system-generated hints over teacher assistance might weaken or reverse, and explain why."*

assess them. greCAPTCHAs address one part of this problem by eliciting manuscript-specific understanding relevant to authors' capacity to verify their contributions. Their usefulness depends on whether responses provide valid evidence of that capacity, how assessment errors and opportunities for gaming are handled, and who bears the resulting burden.

2.2 Designing Assessments of the Capacity to Verify

The previous policies establish a need for evidence about authors' understanding, but leave open how to elicit and interpret that evidence. We approach the design of greCAPTCHAs as a form of *assessment*, motivated by established practices such as oral examinations, thesis defenses, and interactive questioning. These practices ask people to explain and critically examine work for which they claim responsibility. Accordingly, greCAPTCHA seeks evidence of manuscript-specific understanding and judgment through participants' responses to questions about their work.

Following Dawson et al. [Daw+24], the validity of an assessment concerns whether evidence supports a particular *interpretation and use* of performance. Our intended interpretation concerns *capacity to verify*: the knowledge and reasoning needed to critically assess the claims, methods, and evidence underlying one's contributions to a manuscript.

Our construct is based on the concept of *evaluative judgment* [Tai+18]. This framework states that two main components must be present in order for someone to make decisions about the quality of their work: (1) understanding what constitutes quality; and (2) applying this understanding through an appraisal of work. Applied to manuscript assessment, this perspective motivates questions that elicit both comprehension of the work and justification of its underlying choices. We therefore ask participants to demonstrate relevant knowledge and use it to explain decisions, identify problems, and assess limitations.

Specifically, we use the revised Bloom's taxonomy to guide the knowledge and cognitive demands of the assessment [Kra02]. For this prototype, we focus on factual, conceptual, and procedural knowledge. We adapt these dimensions to the manuscript context as follows:

- **Factual knowledge:** relevant terminology, specific details, and reported findings.
- **Conceptual knowledge:** relationships among the manuscript's claims, underlying concepts, assumptions, and supporting evidence.
- **Procedural knowledge:** how methods and analyses were performed, and the conditions under which particular procedures are appropriate.

These dimensions inform four question families: planted-error, unstated rationale, background knowledge, and failure mode (see Table 2). The families provide different ways to elicit knowledge and reasoning, and individual questions may draw on more than one knowledge dimension. Questions are generated using both the manuscript and the participant's stated contributions to improve their relevance to the work for which the participant claims responsibility. Our first empirical analysis explores whether greCAPTCHA can distinguish author responses from non-author responses at the participant level. This comparison provides initial evidence about assessment performance and construct validity, but its interpretation also depends on how questions are generated, how responses are graded, and how participants experience the assessment. We next consider the implications of automating these processes.

2.3 Automating Authorship Assessment

The greCAPTCHA assessment approach outlined above requires questions tailored to individual manuscripts and authors' contributions. Producing and grading these questions manually would demand subject-matter expertise and could add substantial labor to existing evaluation processes. Automated question generation offers a potential means of supporting this work [MH03; HS10; DSC17; Kur+20]. For greCAPTCHAs, however, the challenge extends beyond generating fluent questions: questions must be relevant to the author's contributions, elicit the intended knowledge and reasoning, and support defensible judgments about responses.

A closely related approach is the AI-mediated *viva voce* assessment developed by Ebrahimzadeh, Shibani, and Buckingham Shum [ESB26]. Their system, VivaBuddy, uses GenAI to generate comprehension questions and provide dialogic feedback about students' work. The authors introduce *coauthorship integrity*, which concerns students' understanding of work produced in collaboration with GenAI. This approach shares our motivation to elicit understanding rather than infer it from the presence or absence of AI-generated text. Our proposed assessment extends their approach to all forms of research manuscripts where specialized subject matter, shared contributions, and professional responsibility shape what authors can be expected to explain.

Using GenAI to generate questions and grade responses introduces additional sources of uncertainty. Generated questions may contain mistaken premises, introduce material outside an author's contribution, or admit reasonable answers that an automated grader fails to recognize. Consequently, differences in scores may reflect properties of the assessment as well as differences in understanding. We treat automated scores as evidence whose interpretation depends on question relevance, grading criteria, and assessment conditions.

The potential consequences of automated assessment also require attention to fairness and legitimacy [Daw+24]. Language proficiency, career stage, disability, anxiety, and unequal testing conditions may all influence what participants can demonstrate. Proctoring introduces further questions concerning privacy, security, accessibility, and cost [Hut+26]. Any consequential deployment would therefore require consideration of accommodations, scrutiny of automated judgments, and meaningful opportunities to contest questions, grades, and resulting decisions. Work on algorithmic decision-making and contestability highlights the importance of these procedures to how people experience and evaluate automated judgments [Bin+18; Kar+24].

We therefore approach greCAPTCHAs as a sociotechnical assessment process whose usefulness depends on both the evidence elicited and the conditions under which it is interpreted and used. This perspective motivates RQ2: examining researchers' experiences and perceptions can reveal concerns about relevance, grading, burden, and legitimacy that performance comparisons alone could not identify. Our mixed-methods

evaluation of greCAPTCHA brings these perspectives together to investigate the promise and limitations of an initial implementation.

3 Assessing Understanding with greCAPTCHA

We describe the general procedure of administering a greCAPTCHA, then discuss the specific prototype implementation used in our study.

3.1 Design

The greCAPTCHA process translates the assessment approach described in Section 2.3 into four stages: **submission**, **generation**, **examination**, and **reporting** (Figure 1). We distinguish two roles: the *examinee*, whose understanding is assessed, and the *administrator*, who configures and administers the assessment.

Submission. The examinee submits their manuscript and a statement describing their specific contributions. The manuscript provides the assessment content, while the contribution statement identifies the parts of the work for which the examinee claims responsibility. Both inputs guide question generation so that the assessment targets knowledge relevant to the examinee's role. This process may vary across publication, coursework, and recruitment settings.

Generation. The system parses the manuscript and generates questions and corresponding grading rubrics using the contribution statement and the administrator's assessment configuration. This configuration specifies the question families, question counts, models, and prompts used for generation and grading. The administrator also establishes examination conditions, including time limits, response requirements, question ordering, and permitted resources. These choices should reflect the submitted work and the purpose of the assessment: the four question families evaluated in this paper are just one configuration of the broader greCAPTCHA approach.

Examination. The examinee answers the generated questions under the specified proctored conditions. The system presents the questions and records the responses for grading. In deployment, dedicated test centers are a possible venue for the assessment, while certified take-home exams—similar to those offered by Duolingo [Duo26] and TOEFL [ETS26]—could provide a lower-burden setup. Assessments might also be implemented at professional events, such as conferences, hackathons, or career fairs.

Reporting. On submission, the responses of the examinee are graded against the corresponding rubrics to produce scores. This process generates an assessment report that contains the participant's overall and per-question scores, their responses, how those responses matched the rubric, any partial-credit allocation, and feedback. The report is bundled with the manuscript and the examinee's contribution statement and then shared with the requesting institution according to the exam arrangements.

3.2 Implementation

We implement greCAPTCHA as a web application with two interfaces: an **administrator dashboard** and an **exam screen** for examinees. The dashboard supports assessment preparation and management, while the exam screen presents the generated questions and collects responses. This separation allows administrators to configure assessments without requiring examinees to interact with the generation settings.

Assessment preparation and management. An authenticated administrator uploads the examinee's manuscript and enters their contribution statement through the dashboard (Figure 2). The administrator selects model settings and randomization options and configures the question families used to generate the assessment. The dashboard supports a configurable number of families, each with its own generation prompt and question settings, such as the number of questions and, for multiple-choice items, the number of distractors. It also provides controls for timing and warm-up questions. The system supports multiple attempts on a question set, organization of question sets into sequential experiments, and question-set overviews.

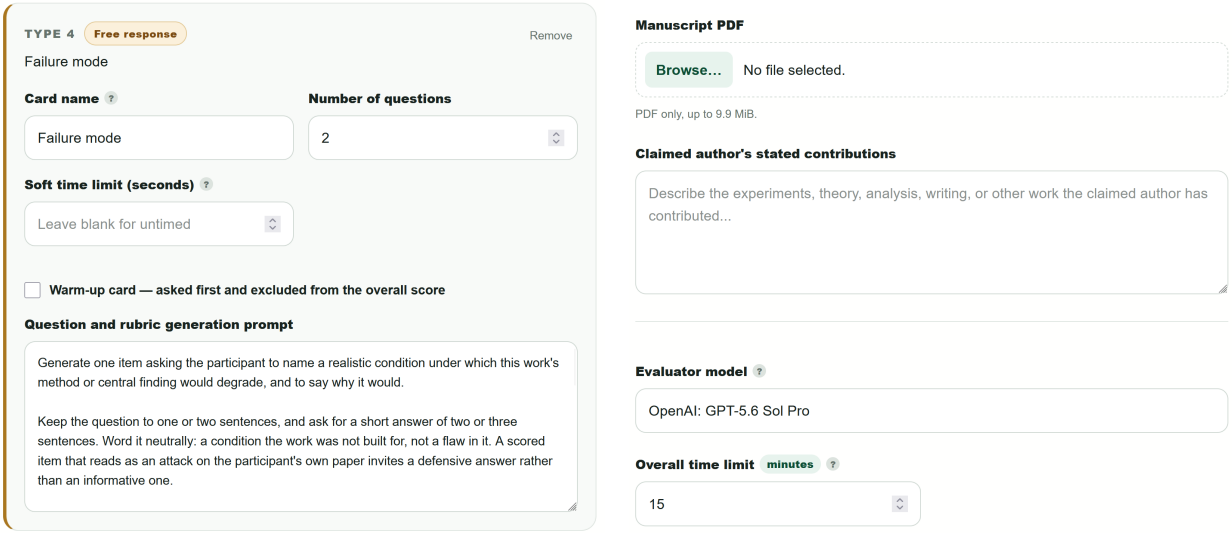

Figure 2. Screenshots from our greCAPTCHA implementation's administrator dashboard. The left image shows options for configuration of a question family, including its generation prompt and question settings. The right image shows manuscript upload, contribution statement, and assessment-level settings.

Question families used in the study. For our evaluation, we configure four question families informed by the assessment rationale in Section 2.2: planted-error detection, unstated rationale, background knowledge, and failure mode. Their response formats and examples are presented in Table 2. The complete generation prompts appear in Appendix A.

Unstated-rationale questions ask examinees to explain choices underlying the manuscript. Background-knowledge questions probe concepts needed to understand the work, while failure-mode questions ask examinees to reason about methodological limitations and circumstances in which an approach may fail. These three families use free-response formats. Planted-error detection uses a multiple-choice format, providing a structured task in which examinees identify the correct option among multiple incorrect options related to the manuscript.

The families are intended to elicit complementary evidence of understanding rather than isolate mutually exclusive types of knowledge. Explaining a design choice, for example, may require knowledge of the manuscript, its domain, and the research process. This overlap follows from the reasoning required by the questions, rather than from the free-response format alone. Planted-error detection provides a comparison with the free-response families, but performance on this task may mostly reflect attention.

4 Evaluation of greCAPTCHA

4.1 Interview Protocol

Given the above setup for greCAPTCHA, we recruit and conduct a mixed-methods evaluation with in-person participants. A timeline of a participant's experience from the start to end of the study is shown in Figure 3 and it has three stages: (1) Sign Up; (2) Session Preparation; and (3) In-Person Session. The study was piloted with four members of the authors' research group.

In the first stage, participants first receive information about the study and provide informed consent. Having consented to taking our study, participants then enter their primary research area, upload their *own paper* on which they contributed substantially, and describe their contributions to the uploaded paper.

In the second stage, we schedule individual one-hour-long in-person sessions with each study participant via email. We then select an *unfamiliar paper* to assign to them. We standardize this selection process

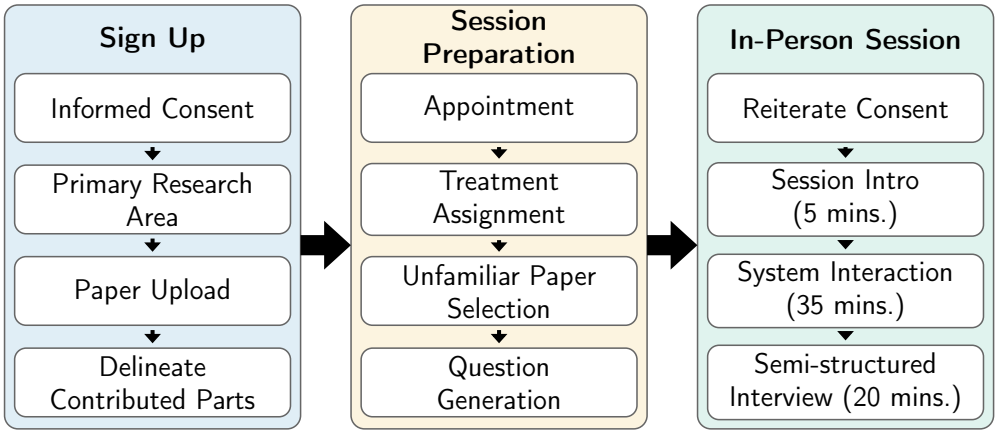

Figure 3. Our study workflow consists of three stages. The first stage requires participants to **Sign Up** using an online form, upload one of their own papers, and briefly explain which parts of the paper they contributed to. The second stage is for **Session Preparation** where we arrange appointments with each participant and assign them to experimental conditions. For the third stage we deliver the **In-Person Session** which involves a 5-minute introduction, a 35-minute interaction with greCAPTCHA, and a 20-minute semi-structured interview.

by prompting GPT-5.6 Sol with a prompt template (see Appendix B) to suggest five potential papers from the preprint server arXiv. As preprints can be low quality, we choose the paper which we judge highest quality based on an initial reading of the papers and the recognizability of their authors' institutions. Participants are then assigned to one of four treatment conditions based on two independent variables (IV)—one within-subjects and one between-subjects:

- **Paper Authorship** [within subjects]: this binary IV has two conditions: (1) own paper and (2) unfamiliar paper. The own paper condition tests participants on questions based on their submitted manuscript and contribution statement. The unfamiliar paper condition tests authors on the unfamiliar paper that was assigned to them. For this condition, the question generation system is told that the participant "did everything for the paper" as the contribution statement. We counterbalance the order of the two conditions.

- **Unfamiliar Paper Field** [between subjects]: this binary IV determines whether a participant is assigned an unfamiliar paper that is closely related to their field (*i.e.,* in-field) or that is distant from their field (*i.e.,* out-of-field). Field familiarity in this case means that the participant's training transfers to the unfamiliar paper. They are familiar with the notation, methods, and background knowledge, and they could read the paper without learning a new apparatus.

This results in a 2×2 design, where each treatment condition is a unique combination of paper authorship order and unfamiliar paper field. One day before the session, the selected unfamiliar papers are emailed to the participants. Prior to the interview, we generate the question sets for both own and unfamiliar paper. We generate eight questions, two from each of the four categories, and randomize the question order to avoid sequence effects.

In the third stage, during the study session, each participant is introduced to the system in 5 minutes. They are then instructed to interact with the greCAPTCHA system with a 15-minute hard time limit per paper, for a maximum total of 35 minutes, including 5 minutes of overhead. Participants are not told the exact purpose of the system beforehand. These examination conditions are chosen to facilitate in-person interviews on an academic budget. In realistic deployments a greCAPTCHA should involve more questions with a longer time limit to reduce variance. It must also be calibrated to the particular context of the examinees, taking into account the fairness, accessibility, and privacy of the participants. We record the following dependent variables:

Table 3. Post-interaction interview themes and representative follow-up probes.

Theme	Representative follow-up probes
Overall reaction	What stood out most or surprised you?
	At what points, if any, were you uncertain about what the system expected?
Grading and feedback	Were any correct or reasonable answers marked incorrect, or weak answers accepted?
	Did the explanations make the grading and supporting evidence understandable?
Question quality	Which questions provided the strongest or weakest evidence of understanding, and why?
	Could someone answer correctly without understanding the paper, or understand it yet answer incorrectly?
Paper-specific assessment	For your paper, did the questions reflect the portions you contributed to or verified?
	For the unfamiliar paper, did the system distinguish field knowledge from familiarity with that specific paper?
Perceived purpose and construct validity	What did the system appear to measure: understanding, verification, personal contribution, or something else?
	Did it unintentionally measure memory, test-taking ability, English fluency, or speed?
Circumvention and strategic behavior	How might authors prepare for or work around the system without verifying their papers?
	Which parts would be hardest to game, and would countermeasures create privacy, fairness, or usability concerns?
Trust and deployment	How would using such a system affect your trust in a conference or journal?
	What role, if any, should it have when it is uncertain or appears to make a mistake?
Recommendation	Would you recommend that a colleague's conference adopt the system?
	What single change would make it more acceptable or effective?

- **Scores:** The score assigned by the automated grader to each of the participant's answers on both own and unfamiliar paper. The average score over all questions in a test is also a measure of interest.
- **Skipped:** Which questions were skipped by the participant.

Following the interaction, we conduct a 20-minute-long semi-structured interview with the participant, covering their overall experiences with the system, their opinions about questions and evaluation quality, and their recommendations about whether and how to deploy the system in the broader academic system, including peer review and admissions. The interview topics are summarized in Table 3, and our full interview script is available at `https://github.com/justinpayan/greCAPTCHA`. Participants were then compensated $30 via bank transfer (using Venmo/Zelle) for their time. This study was approved by the Institutional Review Board of Carnegie Mellon University under study number `2026_333`.

4.2 Participants

A total of 31 participants completed the study. Subjects were recruited via email, social media, message boards, in-person solicitation, and through word-of-mouth from the academic and research institutions in Pittsburgh, PA, USA. A distribution of their relevant background information is shown in Figure 4 with the full underlying information displayed in Appendix C. Participants primarily came from cognitive science and computing disciplines, with at least one each from philosophy, psychology, hydrology, and environmental

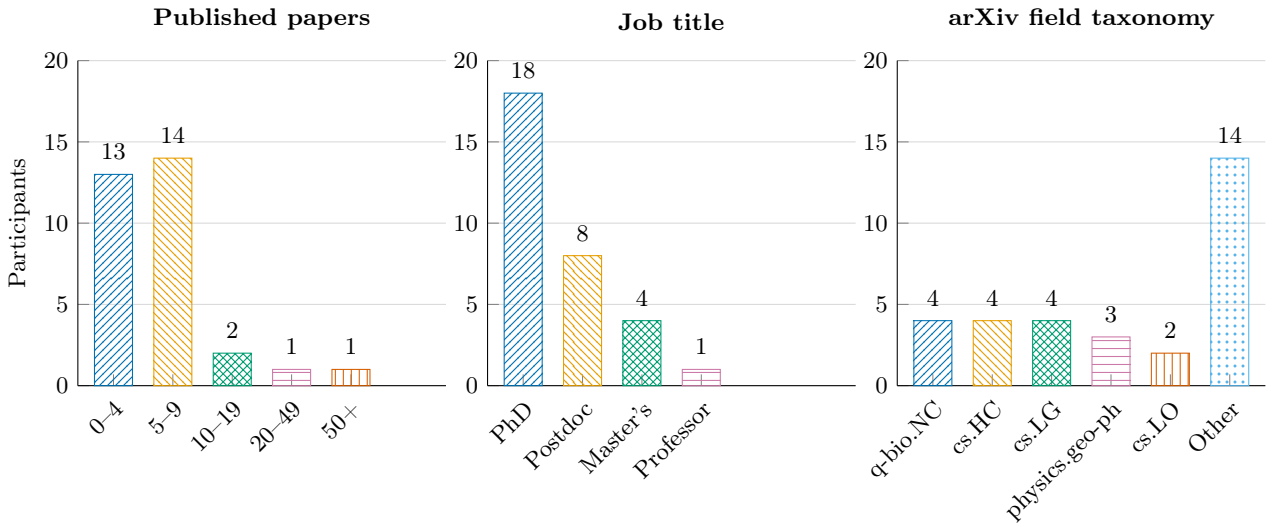

Figure 4. Participant characteristics ($n = 31$). arXiv fields: q-bio.NC = Neurons and Cognition; cs.HC = Human-Computer Interaction; cs.LG = Machine Learning; physics.geo-ph = Geophysics; cs.LO = Logic in Computer Science.

engineering. Most participants were PhD students, followed by postdoctoral researchers (or equivalent industry positions), and the majority of our participants had fewer than 10 public manuscripts. Participants were assigned equally at random to one of our four treatment conditions, giving eight participants per condition, except for the own-paper-first out-of-field condition, which received 7 participants.

4.3 Statistical Methods

We address RQ1 by computing the Receiver Operating Characteristic (ROC) curve and calculating the Area Under this Curve (AUC) from the overall scores assigned to participants. The ROC is especially helpful to understand how well our greCAPTCHA separates true positives (*i.e.,* non-authors classified as non-authors) from false positives (*i.e.,* authors misclassified as non-authors). We also view the distribution of scores assigned within each question family, and construct an ROC curve and AUC for each question family.

To explore the effect of familiarity with the unfamiliar paper's field, our second analysis is restricted to responses only about unfamiliar papers. Given these data, we examine whether performance differs between participants assigned to the in-field and out-of-field conditions. We normalize scores to the range $[0, 1]$ by dividing by 100. Because scores are bounded and clustered around 0 and 1 for the planted-error questions, we fit an ordered-beta mixed-effects regression model to our data [Kub23]. This model predicts authors' scores from fixed effects for the unfamiliar-paper field and question family, and includes random intercepts for each participant: `score ~ field_type + question_family + (1 | participant_id)`. We estimate the contrast between in-field and out-of-field performance, testing the null hypothesis that this contrast is zero. Statistical modeling and figure-generation code were generated using GPT-5.6 Sol and verified by the first authors.

4.4 Qualitative Methods

The individual scores of participants measure their particular performance on our question sets under specific conditions. We contextualize these scores by eliciting participants' views about their experience with greCAPTCHA.

To do this, we make audio recordings of each semi-structured interview. We transcribe and diarize each recording using the WhisperX system [Bai+23]. We further preprocess the transcripts using deterministic rules and a language model constrained to removing only excessive usage of filler words, fixing punctuation,

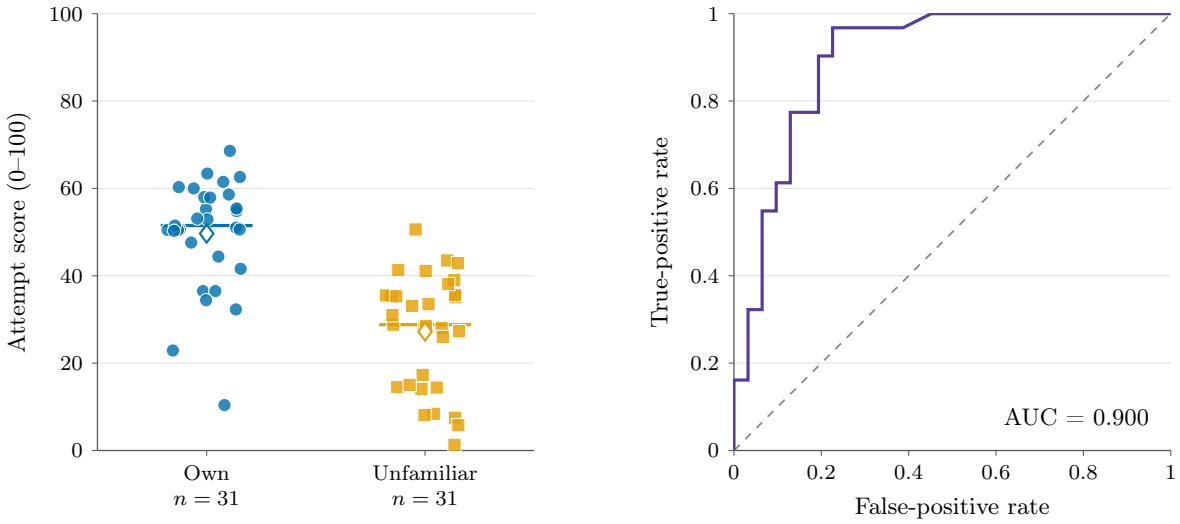

Figure 5. The left plot shows overall score distribution broken down by whether the paper tested is the participant's own paper or the unfamiliar paper ($n = 31$ per category). Lines indicate medians and diamonds indicate means. On the right, we plot the ROC for prediction of "unfamiliar" label, using $(1 - \text{score}/100)$ as the predictor.

and to align the audio timestamps with speaker boundaries. The preprocessed transcripts were then verified against the raw text to ensure that no words were invented or silently altered. Speaker boundaries are manually verified against the recordings by one of the authors when it is ambiguous who is speaking. All deep learning models were run on a local machine.

We perform a hybrid deductive–inductive codebook thematic analysis, organizing our codebook with the open-source software QualCoder [Cur+25]. First, two authors independently developed deductive codebooks based on the interview themes in Table 3 and on skimming the transcripts; these were then consolidated into a single codebook. In addition, we also perform inductive coding. Once data collection finished, the same two authors independently coded the same four randomly selected interviews, then discussed the applications of inductive codes. Finally, the authors merged a combined inductive codebook with their deductive codes. This final codebook, along with descriptions of codes and annotation counts, is available in Appendices F and G.

One author then coded all 31 interview transcripts—recoding the previously selected four transcripts. The first authors then collaboratively organized the coded passages into candidate themes and refined the themes through discussion. We connect these themes back to our quantitative analyses by organizing participants' data into cases and analyzing their distribution along our identified themes. This involves examining participants' accounts within each theme alongside their own-paper and unfamiliar-paper scores and their unfamiliar-paper field assignment. This helps us identify explanations for observed performance patterns as well as analyze cases of disagreement.

5 Quantitative Results

Recall from Section 2.2 that our quantitative analysis set out to explore two related questions. First, we examine the trade-off between correctly identifying non-author responses and incorrectly classifying author responses as coming from non-authors. Second, we look at how domain familiarity may affect the results of greCAPTCHA.

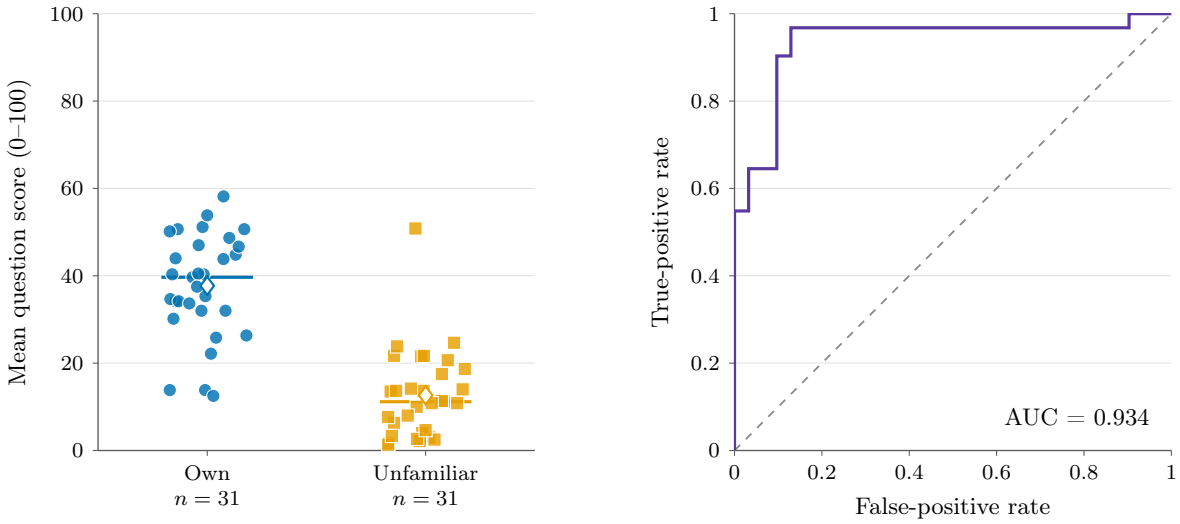

Figure 6. The left plot shows overall score distribution, *excluding planted-error multiple choice questions*, broken down by whether the paper tested is the participant's own paper or the unfamiliar paper ($n = 31$ per category). Lines indicate medians and diamonds indicate means. On the right, we plot the ROC for prediction of "unfamiliar" label, using $(1 - \text{score}/100)$ as the predictor.

5.1 Author Separation Accuracy

We show the distribution of scores on the participants' own papers and on the unfamiliar papers in Figure 5. The mean score on participants' own papers was 49.7, and the mean score on unfamiliar papers was 27.2. On the right side, we compute $(1 - \text{score}/100)$ as the class probability for the "Unfamiliar" label for all 62 tests, plot the ROC curve, and report the AUC. For this data the AUC turns out to be 0.90, indicating high discrimination between own-paper and unfamiliar-paper scores.

We also compute the average score on each of the four question families within each test, and plot the distribution of average score within each question type for own and unfamiliar conditions. These plots are included in Appendix D. We find that the AUC of the planted-error question family (the only multiple-choice family) is very low at 0.595, which is not surprising, since it primarily measures attention: answers can be found by searching the manuscript directly. In Figure 6, we see that the AUC increases to 0.934 when excluding this question family from the score calculation.

These results suggest that our implementation of greCAPTCHA can separate participants who have the capacity to verify from those who do not with a best false-positive rate of around 20%. This rate is lower than typical error rates involved in research assessment. Prior work on the reliability of peer assessment indicates that expert human reviewers can disagree on relative rankings of papers between one-fourth and one-third of the time [Sha+18]. Nevertheless, this rate should be minimized as much as possible to minimize unnecessary manual human verification and improve user experiences. In our qualitative analysis, we show that the main reason for the false-positive rate is the grading rubric rather than question quality.

Supplemental statistical modeling is reported in Appendix E, showing statistically significant differences between scores on all 4 question families across own and unfamiliar paper conditions.

5.2 No Evidence of Differentiating Between In-Field and Out-of-Field Papers

To understand the effect of field familiarity on participant performance, we compare the population-level differences between participants in the in-field unfamiliar-paper condition and the out-of-field condition. The test statistic measuring the difference between in-field and out-of-field scores is estimated to be 0.160 (average in-field scores are higher) with the corresponding p-value estimated at 0.437. We can also see

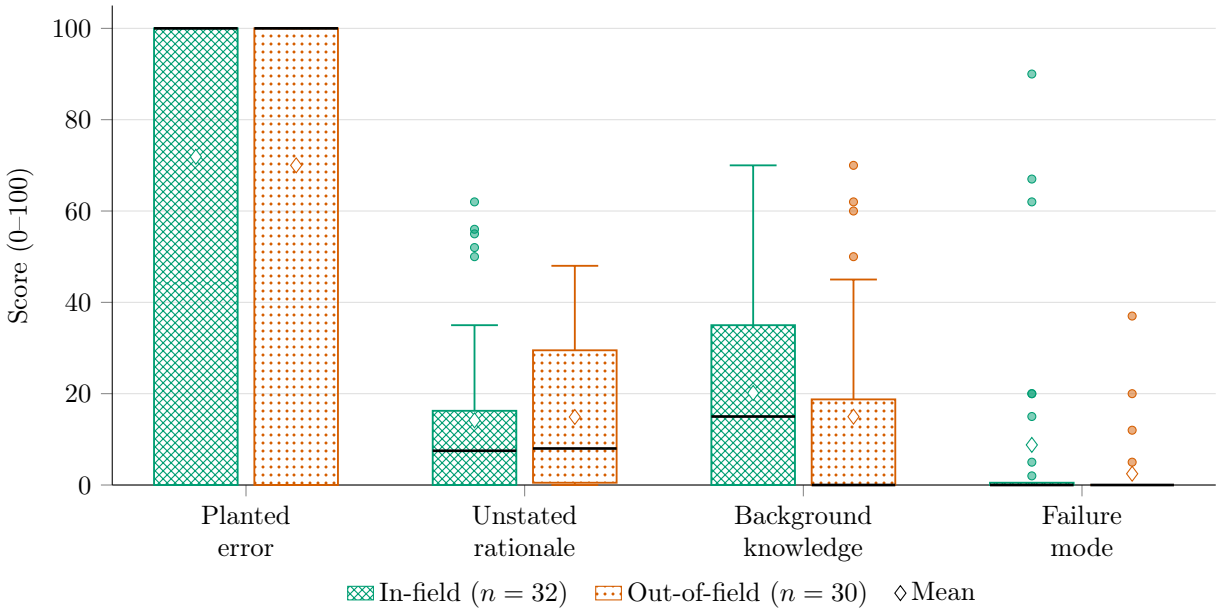

Figure 7. Distribution of unfamiliar-only paper scores grouped by question family and field type. Diamonds indicate means. The overall effect of field type is not statistically significant ($p = 0.437$).

from Figure 7 that the score distributions largely overlap between the in-field unfamiliar-paper and out-of-field unfamiliar-paper conditions. While not definitive proof, this finding does suggest that performance on the question sets generated by our implementation of greCAPTCHA is not dependent on whether the participant is an expert in the field of their assigned unfamiliar paper. In other words, greCAPTCHA seems to be more than just a "field detector".

As explored further in Appendix E, the failure to discriminate between in-field and out-of-field conditions does not appear to be a result of insufficient sample size. A similar analysis shows statistically significant differences between the mean score achieved on the familiar and aggregated unfamiliar papers across all 4 question families.

One question family where we see a clearer separation between field types is background knowledge. Here, the scores of in-field participants tend to be higher than those of out-of-field participants. This finding makes sense intuitively because one would expect domain experts to understand their own field's background better than non-experts. It also highlights the importance of having multiple question families targeting different types of knowledge.

6 Qualitative Analysis

Our qualitative analysis identified 18 deductive codes such as "Perceived validity," "Question clarity," "Trust change," and "Feedback relevance." The full codebook with occurrence frequencies is listed in Appendix F. We inductively identified 61 codes which we organized into 16 categories. The complete inductive codebook is provided in Table 14. Here, we interpret our coded excerpts alongside the 31 participant cases (P1–P31; see Appendix C for IDs).

All codes are assigned to individual statements by participants, so it is possible for all deductive and inductive codes to appear multiple times in a single participant's interview transcript. It is also possible for seemingly contradictory codes to be annotated in the same transcript, at different times. For instance, one participant may comment on the strong clarity of one question while also remarking on the poor clarity of another. Our analysis identified four major themes of feedback from participants regarding:

- **High-level vision (Section 6.1):** Responses about the high-level vision.
- **greCAPTCHA design choices (Section 6.2):** Responses to conceptual design choices, such as the four question families or the overall test structure.
- **Implementation details (Section 6.3):** Responses to detailed implementation choices, such as exact question wording, time pressure, or complaints about the UI design.
- **Deployment (Section 6.4):** Suggestions for how greCAPTCHA should be deployed.

In each category, we lay out summary points for the positive feedback (+) and negative feedback (–), provide the evidence for each summary point, and give additional context and discussion where necessary.

6.1 Vision

(+) greCAPTCHA assesses ability to verify. Participants mostly agreed that greCAPTCHA is a valid way to assess the capacity to verify. Some highlighted that one would have to go through the actual process of doing the research in order to perform well on the exam (P4, P6, P8, P13, P28, P30). Participant P6 said:

> *"It would require someone who understands the paper very well, or at least understands the field extremely well, to be able to answer those questions."*

(+) greCAPTCHA is hard to circumvent. A total of 12 participants noted that it would be challenging to prepare for a greCAPTCHA if they had not written the paper. For instance, P13 reported using AI to walk through the unfamiliar paper beforehand, yet was unable to answer in-depth questions, receiving a score of 1.3 on the unfamiliar paper compared to 51 on their own paper: *"I just gave it to an AI and asked it to walk me through the whole paper"*. Participant P8 commented:

> *"You needed to know the paper or be able to justify why a certain choice was made in the paper. And as someone who did not read the paper, I could not do that."*

(+) Enthusiasm for the idea. 23 participants expressed strong positive overall experiences and enthusiasm for the system, and only 6 expressed negative experiences. P28 thought it was important to:

> *"tackle this problem of people submitting a lot of papers."*

P7 expressed befuddlement at the prospect of people submitting work for which they do not have the capacity to verify:

> *"Why would you go to a conference where people present their papers, but then they don't understand. You might as well just stay at home and read the papers."*

(+) Participants think greCAPTCHA should be deployed. 16 participants expressed some sentiments that greCAPTCHA should be used in deployment, 12 expressed that it should be deployed after some significant changes, and 7 expressed sentiments consistent with not deploying. Participants sometimes expressed contradictory sentiments at different times in the interview. For example, P3 expressed initial frustration due to the low evaluation on their own paper, and commented on the confusing nature of some of the questions. Yet when explicitly asked about deployment later in the interview, P3 stated that the motivation is great, and agreed that a future version of greCAPTCHA should be deployed. Participant P12 said:

> *"If NeurIPS did it, I think it'd be great."*

Participant P18 preferred greCAPTCHA to the version control history solutions currently being employed:

> *"I honestly would rather see something [like] this as opposed to version control."*

(+) Deployment would increase participant trust. 22 participants expressed sentiments consistent with the view that greCAPTCHA's deployment would increase their trust in a peer-review venue. For example, P4 mentioned that deployment:

> *"would definitely raise my opinion on the quality of work."*

and P9 said:

> *"I think it would make me feel a little bit more confident in what I'm reading."*

(+) Participants guessed the purpose of the system. Many participants could guess the purpose of the system, even prior to telling them directly. 13 participants guessed the exact purpose of the system and 12 participants guessed closely, for instance, mentioning the system was intended to measure deep understanding of the paper and/or field.

(−) Some thought reading the paper or having background knowledge of the field was enough. Some participants felt that people may be able to achieve high scores on questions if they had read the paper and/or were familiar with the field.

P13 distinguished understanding a paper from having contributed to it, and P30 provided a counterexample to treating high scores as authorship evidence: after five minutes of preparation, their unfamiliar-paper score was 50.6, close to their own-paper score of 53.1. They recognized the unfamiliar paper as relevant to their research area, scoring 76.0 on its failure-mode questions versus 45.0 on their own. The most senior interviewee by paper count, P12, recalled undergraduate experience while answering an out-of-field medical-physics paper, and scored 55.0 on greCAPTCHA's background-knowledge questions. Variation in performance may depend on transferable knowledge and long-term experience.

(−) Deployment would require extra work from scientists. Although they urged deployment in general, 7 participants also stressed the importance of maintaining human oversight over decisions. This was often expressed as the desire to be able to contest grading decisions, or to have some ranges of scores under which papers would be flagged for human review rather than automatically rejected. However, even this was not unanimous, as P2 stated that they would prefer all decisions to be made autonomously, with results included as "badges" for published papers.

In general, participants emphasized the additional human labor required to maintain human oversight, in addition to the labor of the proctors (P11, P23). However, it is possible that the additional labor required to monitor greCAPTCHA and field complaints would be less than the labor currently expended on full review of large numbers of AI-generated papers.

6.2 Conceptual Design Choices

We also asked participants about the conceptual space explored by the questions, and what the grader should assess.

(+) Free-response questions tested deep understanding. Overall, 19 participants expressed comments that some questions strongly tested deeper concepts, requiring participants to apply, critique, or reconstruct ideas beyond text recall. These questions were usually unstated rationale or failure mode questions. P26 summarized this sentiment acutely:

> *"You're reflecting on your own choices and why did you choose path A instead of path B and to really make us ground our decisions in factual data."*

Participants expressed that questions which test deeper understanding tend to be better at assessing ability to verify. Spans of the `strongly tested` level of the `Question depth` code category overlapped with spans of the `agree` level of the `Perceived validity` code category 9 times in 5 interviews.

(+) Opportunity for learning. The value of greCAPTCHA extended beyond assessment. P25 remarked of a question, *"I'm shocked in the number of times I've presented this research that no one has asked me this question,"* (P25) and indicated that it would be useful to think about further. P20 acknowledged that the system found genuine gaps in their understanding and requested we send them their greCAPTCHA questions for further study: *"Can I get a copy of these questions?"* (P20). P26 similarly interpreted the encounter as a lesson to understand why they made particular decisions. Low scores did not preclude these reactions: P20 had the lowest own-paper score, and P26 scored 34.4.

Some participants even proposed changing the purpose of the system. P15 articulated how their attitude would change accordingly: *"I think if it was teaching me, I would be much more responsive to it."*

> *"If you would generate a paper, I guess you have to learn what's in the paper and then try to learn about the literature that it's drawn from."*—P7, on how to prepare for a greCAPTCHA.

(+ and −) Some participants performed poorly on some/all questions for their own papers. Some participants scored poorly on background knowledge questions in their own papers. P9 acknowledged not fully understanding statistics initially supported by a mentor and recognized that some gaps remained. P26 described following established statistical practices without fully understanding their rationale: *"We just know we have to use this statistical stuff for our results. So we do it, we don't really understand super why sometimes."* (P26). Their own-paper background-knowledge scores were 0.0 (P9) and 17.5 (P26). These results appear to be a correct assessment of inability to verify concepts. As mentioned above, some participants took their low scores as an opportunity for learning.

In contrast, P3 experienced a sharply different meaning of their own low score (22.9). Confronted with a rubric that rejected their account of something they had devised, they protested, *"I made this up myself, I made this myself!"* (P3). P10's difficulty with a closely related unfamiliar paper resulted in uncertainty about their own competence: *"Am I misunderstanding my field? Am I misunderstanding the papers?"* (P10). Their unfamiliar-paper score was 8.4, with 30 minutes of preparation, and they even suspected that the unfamiliar paper had significant flaws but then hesitated to trust their judgment. This encounter illustrates how an apparently authoritative assessment can make a researcher doubt the expertise needed to challenge it.

Three participants scored worse on their own papers than the unfamiliar papers (P1, P3, P20). These were junior participants. This probably indicates a true lack of understanding of their fields (+), but we should be careful about punishing junior authors too harshly for this. We see some evidence that authoritative and negative AI-generated assessments may be discouraging for junior authors (−).

(−) Some questions were considered irrelevant. P6 described an algorithm question that essentially only they could answer, yet called it a detail that *"doesn't really matter for any of the big conclusions in the paper"* (P6). Other participants also felt that questions were not relevant (mentioned by 16 participants in 41 instances). P29 noted *"[The system] has the tendency of going to technical details and picking out some detail and then ask you just another detail, but miss the bigger picture"* (P29).

Participants were not told the purpose of the test was to assess capacity to verify. Questions which asked about specific details may help to discriminate authorship or ability to verify without being perceived as "relevant" to the author. This trade-off between relevance and assessment capacity is important to consider during deployment, but it does not necessarily detract from the test's primary purpose of assessing understanding.

(−) Multiple-choice questions are too easy. When participants remarked on questions that were too easy, they often mentioned the ability to *"control-F"* (that is, search) terms in the question (P5, P8, P9, P11), specifically citing the multiple-choice questions (P14). This intuition corroborates the quantitative analysis showing that the multiple-choice questions are not discriminative and even *harm* predictiveness of the evaluation. On the basis of both the quantitative and qualitative results, we recommend removing the multiple-choice questions in future iterations of greCAPTCHA.

(−) Need to clarify grading expectations. Participants' comments on the grader were generally negative.

Overall, 23 participants described uncertainty about the expected level of detail in responses. P14 imagined a conversation at a conference poster session, while P6 explained concepts assuming little technical background. P16 gave a succinct summary of the core reason for this variance:

> *"What audience am I explaining these answers to? [...] The way that I would explain it to a student who is in their first course, or am I explaining it to a professor who is talking about this specific nitty-gritty paper?"*—P16, on being unclear who is the expected recipient of their answer.

Participants' individual choices affected what they made explicit. P11 objected to having to guess the expected answer or restate implications they considered obvious. P25 offered a counterpoint: *"the text [manuscript] is all the grader has"*, suggesting that from an examiner's perspective, omitted reasoning cannot automatically be inferred or credited. This is an important gap that concerns the evidence required to demonstrate the capacity to verify: disciplinary conventions may be meaningful to a colleague but are likely to remain insufficient for the grading of a rubric.

Overall, unclear expectations led to 16 participants disagreeing with some aspect of the grading, while only 8 expressed agreement with the grading. Several participants focused on the very strict allegiance to the rubric (P9, P13, P20, P21). P21 summed this up by saying: *"I think if I would imagine a human reading this compared to the tool, I would picture that the human is more charitable."* On the other hand, P8 found the system provided explanations helpful, and P25 even accepted some deductions.

The feedback on the grading suggests several improvements to greCAPTCHA:

- The expected level of detail should be stated clearly at the beginning of the exam, and the questions and rubrics should be generated using this statement.
- If something is included in the grading rubric, it should be clearly signaled in the question text. The paper's author should be able to respond to the questions exactly as phrased and get 100%.
- The grading rubric should be more flexible and award points for answers that are not precisely what was initially expected but are nevertheless correct.
- It may also be possible to assess answers not by their correctness to the letter of the questions, but by the likelihood that the examinee understands the paper. An examinee may misunderstand the question text, but still demonstrate strong understanding of the paper in an ostensibly incorrect answer.

6.3 Implementation Details

We discuss responses to low-level implementation decisions, such as the UI design or exact question phrasing.

(+ and −) Time pressure. Time pressure also appeared in 20 participants' coded accounts. Its consequences depended on both familiarity and strategy. P20 reported spending half the available time on the first own-paper question. Overall, we recorded three own-paper timeouts at an average score of 10.4. P1 described how time pressure restricted them from elaborating familiar material and led to a consequential typo. P15 mistakenly believed skipped questions could be revisited, although the interface clearly stated that skipping a question would result in a grade of 0 on that question. Their unfamiliar-paper attempt contained six skips, resulting in a score of 17.3. P15 also reported profound unfamiliarity, so the low score cannot be attributed solely to the interface. Such cases show how knowledge, pacing, and interface interpretation jointly shape the recorded response. These factors made some participants perform poorly on their own paper and others perform poorly on the unfamiliar paper.

Our time was limited during the study, and we could afford participants only 15 minutes per paper. However, an actual deployment may allow examinees more time to answer the questions.

(−) Question wording was often unclear. Participants also often felt that question wording and the expected response were unclear. 21 participants discussed this topic. For example, P18 mentioned that for one question, *"it's hard to verbalize whether or not you understand the message or even relay: 'yes, I know what this detail means', because it's buried in the question."* P18 discussed how some questions had multiple parts that were not organically connected to each other, making them difficult to answer adequately. As a result, they had to take extra time to think deeply about their response, which can be challenging during a timed exam.

LLMs should be prompted to ask about only one aspect per question and to avoid lengthy sentences with complex structures.

(–) Three participants got questions they couldn't answer but co-authors could. Question specificity may also create difficulties for collaborative authorship. P5 distinguished their theoretical knowledge from a co-author's practical implementation decisions; P19 described how they contributed an implementation without knowing the reasoning behind some metrics; and P29 encountered material written by a co-author in the greCAPTCHA. Their own-paper scores were 55.3, 50.6, and 47.6, respectively. P5's response was *"I should go ask my co-author why he did it this way"*. These accounts highlight the importance of accurately matching participants' stated contributions with greCAPTCHA's questions. We may wish to change the way we ask about contributions to make it clearer that you will be tested on what you state you can be tested about, so you should indicate exactly what technical elements your co-authors contributed that you don't understand. While our participants were blinded to the purpose of the study, knowledge of its purpose may result in participants describing their contributions more clearly in the initial submission form. Finally, in the test itself, we could make it possible to excuse oneself from certain concepts during the exam, for instance by responding "I didn't work on this part" to a question. We can give examinees a limited budget to do this.

6.4 Deployment

As discussed in Section 6.1, participants largely agreed with deploying a future iteration of greCAPTCHA. However, we heard a wide range of opinions about how exactly to implement and deploy greCAPTCHA in practice.

Participants had contrasting visions for how greCAPTCHA reports should be used in real workflows. Some participants expressed support for entirely automating some decisions (P2, P7, P8, P16), especially in the most egregious cases. P2 preferred converting AI-assigned scores into badges for papers, based on defined score thresholds. P26 noted that setting an acceptable threshold would be challenging. Five participants suggested that there should be three categories of papers: those that are automatically desk rejected, those that require additional screening, and those that are automatically passed into the primary reviewing pool (P7, P13, P14, P15, P20). Several participants preferred maintaining an appropriate level of human oversight before making decisions (P7, P14, P26, P28). P14 suggested that humans could edit questions and rubrics prior to assessment, and also suggested that greCAPTCHA scores be used for metareview. P18 emphasized that *"to implement a tool, there would have to be some good faith process to contest"*. Other participants noted that simply implementing the assessment may have a deterrent effect, either due to the extra work involved or due to the cognitive effects of receiving a low score (P3).

> *"If I had to choose between a conference that allowed me to just submit a paper and [...] a conference that required me to take a half an hour test on my own paper, I would have to really be invested in that conference to do so, which might be a good thing because there's oversubmission problems at conferences."*—P18, on the deterrent effect of greCAPTCHA.

These proposals frame disclosure itself as a consequential design decision. P18's call for a good-faith contest process and P28's preference for human evaluation place responsibility for resolving uncertainty with the institution, although P24's comments color this approach: *"it is extra work for people who care, who try to do ethical research. But someone who doesn't care [...] that [greCAPTCHA] is still not going to stop them."* Several participants suggested that reviewers be able to see the system report (P16, P20, P21, P27, P28). However, other participants felt strongly that *"flagging it for reviewers is not helpful because then you unnecessarily bias your reviewers"* (P11).

Other participants suggested modifying the modality of the assessment (3 participants) or allowing back-and-forth interactions (7 participants). They frequently cited a desire to clarify their answers or push back on the assessment. They also discussed a desire to request clarification on question texts during assessment. These concerns would be easier to address in an audio or video format.

Finally, some participants emphasized the potential unintended measurements beyond the capacity to verify. These concern how knowledge is elicited under examination conditions. Three participants raised issues with language use and reading speed (P4, P18, P21) regarding accessibility for non-English speakers and people with disabilities. P4 summarized the distinction: *"just because I can't write it in the same way that the question is expecting doesn't mean that I don't know about the work."* Typing-related remarks appeared in

12 participants' coded accounts. P19 described the task as testing *"both my knowledge of the paper and also my ability to info-dump about it."* Relatedly, three participants explicitly linked performance to how well people can search for information in a document (P7, P18, P27). P18 described the effect on performance of *"how efficiently they can use the search function in a PDF reader"*, while P8 offered a more favorable interpretation: knowing where to find an answer quickly can itself reflect familiarity. Recency may also play a role. P1 associated difficulty with having left their research area for industry, and P19 mentioned that two years had passed between finishing their project and its acceptance, during which intervening projects made decisions harder to recall. P12 provided a counterexample: *"I'm giving myself a lot of free credit for a paper I haven't read in 12 years"*, yet their own-paper score was 58.6, compared with 32.3 for P1 and 50.6 for P19.

7 Conclusions

In this paper, we introduced greCAPTCHA, an assessment method designed to measure an author's understanding of a research manuscript. We propose that this assessment serve as evidence of research authorship. We measure this evidence via the construct of "capacity to verify," which we define as the necessary knowledge and reasoning skills required to critically evaluate the contributions for which authors claim responsibility. We developed a prototype of greCAPTCHA and evaluated its usefulness via an in-person user study and semi-structured interviews with 31 participants. Using GPT-5.6 Sol, greCAPTCHA generates tailored questions to assess each author's distinct contributions to their manuscript, using an evaluation tailored directly to their work. These questions are designed to measure the authors' understanding of a manuscript at four distinct levels: planted errors, unstated rationales, background knowledge, and failure modes.

To evaluate greCAPTCHA, we used qualitative and quantitative measures to assess participants across two scenarios: an unfamiliar paper selected by the research team according to a standardized procedure, and a paper they had authored, with the latter assessment tailored to their self-specified contributions. Our findings show that participants performed significantly better on their own work than on the unfamiliar control papers (with an AUC of 0.90). Our analysis showed no statistically significant overall difference in performance on unfamiliar papers inside or outside a participant's field. This suggests that greCAPTCHA successfully isolates manuscript-specific understanding from general domain familiarity. More specifically, these results demonstrate that our prototype can differentiate between participants who have the capacity to verify and those who do not, achieving a lowest false-positive rate of 20%. We believe that this error rate should be further reduced, and we discuss several directions for doing so. Participants generally viewed greCAPTCHA as a plausible and robust method for assessing manuscript understanding and therefore as evidence of research authorship.

Participants indicated that deploying a well-designed greCAPTCHA would increase their trust in peer review as a verification mechanism. They also identified free-response questions on rationales and failure modes as particularly effective in evaluating deep understanding of their authored manuscript. Conversely, participants noted that ambiguous question wording, unclear grading expectations, overly rigid rubrics, time constraints, and questions extending beyond an author's specified contribution could result in misleadingly low scores. This feedback points to a clear need for rigorous greCAPTCHA design standards to ensure the tool remains both fair and diagnostic.

As an assessment method, greCAPTCHA shifts the focus of authorship attribution from inspecting the final text artifact (the model of tools like Pangram) to validating the intellectual process behind it. By anchoring verification to the author's capacity to understand and critique their manuscript, our proposed method motivates authentic engagement with the research. Translating this approach to real-world environments requires clear governance in which proposed questions to authors must clearly match declared contributions, evaluation rubrics must remain transparent and adequate, and human oversight must govern consequential judgments. Although our study established the efficacy of greCAPTCHA within a controlled, in-person testing environment, the method holds broad promise for academic publishing and university pedagogy alike. We suggest that conferences, journals, educators, and funders begin prototyping localized implementations. From the lessons learned, greCAPTCHA could evolve into a standardized mechanism that accommodates

responsible AI tools while ensuring authors can fully stand behind the claims they publish and receive appropriate credit.

Acknowledgments

We thank our anonymous study participants, as well as Vishisht Rao for piloting the study. B.G.'s work was funded in part by the Institute for Complex Social Dynamics at Carnegie Mellon University. A.K.'s work was funded in part by the AI2050 program at Schmidt Sciences (grant 24-66924). N.S., J.P., and B.G. were funded in part by grants NSF 1942124 and ONR N000142512346. This work was also supported in part by the Alfred P. Sloan Foundation under Grant No. G-2026-79567. We thank the Gemini Academic Program Award for credits to use Gemini.

References

[ACL25] ACL Rolling Review. *ARR Reviewer Guidelines*. 2025. URL: https://aclrollingreview.org/reviewerguidelines (visited on 09/07/2026).

[AIP26] AIP Publishing. *AI Policy*. 2026. URL: https://publishing.aip.org/resources/researchers/policies-and-ethics/ai-policy/ (visited on 09/07/2026).

[Ame24] American Chemical Society Publications. *Artificial Intelligence (AI) Best Practices and Policies at ACS Publications*. Updated December 13, 2024. 2024. URL: https://researcher-resources.acs.org/publish/aipolicy (visited on 09/07/2026).

[arX26] arXiv. *Monthly Submissions*. Aug. 2026. URL: https://arxiv.org/stats/monthly_submissions.

[Asa+26] Akari Asai et al. "Synthesizing Scientific Literature with Retrieval-Augmented Language Models". In: *Nature* 650.8103 (Feb. 2026), pp. 857–863. ISSN: 1476-4687. DOI: 10.1038/s41586-025-10072-4.

[Ass26] Association for Computing Machinery. *ACM Policy on Authorship*. 2026. URL: https://www.acm.org/publications/policies/new-acm-policy-on-authorship.

[Aug+23] Tal August et al. "Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing". In: *ACM Transactions on Computer-Human Interaction* 30.5 (2023), pp. 1–38.

[Bai+23] Max Bain et al. "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *INTERSPEECH 2023* (2023).

[Bin+18] Reuben Binns et al. ""It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–14. DOI: 10.1145/3173574.3173951. URL: https://doi.org/10.1145/3173574.3173951.

[BMJ26] BMJ Author Hub. *AI Use*. 2026. URL: https://authors.bmj.com/policies/ai-use/ (visited on 09/07/2026).

[BOR23] Jordan Boyd-Graber, Naoaki Okazaki, and Anna Rogers. *ACL 2023 Policy on AI Writing Assistance*. 2023. URL: https://2023.aclweb.org/blog/ACL-2023-policy/ (visited on 09/07/2026).

[Chh+26] Arnavi Chheda-Kothary et al. *Querying Multimodal Scientific Papers with AI: Practices and Preferences Across Blind, Low-Vision, and Sighted Scientists*. July 2026. DOI: 10.1145/3797867.3829024. arXiv: 2607.18514 [cs.HC].

[CSK26] Igor Chirikov, Ivan Smirnov, and René F. Kizilcec. "Generative AI Use and Misuse Call for Assessment Reform in Higher Education". In: *Science* 392.6800 (May 2026), pp. 818–820. DOI: 10.1126/science.aec5115.

[Com23] Committee on Publication Ethics. *Authorship and AI Tools*. 2023. URL: https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools (visited on 09/07/2026).

[Com26] Communication Chairs 2026. *AI-Generated Papers in the NeurIPS 2026 Position Paper Track – NeurIPS Blog*. June 2026. URL: https://blog.neurips.cc/2026/06/02/ai-generated-papers-in-the-neurips-2026-position-paper-track/.

[Con25] Conference on Neural Information Processing Systems. *NeurIPS 2025 Policy on the Use of Large Language Models*. 2025. URL: https://neurips.cc/Conferences/2025/LLM (visited on 09/07/2026).

[Cur+25] C. Curtain et al. *QualCoder 3.8.2*. Version 3.8.2. Computer software. 2025. URL: https://github.com/ccbogel/QualCoder/releases/tag/3.8.2.

[Daw20] Phillip Dawson. *Defending Assessment Security in a Digital World: Preventing E-Cheating and Supporting Academic Integrity in Higher Education*. London: Routledge, Oct. 2020. ISBN: 978-0-429-32417-8. DOI: 10.4324/9780429324178.

[Daw+24] Phillip Dawson et al. "Validity Matters More than Cheating". In: *Assessment & Evaluation in Higher Education* 49.7 (Oct. 2024), pp. 1005–1016. ISSN: 0260-2938. DOI: 10.1080/02602938.2024.2386662.

[Die26] Thomas G. Dietterich. *Attention @arxiv Authors: Our Code of Conduct*. May 2026. URL: https://x.com/tdietterich/status/2055000956144935055.

[DSC17] Xinya Du, Junru Shao, and Claire Cardie. "Learning to Ask: Neural Question Generation for Reading Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1342–1352. DOI: 10.18653/v1/P17-1123. URL: https://aclanthology.org/P17-1123/.

[Duo26] Duolingo. *How to set up your space for the Duolingo English Test*. 2026. URL: https://blog.englishtest.duolingo.com/duolingo-english-test-setup-secondary-camera-room-scan/.

[ESB26] Mohsen Ebrahimzadeh, Antonette Shibani, and Simon Buckingham Shum. "Coauthorship integrity: Reconceptualising assessment validity for the age of generative artificial intelligence". In: *Computers and Education: Artificial Intelligence* 10 (2026), p. 100609. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2026.100609. URL: https://www.sciencedirect.com/science/article/pii/S2666920X26000718.

[Els26] Elsevier. *Generative AI Policies for Journals*. 2026. URL: https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals (visited on 09/07/2026).

[ETS26] ETS. *TOEFL iBT Home Edition*. 2026. URL: https://www.ets.org/toefl/test-takers/ibt/about/testing-options.html.

[FPB26] Annette Flanagin, Roy H. Perlis, and Kirsten Bibbins-Domingo. "Updated Guidance for Author Use of AI in Medical Publication". In: *JAMA* (2026). Published online August 10, 2026. DOI: 10.1001/jama.2026.16613. URL: https://doi.org/10.1001/jama.2026.16613.

[Gao+24] Jie Gao et al. "CollabCoder: A Lower-Barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–29.

[Gar+26] Claudine Gartenberg et al. "More versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review". In: *Organization Science* 37.3 (2026), pp. 795–812.

[HS10] Michael Heilman and Noah A. Smith. "Good Question! Statistical Ranking for Question Generation". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, June 2010, pp. 609–617. URL: https://aclanthology.org/N10-1086/.

[HZH25] Hengzhi Hu, Qing Zhou, and Harwati Hashim. "Negotiating Identity in the Age of ChatGPT: Non-Native English Researchers' Experiences with AI-assisted Academic Writing". In: *Humanities and Social Sciences Communications* 12.1 (July 2025), p. 965. ISSN: 2662-9992. DOI: 10.1057/s41599-025-05351-4.

[Hut+26] Adryana Hutchinson et al. "Re-Examining the Examiners: Changes in Privacy and Security Perceptions of Exam Proctoring". In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. CHI '26. New York, NY, USA: Association for Computing Machinery, 2026. DOI: 10.1145/3772318.3790794. URL: https://doi.org/10.1145/3772318.3790794.

[ICL26] ICLR 2026 Program Chairs. *A Retrospective on the ICLR 2026 Review Process – ICLR Blog*. Mar. 2026.

[IEE24] IEEE Open. *Author Guidelines for Artificial Intelligence (AI)-Generated Text*. Apr. 2024. URL: https://open.ieee.org/author-guidelines-for-artificial-intelligence-ai-generated-text/ (visited on 09/07/2026).

[IEE25a] IEEE/CVF Conference on Computer Vision and Pattern Recognition. *CVPR 2025 Author Guidelines*. 2025. URL: https://cvpr.thecvf.com/Conferences/2025/AuthorGuidelines (visited on 09/07/2026).

[IEE26] IEEE/CVF Conference on Computer Vision and Pattern Recognition. *CVPR 2026 Reviewer Guidelines*. 2026. URL: https://cvpr.thecvf.com/Conferences/2026/ReviewerGuidelines (visited on 09/07/2026).

[IEE25b] IEEE/CVF International Conference on Computer Vision. *ICCV 2025 Reviewer Guidelines*. 2025. URL: https://iccv.thecvf.com/Conferences/2025/ReviewerGuidelines (visited on 09/07/2026).

[Int26a] International Committee of Medical Journal Editors. *ICMJE | Recommendations | Defining the Role of Authors and Contributors*. 2026. URL: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html#two.

[Int26b] International Committee of Medical Journal Editors. *Use of AI by Authors*. 2026. URL: https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html (visited on 09/07/2026).

[Int25] International Conference on Learning Representations. *ICLR 2025 Reviewer Guide*. 2025. URL: https://iclr.cc/Conferences/2025/ReviewerGuide (visited on 09/07/2026).

[Int23] International Conference on Machine Learning. *Clarification on Large Language Model Policy 2023*. 2023. URL: https://icml.cc/Conferences/2023/llm-policy (visited on 09/07/2026).

[Int26c] International Conference on Machine Learning. *ICML 2026 Policy for LLM Use in Reviewing*. 2026. URL: https://icml.cc/Conferences/2026/LLM-Policy (visited on 09/07/2026).

[IOP26a] IOP Publishing Author Support. *Generative AI Tools*. Author policy. 2026. URL: https://publishingsupport.iopscience.iop.org/questions/generative-ai-tools/ (visited on 09/07/2026).

[IOP26b] IOP Publishing Reviewer Support. *Generative AI (Including ChatGPT)*. Reviewer policy. 2026. URL: https://publishingsupport.iopscience.iop.org/questions/generative-ai-for-reviewers/ (visited on 09/07/2026).

[Kar+24] Naveena Karusala et al. "Understanding Contestability on the Margins: Implications for the Design of Algorithmic Decision-Making in Public Services". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. DOI: 10.1145/3613904.3641898. URL: https://doi.org/10.1145/3613904.3641898.

[Kra02] David R. Krathwohl. "A Revision of Bloom's Taxonomy: An Overview". In: *Theory Into Practice* 41.4 (Nov. 2002), pp. 212–218. ISSN: 0040-5841. DOI: 10.1207/s15430421tip4104_2.

[Kub23] Robert Kubinec. "Ordered Beta Regression: A Parsimonious, Well-Fitting Model for Continuous Data with Lower and Upper Bounds". In: *Political Analysis* 31.4 (Oct. 2023), pp. 519–536. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2022.20.

[Kur+20] Ghader Kurdi et al. "A Systematic Review of Automatic Question Generation for Educational Purposes". In: *International Journal of Artificial Intelligence in Education* 30 (2020), pp. 121–204. DOI: 10.1007/s40593-019-00186-y. URL: https://doi.org/10.1007/s40593-019-00186-y.

[Lia+23a] Weixin Liang et al. *Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis*. Oct. 2023. DOI: 10.48550/arXiv.2310.01783. arXiv: 2310.01783 [cs.LG].

[Lia+23b] Weixin Liang et al. "GPT Detectors Are Biased against Non-Native English Writers". In: *Patterns* 4.7 (July 2023). ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100779.

[Liu+25] Jialin Liu et al. *AI-Assisted Writing Is Growing Fastest Among Non-English-Speaking and Less Established Scientists*. Nov. 2025. DOI: 10.48550/arXiv.2511.15872. arXiv: 2511.15872 [cs.DL].

[MH03] Ruslan Mitkov and Le An Ha. "Computer-Aided Generation of Multiple-Choice Tests". In: *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*. 2003, pp. 17–22. URL: https://aclanthology.org/W03-0203/.

[Nat26] Nature Portfolio. *Artificial Intelligence (AI): A Risk-Assessment Framework for Responsible AI Use in Research Publishing*. 2026. URL: https://www.nature.com/nature-portfolio/editorial-policies/ai (visited on 09/07/2026).

[NM26] Batuhan Nursal and Cassie S. Mitchell. "Understanding LLMs' Summarization Capabilities: An Analysis of Biomedical Abstract and Lay Summary Generation". In: *Findings of the Association for Computational Linguistics: ACL 2026*. Ed. by Maria Liakata et al. San Diego, California, United States: Association for Computational Linguistics, July 2026, pp. 11393–11417. ISBN: 979-8-89176-395-1. DOI: 10.18653/v1/2026.findings-acl.554.

[Par+25] Angelina Parfenova et al. "Text Annotation via Inductive Coding: Comparing Human Experts to LLMs in Qualitative Data Analysis". In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by Luis Chiruzzo, Alan Ritter, and Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 6471–6484. ISBN: 979-8-89176-195-7. DOI: 10.18653/v1/2025.findings-naacl.361.

[Pub26] Public Library of Science. *Ethical Publishing Practice*. Portfolio-wide policy. 2026. URL: https://journals.plos.org/plosone/s/ethical-publishing-practice (visited on 09/07/2026).

[Qia+25] Shan Qiao et al. "Generative AI for Thematic Analysis in a Maternal Health Study: Coding Semistructured Interviews Using Large Language Models". In: *Applied Psychology: Health and Well-Being* 17.3 (2025), e70038. ISSN: 1758-0854. DOI: 10.1111/aphw.70038.

[Rad+26] Marissa Radensky et al. *Human-LLM Compound System for Scientific Ideation through Facet Recombination and Novelty Evaluation*. May 2026. DOI: 10.48550/arXiv.2409.14634. arXiv: 2409.14634 [cs.HC].

[Rao+25] Vishisht Srihari Rao et al. "Detecting LLM-generated Peer Reviews". In: *PLOS ONE* 20.9 (Sept. 2025), e0331871. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0331871.

[Rat+25] Marc Ratkovic et al. *Harnessing GPT for Enhanced Academic Writing: Evidence from a Field Experiment with Early-Career Researchers in the Social Sciences*. SSRN Scholarly Paper. Rochester, NY, June 2025. DOI: 10.2139/ssrn.5313034. Social Science Research Network: 5313034.

[Sag26] Sage. *Artificial Intelligence Policy*. 2026. URL: https://www.sagepub.com/journals/publication-ethics-policies/artificial-intelligence-policy (visited on 09/07/2026).

[SSM80] Shayle R Searle, F Michael Speed, and George A Milliken. "Population marginal means in the linear model: an alternative to least squares means". In: *The American Statistician* 34.4 (1980), pp. 216–221.

[Sha+18] Nihar B. Shah et al. "Design and Analysis of the NIPS 2016 Review Process". In: *Journal of Machine Learning Research* 19.49 (2018), pp. 1–34. URL: http://jmlr.org/papers/v19/17-511.html.

[Ska+24] Michael D. Skarlinski et al. *Language Agents Achieve Superhuman Synthesis of Scientific Knowledge*. Sept. 2024. DOI: 10.48550/arXiv.2409.13740. arXiv: 2409.13740 [cs.CL].

[Tai+18] Joanna Tai et al. "Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work". In: *Higher Education* 76.3 (Sept. 2018), pp. 467–481. ISSN: 1573-174X. DOI: 10.1007/s10734-017-0220-3.

[Tay26] Taylor & Francis. *AI Policy*. 2026. URL: https://taylorandfrancis.com/our-policies/ai-policy/ (visited on 09/07/2026).

[Tho23] H. Holden Thorp. "ChatGPT Is Fun, but Not an Author". In: *Science* 379.6630 (2023), pp. 313–313. DOI: 10.1126/science.adg7879. URL: https://doi.org/10.1126/science.adg7879.

[Tra26] Transactions on Machine Learning Research. *Annual Author Submission Quotas for TMLR*. July 2026. URL: https://medium.com/@TmlrOrg/annual-author-submission-quotas-for-tmlr-1db785e51548.

[USE26] USENIX Association. *Policy on the Use of Generative AI*. VehicleSec 2026. 2026. URL: https://www.usenix.org/conference/vehiclesec26/ai-policy (visited on 09/07/2026).

[Wan+26] Binglu Wang et al. *Human-AI Collaboration in Science at Scale: A Global Large-scale Randomized Field Experiment*. May 2026. DOI: 10.48550/arXiv.2605.24180. arXiv: 2605.24180 [physics.soc-ph].

A Question Generation Prompts

This section contains Listings 1 to 4 which each contain the verbatim prompts used with greCAPTCHA to generate the questions sets and grading rubrics of our experiment.

Listing 1: The greCAPTCHA prompt used for generating **planted-error detection** questions.

```
Generate planted-error detection items. Each item states a specific claim about this manuscript, and the participant must identify which version of the claim is what the paper actually reports.

Build each item from a single atomic, verifiable fact that appears exactly once in the paper: a numeric result, an ablation delta, a dataset or baseline name, a hyperparameter, a section or table attribution. Anchor the item to where that fact lives, such as "the ablation in Section 5.2" or "Table 3", so the key is checkable against a specific passage. Never include Section or Table names in the question text to avoid easily looking up the information.

Perturb a value or a referent, never a qualitative direction. Reassigning which component an effect belongs to, or which condition a number describes, is the target. Flipping "improves" to "degrades" is too easy and must not be used. Every incorrect option must be plausible enough that a reader who does not know the work has to locate and read the relevant passage to rule it out.

Never build an item on a fact the paper restates elsewhere, including in the abstract, a figure caption, or the conclusion. A restatement gives a non-author a cheap second place to check. Never build an item whose correct option can be identified by keyword overlap with the prompt, by grammar, by option length, or by being the most specific or most hedged option. Options must be mutually exclusive and comparable in length, specificity, and technical register.

Discard any candidate answerable from the title and abstract alone, and any candidate that someone who knows this field but has never read this paper would get right.
```

Listing 2: The greCAPTCHA prompt used for generating **unstated rationale** questions.

```
Generate unstated-rationale items. Each item asks the participant to justify a methodological or design choice whose reason is not stated in the manuscript.
```

```

Locate a choice the paper makes but does not explain: a test or estimator selected over an obvious alternative, a threshold or cut-off, an excluded condition, an ordering of stages, a control that is present or conspicuously absent. Ask why that choice is appropriate here and, where it sharpens the item, why it would not be appropriate for a specific other part of the paper.

The reason must be genuinely absent from the manuscript, so that no amount of searching the PDF produces it. If the paper explains the choice anywhere, including a footnote, an appendix, or a limitations paragraph, discard the item.

Ground every item in a specific named element of this paper: this comparison, this table, this preprocessing step. A question that could be asked of any paper in the field is testing field knowledge rather than authorship of this artifact, and must be discarded.

Build the rubric to award credit for reasoning that connects the choice to the specific properties of this study's data or design, and to withhold credit for a generically correct textbook justification that never touches this paper. Accept substantively equivalent reasoning and multiple valid explanations. Award no credit for fluency, length, confidence, or command of English; score only the content of the reasoning.
```

Listing 3: The greCAPTCHA prompt used for generating **background knowledge** questions.

```
Generate background-concept items. Each item asks the participant to define or explain, in their own words, one concept the paper depends on but does not itself define - the presupposed background a competent member of this field carries into reading it, not the paper's own contribution.

Choose concepts that are load-bearing. If the participant misunderstood the concept, the paper's design, its choice of method, or its interpretation of its results would no longer make sense. A term mentioned once in passing is not load-bearing; the standard is that you could name a specific design decision or claim that depends on it.

The most important filter is that the paper must not define the concept. The participant answers with the paper in front of them, so a concept the paper explains in its own words is a lookup rather than a question. Prefer concepts the paper names and uses as though they need no introduction. Where the paper's own topic is a concept it does define, choose an adjacent concept it presupposes instead: for a paper about p-hacking that defines p-hacking, ask what a p-value means under the null hypothesis, what the familywise error rate is, or what pre-registration is for.

Never name a section, table, figure, or page in the question text, and never reuse the paper's own phrasing of the concept. Both point the participant at a passage to copy.

Ask for two things in each item, in this order: the definition or explanation itself, and one sentence on why the concept matters for the work reported here. The first half is what the item is for. The second half is what someone who does not understand the work cannot supply, and it is where an answer assembled from general knowledge alone reads as generic.

Build the rubric as an enumeration of the elements a correct definition must contain, each its own criterion with its own points, totalling 70, plus one criterion worth 30 for the connection to this paper. Name the required elements explicitly in the guidance, so grading is a check against a list rather than a judgement of quality. For at least one criterion, state a specific plausible-but-wrong answer that earns nothing - the near-miss someone with topic-adjacent familiarity would give - so an answer that circles the concept without stating it cannot collect points.

Award nothing for fluency, length, hedging, confidence, or restating the question. Accept any wording that carries the required elements, including informal phrasing, an example that entails the definition, and notation in place of prose. Do not require the participant's terminology to match the paper's.

Discard any candidate concept whose definition appears anywhere in the paper, any answerable from the title alone, and any that duplicates a concept an earlier card already covers.
```

Listing 4: The greCAPTCHA prompt used for generating **failure mode** questions.

```
Generate one item asking the participant to name a realistic condition under which this work's method or central finding would degrade, and to say why it would.

Keep the question to one or two sentences, and ask for a short answer of two or three sentences. Word it neutrally: a condition the work was not built for, not a flaw in it. A scored item that reads as an attack on the participant's own paper invites a defensive answer rather than an informative one.

The condition must be specific to this setup and plausible in this domain - a property of the data, a regime, a scale, a population, or an interaction this pipeline would mishandle. Its mechanism must follow from how this work is built, not from general methodological caution.

Do not use anything the paper names itself: its limitations, its future work, its statements about what it did not test, or anything in the abstract. The participant reads with the paper open, so those are lookups. Never name a section, table, or figure in the question text.

In the guidance, list the qualifying conditions for this paper, each with the mechanism that makes it fail. Any one of them earns the naming points, since reasonable people will pick different edges.
```

```
10
11 Then list the answers that earn nothing: that the sample is small, that the results may not generalise, that more data
   ↪ or more baselines are needed, that the method is untested in other settings, and anything else that could be written
   ↪ without having read this paper. A fluent, confident answer of exactly that kind is the most likely wrong answer here,
   ↪  and it must score zero.
12
13 Weight the mechanism above the condition. Award nothing for fluency, length, hedging, or for restating what the paper
   ↪ already says about its own scope.
```

B Paper Selection Prompt Template

The prompt used to suggest five potential candidates for the unfamiliar paper of the participant, shown in Listing 5. This prompt generates a full report with justification of why each of the five candidate papers are appropriate.

Listing 5: The prompt used to select five candidate unfamiliar papers for a participant.

```
You are helping select a paper for a controlled study. Attached is a manuscript written by {{PARTICIPANT_NAME}}, one of its authors. I need five candidate papers by other people that this person is unlikely to have read, all of them {{IN-FIELD/OUT-OF-FIELD}} for them.

Definitions, which govern everything below.

IN-FIELD means their training transfers: the notation, methods and background concepts are ones they already hold, and they could read the paper without learning a new apparatus. Same research area, different research group.

OUT-OF-FIELD means their training does not transfer: the background concepts are ones they have no reason to know. It does not mean unreadable. The paper must still be a normal empirical or theoretical paper that an intelligent outsider can follow at the level of its claims, because a participant who cannot parse the abstract at all tells us nothing.

STEP 1 - Identify the person, and say how.

Use the attached manuscript as ground truth: its author list, affiliation, topic and references are the anchor. Then search for this person's publication record.

Report what you found as a short profile: their research area, subfield, the methods they work with, their affiliation and career stage, their most cited work, and the venues they publish in. Cite a source for each claim - their scholar profile, lab page, arXiv listing, or a specific paper.

Name disambiguation is the failure mode that ruins everything downstream. If more than one researcher shares this name, say so explicitly, state which one you concluded is the author of the attached manuscript, and give the evidence that settles it. If you cannot settle it confidently, stop and tell me rather than guessing.

STEP 2 - Define their familiarity neighbourhood.

Before choosing anything, list what this person is likely to have already read:
- their own papers and their coauthors' papers;
- work by their advisor, their lab, and their institution's group in this area;
- the papers their attached manuscript cites, and well-known papers that cite it;
- the canonical or widely discussed papers of their subfield, including anything that circulated heavily on social media or won an award;
- the standard venues they attend, whose proceedings they will have browsed.

Everything in that neighbourhood is disqualified. Say what you excluded and why, briefly.

STEP 3 - Select five candidates.

Constraints on every candidate:
- A real paper, posted to arXiv in 2025 or 2026, with a resolvable arXiv ID. Quote the first sentence of its abstract verbatim as evidence you actually retrieved it. Do not offer a paper you cannot verify exists; five verified candidates matter more than five interesting titles.
- Not authored by this person, their coauthors, their advisor, or their institution.
- Outside the familiarity neighbourhood from Step 2.
- Low visibility: few citations, no award, no viral thread. A paper they would have encountered by accident is useless to me.
- Comparable in difficulty to the attached manuscript - similar length, similar density of mathematics, a similar number of experiments or results. If the unfamiliar paper is markedly harder than their own, any score difference I measure is about difficulty rather than about authorship, which destroys the comparison.
- Self-contained enough to be read cold: it states its own research question and what it found, rather than depending on a companion paper.

Make the five diverse, so that one wrong guess about this person does not invalidate the whole set. For in-field, five different research groups. For out-of-field, five different fields.

STEP 4 - Report.

Rank them best fit first, and for each give:
1. Title, authors, arXiv ID and link, posting date, primary arXiv category.
2. The first sentence of the abstract, quoted.
3. One sentence on what the paper is about.
4. Why it is {{IN-FIELD/OUT-OF-FIELD}} for this person specifically, referring to the profile from Step 1 rather than to the field in general.
```

```
5. Why they are unlikely to know it, referring to the neighbourhood from Step 2.
6. Difficulty match to the attached manuscript: length, mathematical density, number of results, and whether you judge it easier, comparable, or harder.
7. The single most likely reason this person might in fact already know it. Every candidate gets one; if you cannot think of one you have not looked hard enough.
8. Your confidence that they have not read it, as high, medium or low.

End with the one candidate you would pick if you had to choose, and the one you consider riskiest, each in a sentence.

Do not pad the list to five with papers you doubt. If you can only verify three that meet the constraints, give me three and say what blocked the others.
```

C Participant Summary

Table 4. Participant characteristics and metadata for participants' own and unfamiliar papers. Preparation time is the amount of time the participant reported reviewing the unfamiliar paper, in minutes.

	Participant background		Own paper		Unfamiliar paper		
ID	**Position**	**Pubs.**	**Field**	**Year**	**Field**	**Pages**	**Prep. time**
1	Master's	0–4	Machine Learning	2026	Astrophysics	1–10	60
2	Postdoc	5–9	Social and Information Networks	2026	Human AI Interaction	11–20	0
3	Master's	0–4	Logic	2026	Geophysics	11–20	0.5
4	PhD	0–4	Robotics	2026	Geophysics	11–20	0
5	PhD	5–9	Logic in Computer Science	2026	Quantitative Biology	11–20	30
6	PhD	0–4	Programming Languages	2026	Chemical Physics	21–30	5
7	PhD	0–4	Computer Science and Game Theory	2024	Artificial Intelligence	11–20	5
8	PhD	5–9	Geophysics	2026	Image and Video Processing	11–20	10
9	PhD	0–4	Human-Computer Interaction	2026	Education Technology	1–10	6
10	Postdoc	5–9	Signal Processing	2024	Computer Vision and Pattern Recognition	11–20	30
11	PhD	5–9	Distributed, Parallel, and Cluster Computing	2026	Theoretical Computer Science	21–30	20
12	Professor	50+	Neurons and Cognition	2015	Medical Physics	21–30	0
13	Master's	0–4	Computation and Language	2026	Computational Linguistics	11–20	10
14	PhD	5–9	Artificial Intelligence	2026	LLMs/Agents	21–30	5
15	Postdoc	20–49	Human-Computer Interaction	2021	Astrophysics	11–20	0
16	Postdoc	5–9	Computation and Language	2025	Computation and Language	21–30	25
17	Postdoc	10–19	Geophysics	2025	Robotics	1–10	30
18	Postdoc	5–9	Neurons and Cognition	2026	Atmospheric and Oceanic Physics	41–50	0
19	PhD	0–4	Robotics	2025	Robotics	1–10	5
20	PhD	0–4	Machine Learning	2026	Fluid Mechanics	21–30	0.5
21	Postdoc	5–9	Human-Computer Interaction	2026	Human-Computer Interaction	11–20	20
22	PhD	5–9	Machine Learning	2026	Clinical AI	21–30	0
23	PhD	5–9	Human-Computer Interaction	2022	Neurons and Cognition	11–20	0
24	PhD	5–9	Cryptography and Security	2025	Populations and Evolution	21–30	20
25	Postdoc	5–9	Neurons and Cognition	2026	(Computational) Linguistics	21–30	10
26	PhD	0–4	Logic in Computer Science	2026	Biological Physics	11–20	5
27	PhD	0–4	Machine Learning	2026	Soft Condensed Matter	11–20	0
28	PhD	10–19	Cryptography and Security	2026	Microfluidics	11–20	5
29	Master's	0–4	Neurons and Cognition	2026	Logic in Computer Science	11–20	0
30	PhD	5–9	Computers and Society	2026	Education Technology	11–20	5
31	PhD	0–4	Geophysics	2024	Geophysics	51–60	3

D Distribution of Scores by Question Family

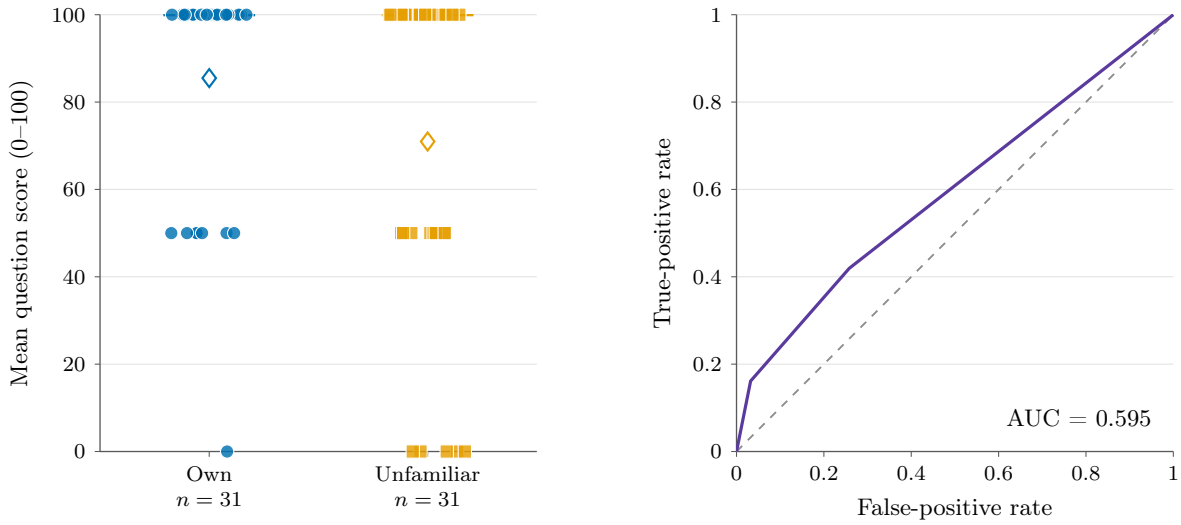

Figure 8. Distribution of the scores for the planted-error question family for participants' own papers and unfamiliar papers, and the ROC curve and AUC when classifying unfamiliar papers using only average scores on **planted error**.

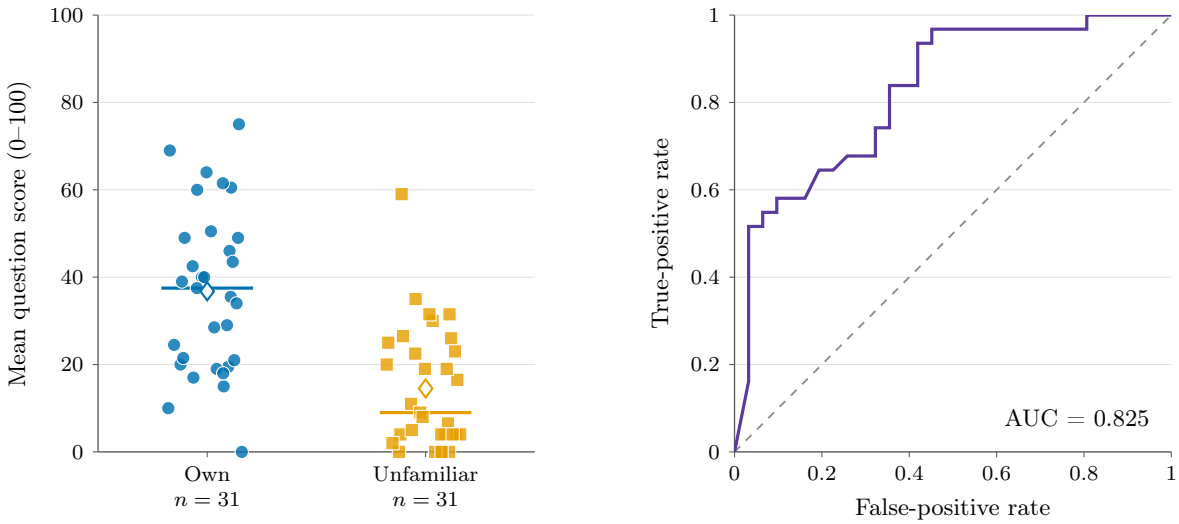

Figure 9. Distribution of the scores for the **unstated-rationale** question family for participants' own papers and unfamiliar papers, and the ROC curve and AUC when using only average scores on unstated rationale.

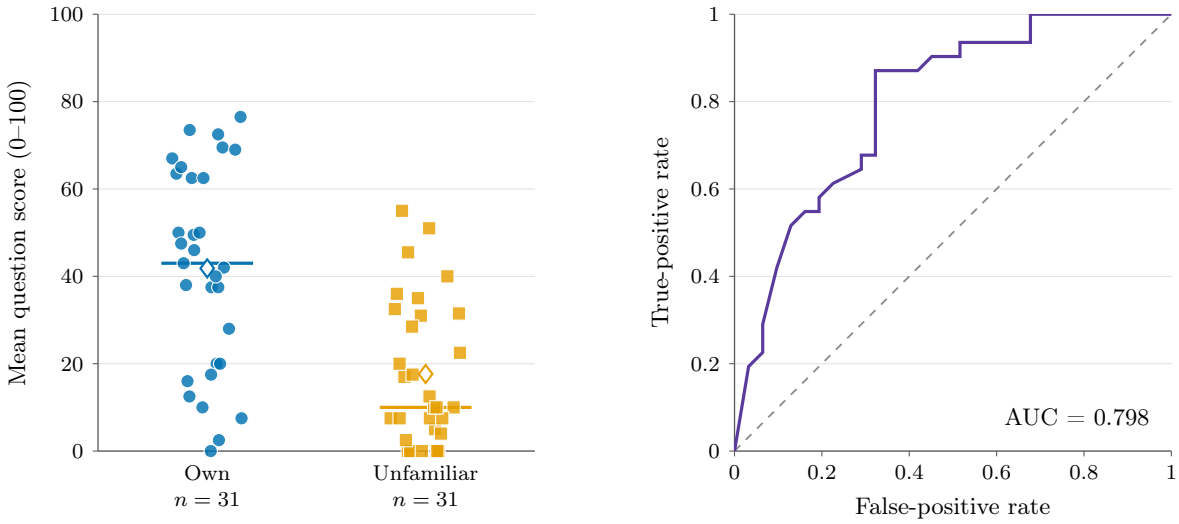

Figure 10. Distribution of the scores for the background-knowledge question family for participants' own papers and unfamiliar papers, and the ROC curve and AUC when classifying unfamiliar papers using only average scores on **background knowledge**.

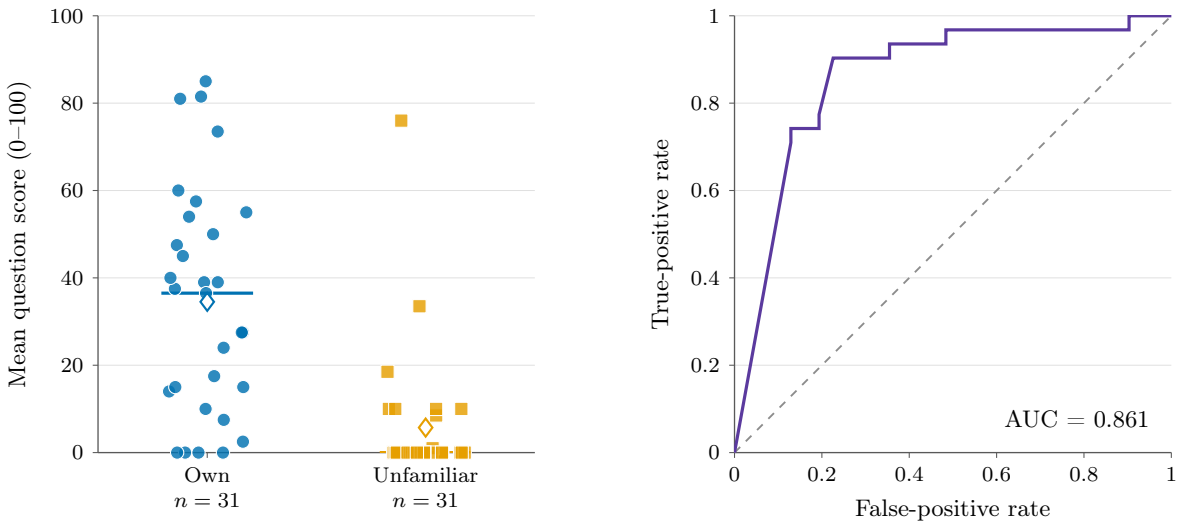

Figure 11. Distribution of the scores for the failure-mode question family for participants' own papers and unfamiliar papers, and the ROC curve and AUC when classifying unfamiliar papers using only average scores on **failure mode**.

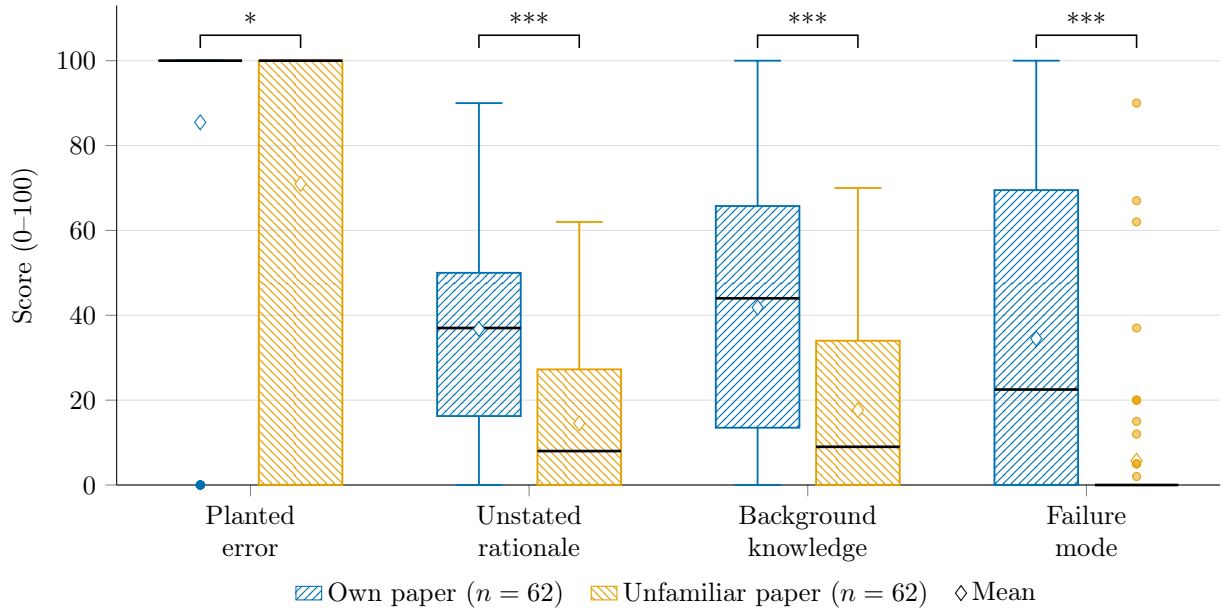

Figure 12. Score distributions by question family and paper type ($n = 62$ per box). Diamonds indicate means; brackets mark significant differences using Holm-adjusted p-values (* $p < .05$, *** $p < .001$).

E Additional Statistical Tests

For further analysis we define H_1 to be the null hypothesis that, within each question family, participants cannot significantly be separated based on whether they were assessed on their own paper vs. on an unfamiliar paper. To test this hypothesis, we use the model: `score` $\sim$ `paper_type * question_family + (1 | participant_id)`. We fit it on all 496 question responses, from 2 tests with 8 questions each over 31 participants. Questions which were unanswered due to skipping or test time-out were marked as having a score of 0. Details of this data are shown in Table 5.

We normalize scores to the range $[0, 1]$ by dividing by 100. Scores are bounded and clustered around 0 and 1, especially for the planted-error question family. We therefore fit an ordered-beta regression model to our data [Kub23]. For each question family q, the maximum likelihood ordered-beta regression model fits logit-link values $\beta_{\text{own},q}$ and $\beta_{\text{unfamiliar},q}$ indicating the connections between the question family and score for each paper type.

For each question family q, we construct the statistic $C_q \doteq \beta_{\text{own},q} - \beta_{\text{unfamiliar},q}$. Our null hypothesis for each question family q is $H_0 : C_q = 0$. We calculate marginal means using the "emmeans" R package [SSM80], and correct for multiple comparisons using 4-way Holm correction. Tables 6 and 7 and Figure 12 show the full statistical results of the test.

Our statistical analysis rejects the null hypothesis for all four question families under Holm correction. We find that p-values are significant at the 0.05 level for planted error, and at the 0.001 level for all free-response question families. This is strong evidence to indicate that the four question families are predictive of the capacity to verify one's paper.

We also create an alternative model in which we exclude all skipped and timed-out questions from our analyzed data. The results for this are reported in Tables 8 and 9 and Figure 13. Under this model, all free-response question families show significant contrasts between scores in the own and the unfamiliar settings ($p < 0.001$; Holm-adjusted). Nevertheless, the effect fails to meet the threshold for significance on planted error (multiple choice) questions. For this question family, this additional result suggests a different driver behind the previously witnessed effect: participants seem more likely to skip or not reach the planted error questions in the unfamiliar-paper setting. When they do answer these questions, they are not significantly

less accurate than in the own-paper setting.

We also present additional results for H_2, the null hypothesis that there are no significant differences between in-field and out-of-field unfamiliar paper scores. These results are described in Tables 10 to 12.

Table 5. Descriptive metrics of question-scores overall, grouped by paper ownership and by question family. Scores range from 0 to 100. For every group, the total number of observations equals n, and no scores are missing. Zero, Full, and Nonresponse denote the percentages of observations with a score of 0, a score of 100, and no response, respectively.

		Score			Rate (%)		
Group	n	Mean	SD	Median	Zero	Full	Nonresponse
Overall	496	38.43	39.85	25.0	34.1	21.0	14.3
By paper ownership							
Unfamiliar paper	248	27.21	38.05	5.0	46.8	17.7	22.6
Own paper	248	49.65	38.50	50.0	21.4	24.2	6.0
By question family							
Planted error	124	78.23	41.44	100.0	21.8	78.2	10.5
Unstated rationale	124	25.64	24.36	20.0	21.0	0.0	12.9
Background knowledge	124	29.73	29.69	20.0	33.1	2.4	15.3
Failure mode	124	20.12	31.97	0.0	60.5	3.2	18.5
Unfamiliar paper, by question family							
Planted error	62	70.97	45.76	100.0	29.0	71.0	17.7
Unstated rationale	62	14.52	17.76	8.0	32.3	0.0	19.4
Background knowledge	62	17.65	22.23	9.0	45.2	0.0	21.0
Failure mode	62	5.73	16.89	0.0	80.6	0.0	32.3
Own paper, by question family							
Planted error	62	85.48	35.51	100.0	14.5	85.5	3.2
Unstated rationale	62	36.76	25.10	37.0	9.7	0.0	6.5
Background knowledge	62	41.82	31.39	44.0	21.0	4.8	9.7
Failure mode	62	34.52	36.80	22.5	40.3	6.5	4.8

Table 6. Estimated marginal mean scores for H_1 by question family and paper ownership. Scores are reported on a 0–100 scale, with 95% confidence intervals (CIs).

	Unfamiliar paper		Own paper	
Question family	Mean	95% CI	Mean	95% CI
Planted error	91.33	[83.99, 95.49]	97.19	[93.89, 98.74]
Unstated rationale	26.81	[21.82, 32.46]	42.79	[37.13, 48.64]
Background knowledge	31.36	[25.44, 37.96]	49.30	[43.03, 55.60]
Failure mode	14.43	[9.46, 21.39]	47.01	[39.95, 54.19]

Table 7. Contrasts calculated for H_1 between own vs. unfamiliar paper conditions on the model's link scale. Positive estimates indicate higher scores for own papers. We apply Holm adjustment across the four contrasts. Bold estimates and stars indicate significance using the Holm-adjusted p-values: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

			p-value	
Question family	Estimate	SE	Raw	Holm-adjusted
Planted error	**1.190***	0.475	0.0122	0.0122
Unstated rationale	**0.714*****	0.165	< 0.0001	< 0.0001
Background knowledge	**0.755*****	0.181	< 0.0001	< 0.0001
Failure mode	**1.660*****	0.267	< 0.0001	< 0.0001

Table 8. Estimated marginal mean scores for H_1 using questions that were answered only, grouped by question family and paper ownership. Scores are reported on a 0–100 scale, with 95% confidence intervals (CIs).

	Unfamiliar paper		Own paper	
Question family	Mean	95% CI	Mean	95% CI
Planted error	98.08	[95.26, 99.24]	98.44	[96.17, 99.38]
Unstated rationale	27.24	[22.13, 33.04]	42.34	[36.62, 48.28]
Background knowledge	32.45	[26.27, 39.31]	49.80	[43.38, 56.22]
Failure mode	15.79	[10.17, 23.70]	46.18	[38.98, 53.55]

Table 9. Contrasts calculated for H_1 between own vs. unfamiliar paper conditions on the model's link scale but using answered questions only. Skipped, timed-out, and unknown-status questions are excluded. Positive estimates indicate higher scores for own papers. Holm adjustment is applied across the four contrasts. Bold estimates and stars indicate significance using Holm-adjusted p-values: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

			p-value	
Question family	Estimate	SE	Raw	Holm-adjusted
Planted error	0.210	0.580	0.7171	0.7171
Unstated rationale	**0.674*****	0.170	< 0.0001	0.0002
Background knowledge	**0.725*****	0.189	0.0001	0.0002
Failure mode	**1.521*****	0.283	< 0.0001	< 0.0001

Table 10. Estimated marginal mean scores for H_2 on unfamiliar papers by whether their field matches that of the participant. Scores are reported on a 0–100 scale, with 95% confidence intervals (CIs).

Field match	Mean	95% CI
Out-of-field	36.77	[28.22, 46.23]
In-field	40.55	[31.93, 49.79]

Table 11. H_2 in-field vs. out-field contrast on the model's link scale for unfamiliar papers. A positive estimate indicates higher scores for in-field papers.

Estimate	SE	p-value
0.160	0.205	0.4375

Table 12. Participant-level robustness tests comparing unfamiliar-paper scores by whether they match the field of the participant. Confidence intervals (CIs) are reported at the 95% level.

Test	Statistic	p-value	95% CI
Welch two-sample t-test	-0.67	0.5084	[−12.808, 6.491]
Wilcoxon rank-sum test	100.00	0.4407	[−12.750, 6.125]

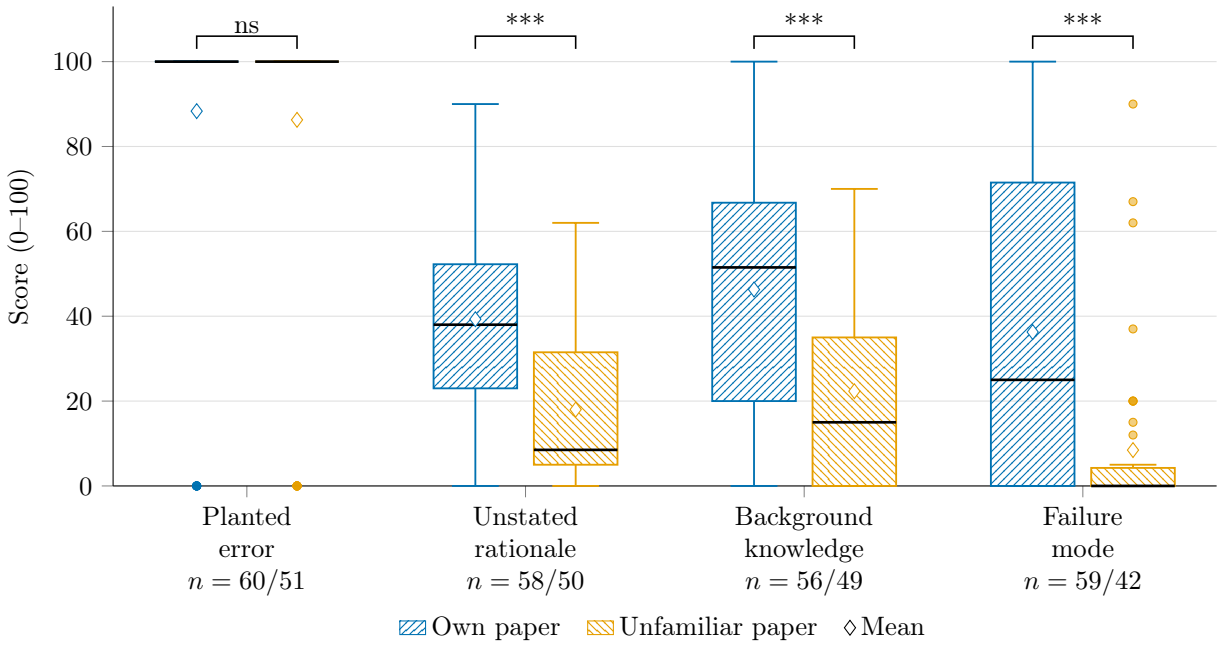

Figure 13. Score distributions for answered questions only, grouped by question family and paper type. Sample sizes show the counts of responses to question sets based on own/unfamiliar paper. Diamonds indicate means. Brackets report Holm-adjusted comparisons (*** $p < .001$; ns, not significant).

F Deductive Codebook

Table 13 shows our deductive codebook with occurrence counts after coding.

Table 13. Our final deductive codebook with columns for the **Code** name, its **Description**, the different **Type and values** into which we have subdivided each code, and the total number of occurrence **Count** across all transcripts.

Code	Description	Type and values	Count
Perceived validity	The participant perceived the system as providing valid evidence to the fact that someone has understood the paper.	*Category:* Disagree, Neither agree nor disagree, Agree	91
Question clarity	Whether the wording and requested response were understandable.	*Category:* Unclear; Partly clear; Clear	79
Question relevance	The participant found the question relevant.	*Category:* Disagree, Neither agree nor disagree, Agree	71
Background familiarity	Did the participant feel their background prepared them to answer questions about the unfamiliar paper?	*Category:* Not at all, a little bit, very much.	70
Question difficulty	The participant's overall perception of question difficulty.	*Category:* Too easy; Appropriate; Too difficult	64
Rubric appropriate	The participant thought the grading rubrics were appropriate.	*Category:* Disagree, Neither agree nor disagree, Agree	61
Deployment	Should a system like this be deployed in conferences or journals?	*Category:* No, Maybe with Significant Changes, Yes	60
Perceived purpose	Did the participant guess the intended purpose of the system?	*Category:* No, Close Guess, Yes	55
Time pressure	Did the participant feel pressured for time?	*Binary:* Present, Absent	55
Overall preference	How did the participant like the overall experience with the system?	*Category:* Disliked, Neither liked nor disliked, Liked	54
Grading agreement	The participant agreed with their grades.	*Category:* Disagree, Neither agree nor disagree, Agree	51
Question depth	Whether the question requires applying, extending, critiquing, or reconstructing ideas beyond recalling the text.	*Category:* Not tested; Weakly tested; Strongly tested	44
Trust change	How would the use of our system change the participant's trust in a submission process?	*Category:* Decrease Trust, Neither increase nor decrease trust, Increase trust	41
Circumvention feasibility	How feasible the participant thinks it would be to prepare strategically without genuinely verifying the paper.	*Category:* Low; Moderate; High	31
Feedback relevance	The participant found the feedback provided by the system relevant.	*Category:* Disagree, Neither agree nor disagree, Agree	28
Contribution coverage	Whether important parts of the participant's contribution were tested.	*Category:* Important omissions; Partial coverage; Strong coverage	27
Grading false positive	The participant reports that a weak or incorrect answer received too much credit.	*Binary:* Present, Absent	7
Grading false negative	The participant reports that a correct or reasonable answer received too little credit.	*Binary:* Present, Absent	6

G Inductive Codebook

Table 14. Inductive codes and counts, grouped by category. Categories are ordered by the sum of their code counts; codes are ordered by count within each category.

Category and description	Code	Count
Clarify expectations *Uncertainty about the expected answer, level of detail, or criteria for an acceptable response.*	System expects particular answer	52
	Unclear how much detail expected	44
	Transparency of expected outcome	13
Negative emotions *Emotional reactions to assessment and feedback.*	Reflection on performance	32
	Frustration	26
	Feels like on trial	14
	Black mirror	11
	Breakdown of trust	2
How to prepare *Strategies for preparing for the assessment through reading, careful note-taking, AI assistance, and critical examination of the paper.*	Prepare by skimming	24
	Prepare with LLM	20
	Prepare by documentation	17
	Inspect references	11
	Find inconsistencies	4
Content limit *Constraints on demonstrating understanding by time pressure and response-length limits.*	(Deductive: Time pressure)	55
	Length limit	13
Change modality or test structure *Proposed changes to questions, rubrics, interaction, and response modalities to improve the assessment during deployment.*	Question changes	20
	Interaction with system	19
	Rubric changes	17
	System modality	6
Deployment role *Proposed roles for greCAPTCHA results in downstream decisions, including who should receive them.*	Desk reject or ban	18
	Threshold choice	17
	Show to reviewers	12
	Just show score to authors	5
	Opt-in choice	3
	Show to metareviewers	2
Deployment concern *Practical and governance concerns about deployment.*	Deployment overhead	21
	Requires human oversight	17
	Need for deployment transparency	7
	Deployment with many projects	5
	Deployment with many collaborators	3
	Seniority effects	1
Unintended measurements *Concerns that scores reflect unintended capabilities such as test-taking skills, which are beyond the participant's control.*	Measures typing ability	17
	Measures test aptitude	10
	Measures speed of thinking	8
	Measures English fluency	6
	Measures searching ability	5
	Measures paper age	3

Continued on next page.

Table 14 continued.

Category and description	Code	Count
Experience confound *Perceived influences of disciplinary background, field familiarity, and specific research contributions on greCAPTCHA performance.*	Background dependence	21
	Field dependence	12
	Contribution dependence	11
Learning *Opportunities to use the assessment for learning, reflection, and identifying gaps in understanding one's own research.*	Use case for learning	25
	Learn about own paper	19
Uncategorized *Additional observations about assessment experiences, alternative approaches, and contextual factors not captured by other categories.*	Baffled by unfamiliar paper	13
	Alternative systems for authorship check	7
	Question order	6
	Extenuating reason for failure	4
	Response time	3
	Communication importance	2
Preparation time *Accounts of the amount of time spent preparing for the unfamiliar-paper assessment.*	Short preparation	17
	Long preparation	9
False positives and false negatives *Perceived errors in awarding or withholding credit, and concerns about incorrectly marking authors as non-authors.*	(Deductive: Grading false positive)	7
	(Deductive: Grading false negative)	6
	System weeds out wrong people	5
Answer source *Resources used or discussed regarding answering questions, including background knowledge, the manuscript, and cheating.*	Use background	9
	Cheating	4
	Use paper	2
Perceived validity: field *Views on whether the assessment provides valid evidence of understanding the manuscript's research field.*	Agree	7
	Disagree	4
	Neither agree nor disagree	0
UI design *Experiences of interface clarity and how it supports or obstructs interaction with the assessment.*	UI unclear	9
	UI clear	2