Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Recipient-Focused Hallucinations in AI-Generated Professor Outreach Emails: A Pilot Factorial Evaluation of Grounding and Verification

Azhdar Mammadov

azhdar.mammadov06@gmail.com

University of Nottingham https://orcid.org/0009-0009-4326-5838

Abhinav Singh

Birla Institute of Technology and Science, Pilani https://orcid.org/0009-0007-0604-0782

Phong Cao

Worcester Polytechnic Institute https://orcid.org/0009-0008-1951-1560

Kanan Gasimov

Independent Researcher https://orcid.org/0009-0002-2140-4154

Research Article

Keywords: Large Language Models, Hallucination, Academic outreach, Grounded generation, Scientific claim verification, Personalized text generation, Factuality evaluation

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11749150/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Recipient-Focused Hallucinations in AI-Generated Professor Outreach Emails: A Pilot Factorial Evaluation of Grounding and Verification

†Azhdar Mammadov[1], Abhinav Singh[2], Kanan Gasimov[3], and Phong Cao[4]

[1] Department of Electrical and Electronic Engineering, University of Nottingham, Nottingham, United Kingdom

[2] Department of Computer Science, BITS Pilani, Pilani, Rajasthan, India

[3] Independent Researcher, Baku, Azerbaijan

[4] Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts, USA

† Corresponding author: Azhdar Mammadov, azhdar.mammadov06@gmail.com

ORCID iDs: A. Mammadov: https://orcid.org/0009-0009-4326-5838 | A. Singh: https://orcid.org/0009-0007-0604-0782 | K. Gasimov: https://orcid.org/0009-0002-2140-4154 | P. Cao: https://orcid.org/0009-0008-1951-1560

ABSTRACT

Background: LLMs increasingly draft personalized academic outreach, which depends on describing a real professor's research. Unsupported claims about a recipient's topics, methods, or findings misrepresent their scholarly work.

Methods: In an offline pilot, a 2 × 2 design varied whether the writer received a publication evidence packet during drafting and whether a verifier revised the draft afterwards (conditions A–D). We generated 96 emails for 12 professors across three fields, decomposed them into 360 atomic research claims, and labelled each against the professor's evidence packet with an automatic (LLM) evaluator. All primary estimates are pilot estimates from automatic labels, not human-verified measurements. Reliability was assessed against a blinded 36-claim human-labeled subset (10% of claims) and a second automatic evaluator from a third model family.

Results: Under the primary evaluator, the closed-book baseline (A) had a Severe Error Rate (SER) of 0.344 (21/61 claims; exact 95% CI 0.227–0.477). Grounding the writer (B) produced no severe errors in 90 judged claims (0.000, CI 0.000–0.040) and raised Supported Personalization Density (SPD) from 0.65 to 1.71 supported claims per 100 words. Verify-only (C) left 1 severe error in 50 claims (0.020; SPD 1.04); grounded-plus-verified (D) produced none (0/87) at the highest specificity (1.76).

Interpretation and limits: Primary-evaluator agreement with the human annotator was only fair (Cohen's $\varkappa = 0.33$, CI 0.16–0.49). The evaluator was too lenient in 14 of 16 disagreements, suggesting that the automatic severe-error rates may underestimate error prevalence. However, the small human-labelled subset does not establish the magnitude or direction of bias across the complete dataset. The zeros are evaluator-specific: a second evaluator relabelling the same claims found 10 severe errors in B and 4 in D, closed-book 0.455 - a large reduction under grounding, not elimination. And 17 of the 21 closed-book severe errors came from the four psychology professors; excluding that field, closed-book SER falls to 0.095. Both automatic evaluators reported lower pooled severe-error rates for grounded conditions than for ungrounded conditions. Grounding during drafting also showed a lower observed pooled SER than post-generation verification alone.

Conclusion: Verified publication evidence sharply reduced severe misrepresentation - to zero under the primary evaluator, roughly four-fold under the second - while making emails more specific, and outperformed correcting drafts afterwards. Grounding, not correction alone, is the more promising safeguard; establishing the effect size needs human labelling at scale and a validated judge.

Keywords: Large Language Models; Hallucination; Academic outreach; Grounded generation; Scientific claim verification; Personalized text generation; Factuality evaluation.

1. Introduction

Large Language Models (LLMs) are increasingly used as writing assistants for personalized text, including email and academic writing. Models can adapt writing style from very limited user data, and personalized generation cannot be judged well by generic lexical metrics alone. [1, 2]

Academic outreach is distinctive because personalization is not only about tone. A message to a professor typically demonstrates fit by referring to the recipient's research topics, publications, methods, datasets, findings, or collaborations. That makes it more sensitive than ordinary email polishing: a message may sound specific and respectful while attributing the wrong method, overstating a result, inventing an impact claim, or combining unrelated papers into a misleading description. Unlike ordinary summarization errors, these mistakes concern identifiable researchers whose published work provides an evidence base against which factual claims can be assessed. However, publication evidence packets are necessarily incomplete, and lack of support from a packet does not establish that a claim is false.

Although previous research has examined personalized text generation, factuality evaluation, and grounded generation separately, recipient-specific factuality in academic outreach remains comparatively underexplored. We examine this setting through claim-level evaluation and measure whether factuality safeguards preserve supported personalization. This study addresses that gap with a controlled 2 × 2 pilot comparing evidence supplied during drafting against evidence-based post-generation verification.

The primary research question is how publication grounding during drafting and evidence-based post-generation verification affect the factual accuracy of research-related claims in AI-generated outreach emails to professors. The secondary objective is to measure the factuality-specificity trade-off - whether safer outputs remain genuinely personalized, or whether safety is achieved by deleting the details that give an email value.

This is a pilot: its contribution is a measurement framework, a reproducible pipeline, and a first effect estimate with explicit uncertainty - not a settled effect size.

2. Related Work & Research Objective

2.1 Personalized text and email generation

Language models can adapt wording, tone, and perceived relevance to a target reader. Panza shows personalized email assistants learn sender-style patterns from small email collections and use retrieval-augmented generation to condition on examples.[1] Organizational bulk-email research shows personalization affects engagement outcomes such as recognition and detailed reading.[3] Both evaluate style, relevance, or engagement rather than whether claims about a real recipient are supported. We instead focus on recipient-grounded personalization: the email must describe a professor's actual research accurately, not merely sound tailored.

2.2 Automated personalized-text evaluation

AuPEL argues personalized generation should be evaluated across personalization, quality, and relevance, not by surface metrics alone.[2] Work on personalization-induced hallucination shows personalization can itself distort factual reasoning, creating tension between being specific and staying correct.[4] Both motivate measuring whether an email's research claims are evidence-supported, not just whether it sounds personalized.

2.3 Atomic factuality evaluation

FActScore evaluates long-form generation by decomposing text into atomic factual claims and measuring the proportion supported by reliable sources.[5] This matters here because one fluent sentence may carry a topic, method, dataset, and impact claim at once; judging the email as simply correct or incorrect would hide those differences. We adapt atomic factuality evaluation to research claims in academic outreach.

2.4 Scientific claim verification

SciFact checks scientific claims against paper abstracts, labelling whether evidence supports, refutes, or fails to establish them.[6] SciFact-Open extends this to an open domain, showing verification hardens as retrieval scale grows and that abstract-only evidence supports only a limited version of some claims.[7] Hence our evidence policy: packets carry publication metadata and abstracts plus selected passages for focal papers where method-, result-, or scope-sensitive claims need more detail. These benchmarks, however, evaluate prepared claims rather than naturally generated emails, where claims arise inside persuasive text that may overstate novelty or infer collaborations.

2.5 Grounded generation and post-generation correction

RARR checks outputs after generation and minimally revises them to improve attribution while preserving original text.[8] FLEEK pipelines claim extraction, evidence gathering, judgement, and correction.[9] SciFix corrects false scientific claims.[10] These motivate our verification arm. On the grounding side, AGREE shows explicit grounding improves generation and citation quality over weaker prompting or post hoc citation,[11] and ScholarCopilot shows retrieval-integrated academic writing improves citation quality.[12] Together, these studies motivate a direct comparison between evidence supplied during drafting and evidence-based revision after generation.

2.6 Remaining gap

These strands study separate parts of the problem. Email-personalization work rarely evaluates factual accuracy about a real recipient; scientific verification rarely evaluates naturally generated outreach; grounded generation does not measure whether safeguards preserve professor-specific personalization. The gap is recipient-focused factuality in AI-generated academic outreach.

3. Methods

3.1 Study Design

This is an offline pilot; no generated emails were sent. The design is a 2 × 2 factorial with two binary factors: whether the writing model receives an evidence packet during drafting, and whether a separate verifier revises the draft afterwards (Table 1). This design permits exploratory comparisons of the effects of grounding and post-generation verification, including their possible interaction. Given the pilot sample size, these comparisons are primarily descriptive. The B-versus-C comparison is central, testing whether evidence is more useful during generation or only during repair.

Table 1: Experimental conditions

Condition	Writer receives evidence	Post-generation verification	Purpose
A	**No**	**No**	**Closed-book baseline**
B	**Yes**	**No**	**Effect of grounding during writing**
C	**No**	**Yes**	**Effect of correction after ungrounded writing**
D	**Yes**	**Yes**	**Combined safeguards**

3.2 Sampling Frame & Evidence Packet

We sampled 12 professors across computer science/AI, psychology/cognitive science, and biomedicine/public health - four per field - reported under anonymous IDs (CS-01, PSY-01, BIO-01, …). Professors were selected from OpenAlex[13] using ORCID-linked researcher records. The sample focused on mid-visibility researchers in an attempt to reduce potential familiarity from model pretraining; however, pretraining exposure was not directly measured.

Each packet contains a faculty-page or lab-profile research summary; eight representative papers with title, abstract, year, authors, venue, DOI, and source link; one or two selected passages for a focal paper; the retrieval date; and provenance notes. Abstracts were reconstructed from OpenAlex and normalized to remove residual markup.

The packet is the agreed evaluation source, not a complete record of a career. A claim can therefore be unsupported relative to the packet without being false in the world, and claims it cannot judge fairly are labelled outside evidence scope rather than forced into the unsupported category. These are provisional pilot packets; full multi-source identity verification is future work.

3.3 Generation, Verification & Annotation Procedures

For each professor and condition we generated two independent drafts, giving 12 × 4 × 2 = 96 emails. The writer was OpenAI gpt-4o-mini (temperature 0.7, 220-word limit), using one prompt across all conditions differing only in the presence of the evidence block.

In C and D a separate verifier (Anthropic claude-haiku-4.5, temperature 0) received the draft and the packet under a minimal-edit policy: extract research claims, check them against the packet, preserve supported content, soften overstatements, remove or generalize unsupported or contradicted content. Every deletion and rewrite was logged separately, so the analysis can test whether factuality improved by preserving supported specificity or by erasing risky

content. Models were accessed via OpenRouter; every run recorded model, provider, prompt version, seed, and temperature.

3.4 Claim extraction and automatic evaluation

Each email was decomposed into atomic research claims - the smallest independently checkable statements about the professor's research - by a claim extractor (Anthropic claude-3-haiku), with extraction and labelling as separate steps under separate prompts. Each claim was then labelled against the packet by an automatic evaluator (also claude-3-haiku) using the seven labels in Table 2.

Table 2: Claim labels used for annotation

Label	Meaning	Use in metrics
Supported	The packet directly supports the claim as written	Counts toward supported specificity
Partially supported	Core idea related, but wording broader or less precise than the evidence	Secondary factuality analysis
Overstated	Directionally related but stronger than the evidence permits	Not reported (see §3.6)
Unsupported	No support for a claim the packet should be able to support if justified	Counts as a severe error
Contradicted	The packet directly conflicts with the claim	Counts as a severe error
Outside evidence scope	The packet cannot judge the claim fairly	Excluded from the factual denominator
Subjective/generic	Praise or generic wording with no checkable research fact	Excluded from the factual denominator

The writer (OpenAI) and evaluator (Anthropic) are different families, so the evaluator does not grade its own family's text. This separation does not hold in C and D, where the verifier (claude-haiku-4.5) and evaluator (claude-3-haiku) are both Anthropic models; §5.3 treats this as a limitation. Condition B, which carries the central claim, uses no verifier and is unaffected.

3.5 Reliability assessment

Automatic labels require validation, so we assessed the evaluator two ways.

A human annotator labelled a blinded 36-claim subset (10% of claims) spanning all conditions and fields, working from a name-scrubbed evidence reference and blind to condition and to whether text was an original or verified draft. We report raw agreement, Cohen's κ with a 95% interval, the full confusion matrix, and per-condition coverage. Separately, all 360 claims were relabelled by a second automatic evaluator from a third model family (gemini-2.5-flash-lite). Human annotations served as reference labels for assessing evaluator reliability. They were analysed separately and did not replace the automatic labels used to calculate the full-corpus pilot estimates reported in the primary results. The human subset is small and unevenly distributed across conditions; §4.2 reports that distribution explicitly rather than aggregating over it.

3.6 Metrics and statistical analysis

The primary factuality metric is Severe Error Rate (SER) = (unsupported + contradicted) / judged factual claims. The primary specificity metric is Supported Personalization Density (SPD) = supported research-specific claims per 100 words. Secondary measures are the email-level any-severe-error rate and verifier deletion/rewrite frequencies.

We had pre-specified an overstatement rate. We do not report it: neither automatic evaluator ever emitted the overstated label on any of the 360 claims, so the metric is identically zero by construction rather than by finding. The human annotator did use it. We treat this as a defect of the instrument (§4.2) rather than present a structurally empty metric.

Confidence intervals. Claims are nested within emails within professors, so claims are not independent. We report Clopper–Pearson 95% confidence intervals calculated from pooled claim counts as descriptive binomial intervals. These intervals remain computable when no severe errors are observed, but they assume independent observations and do not account for clustering of claims within emails or professors. Therefore, they should not be interpreted as fully cluster-adjusted measures of uncertainty.

The cluster bootstrap is uninformative for a zero-event condition: if no email contains a severe error, every resample returns zero and the interval collapses to [0, 0]. That collapse is an artifact of the resampling procedure, not evidence the underlying rate is zero, and reporting it as a confidence interval would assert unsupportable certainty. For conditions B and D, we therefore report separate exact intervals for the claim-level severe-error rate and the email-level probability of at least one severe error. These descriptive intervals do not account for clustering within professors and should not be interpreted as bounds on the same underlying parameter.

A logistic mixed-effects model with professor and email random effects was considered but is not estimable at this sample size, particularly for the zero-event conditions.

3.7 Anonymization, blinding, and reproducibility

To reduce the risk of associating potentially fabricated research claims with identifiable individuals, the released analytical materials use anonymized professor identifiers, and the name-to-identifier mapping is withheld. Emails, claims, and the annotation sheet are name-stripped, and annotators see neither the condition nor whether a text is original or verified. Code, prompts, de-identified claims and labels, and figures are released to support reproducibility. The generated email texts and professor identity mapping are withheld.

4. Results

4.1 Primary pilot estimates

We generated 96 emails across 12 professors, four conditions, and two seeds, yielding 360 atomic claims. The automatic label distribution was 246 supported, 62 outside evidence scope, 21 unsupported, 20 partially supported, 10 subjective/generic, and 1 contradicted - a factual denominator of 288 judged claims and 22 severe errors. All estimates here are automatic pilot labels; §4.2 quantifies how far they can be trusted.

Two details in Table 3 matter for interpretation. First, the email-level denominator is the number of emails carrying at least one judged factual claim, not the number generated: in condition A, 2 of 24 emails produced only outside-scope or subjective/generic claims and are excluded, giving 10/22 rather than a rate over 24. Second, B and D contain no severe errors, so their intervals are upper bounds only.

Table 3: Per-condition factuality and supported personalization (automatic pilot labels)

Intervals are exact (Clopper–Pearson) 95% intervals. B and D observed zero severe errors; their intervals are upper bounds and must not be read as demonstrated zeros.

Condition	Emails Generated	Emails with ≥ 1 factual claim	Factual claims	Severe Errors	SER	95% CI	SPD	Email-level ≥ 1 severe	95% CI
A (closed-book)	24	22	61	21	0.344	0.227–0.477	0.65	10/22 (0.455)	0.244–0.678
B (grounded)	24	24	90	0	0.000	0.000–0.040	1.71	0/24 (0.000)	0.000–0.142
C (verify-only)	24	24	50	1	0.020	0.001–0.106	1.04	1/24 (0.042)	0.001–0.211
D (grounded + verified)	24	24	87	0	0.000	0.000–0.042	1.76	0/24 (0.000)	0.000–0.142

The closed-book baseline had a SER of 0.344, with at least one severe error in 10 of 22 judged emails. Grounding (B) produced no severe errors in 90 judged claims under this evaluator - §4.2 shows a second evaluator does not reproduce that zero - and raised SPD from 0.65 to 1.71. Verifying ungrounded drafts (C) left 1 severe error in 50 claims (SPD 1.04). The combined condition (D) also produced none, at the highest specificity (1.76). Figure 1 plots the rates with their intervals; Figure 2 places each condition on the Supported-Specificity Frontier.

The 22 severe errors were 21 in condition A and 1 in condition C; B and D produced none. Factual claims per condition varied (A fewest at 61, B most at 90), reflecting that grounded emails made more, and more specific, checkable statements rather than generic praise - the same effect captured by the rise in SPD.

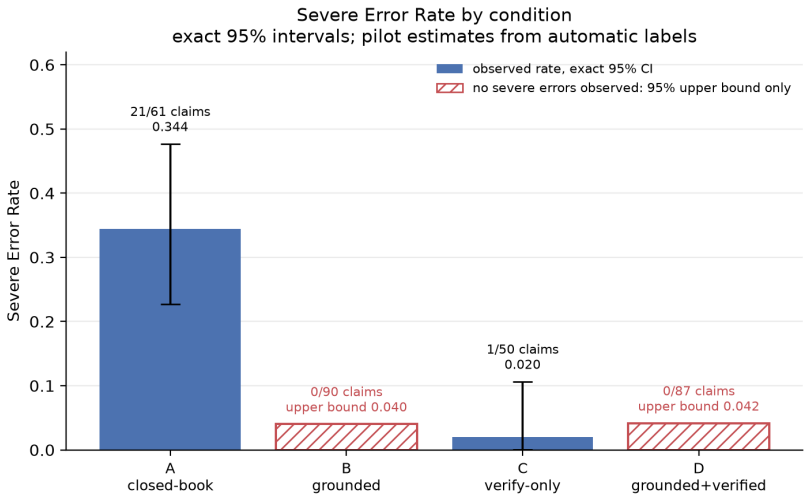


Figure 1. Severe Error Rate by condition, with exact 95% intervals. Solid bars are observed rates (A, C). Hatched bars (B, D) mark conditions where no severe error was observed; the bar height is the 95% upper bound, not a measured rate. All values are pilot estimates from automatic labels.

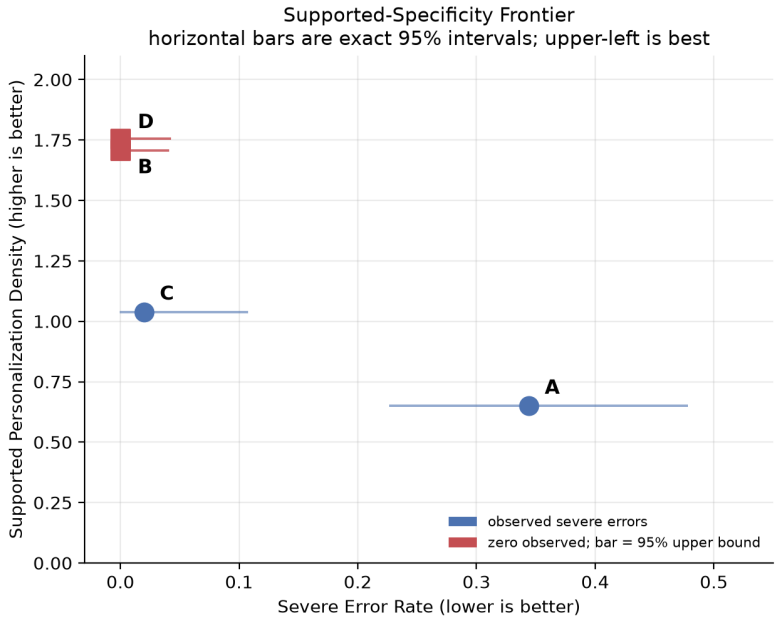


Figure 2. Supported-Specificity Frontier. Each condition is plotted at (x = Severe Error Rate, lower is better; y = Supported Personalization Density, higher is better), with horizontal bars showing exact 95% intervals. A sits at the lower right; B and D at the upper left, where the bar denotes the interval their zero observation is consistent with; C is intermediate.

4.2 Evaluator reliability and the uncertainty it adds

The estimates above come from an LLM judge. This subsection reports how well it tracks human judgment (Table 4), placed here rather than in the Discussion because it bounds every number above.

Table 4. Agreement between annotators. $\varkappa$ intervals are Wald intervals on Cohen's $\varkappa$.

Comparison	Claims	Raw Agreement	Cohen's $\varkappa$	95% CI	Interpretation
Primary evaluator (Claude-3-Haiku) vs human	36	0.556	**0.33**	0.16–0.49	fair
Primary evaluator vs second evaluator (gemini-2.5-flash-lite)	360	0.725	0.50	0.44–0.57	moderate
Second evaluator vs human	36	0.694	**0.54**	0.35–0.74	moderate

Four points follow:

First, the primary evaluator's agreement with the human annotator is only fair, and the entire interval (0.16–0.49) lies below the $\varkappa \geq 0.6$ threshold conventionally required before a rater is treated as reliable. The primary labels are not a validated instrument.

Second, the second automatic evaluator agreed with the human better than the primary did ($\varkappa = 0.54$ versus 0.33). The primary estimates are drawn from the weaker of the two automatic raters; re-running the corpus with a stronger judge is a priority.

Third, and most consequentially, the zeros in B and D are specific to the primary evaluator. Running the second evaluator over the same 360 claims does not reproduce them (Table 5).

Table 5. Per-condition Severe Error Rate under each automatic evaluator. Both labelled all 360 claims independently; factual denominators differ because the evaluators assign the excluded categories differently. Intervals are nominal 95% Clopper–Pearson binomial confidence intervals calculated from pooled claims; they do not account for clustering within emails or professors.

Evaluator	A (closed-book)	B (grounded)	C (verify-only)	D (grounded + verified)
Primary (claude-3-haiku)	21/61 = 0.344 (0.227–0.477)	0/90 = 0.000 (0.000–0.040)	1/50 = 0.020 (0.001–0.106)	0/87 = 0.000 (0.000–0.042)
Second (gemini-2.5-flash-lite)	35/77 = 0.455 (0.341–0.572)	10/94 = 0.106 (0.052–0.187)	11/54 = 0.204 (0.106–0.335)	4/90 = 0.044 (0.012–0.110)

The second evaluator finds 10 severe errors in condition B and 4 in D, whereas the primary found none. What survives across both evaluators is the structure of the result, not the zero: the closed-book baseline is worst under both; both grounded conditions beat both ungrounded ones under both; and B < C, the study's key comparison, holds under both (0.000 vs 0.020; 0.106 vs 0.204). What does not survive is the claim that grounding eliminated severe errors, and the B-versus-D ordering (the second evaluator reports a lower SER for D, whereas the primary evaluator reports identical observed rates for B and D). We therefore treat the grounding effect as a large reduction in severe misrepresentation whose magnitude is judge-dependent - roughly four-fold on the second evaluator's labels - rather than as elimination. This is the strongest reason not to read the zeros in Table 3 as a property of the emails.

Fourth, the human/machine disagreement has a specific and consequential structure (Table 6).

Table 6. Human (rows) × primary automatic evaluator (columns), 36 claims.

Human label ↓	Contradicted	Outside scope	Overstated	Partially supp.	Subj./ generic	Supported	Unsupported	Total
Contradicted	0	**1**	0	0	0	0	0	1
Outside evidence scope	0	3	0	0	0	0	0	3
Overstated	0	0	0	0	0	**1**	0	1
Partially supported	0	0	0	1	0	1	0	2
Subjective/generic	0	0	0	0	0	**6**	0	6
Supported	0	0	0	1	0	16	0	17
Unsupported	0	**6**	0	0	0	0	0	6
Total	0	10	0	2	0	24	0	36

Of the 16 disagreements, the evaluator was too lenient in 14, too strict in 0, and orthogonally different in 2. This pattern suggests that the primary evaluator may underestimate severe-error prevalence, although the small human-labelled subset does not establish the direction or magnitude of bias across the full dataset. The disagreement also reveals a structural problem beyond simple leniency:

Within the human-labelled subset, the primary evaluator classified seven severe errors identified by the human annotator (six unsupported and one contradicted) as outside evidence scope, excluding these claims from the SER denominator. The judge therefore biases the denominator, not only the numerator, so SER is not simply an attenuated version of the true rate.

The evaluator also overestimated supported personalization. Six claims labelled subjective/generic by the human annotator were classified as supported by the automatic evaluator. Of the 24 claims classified as supported by the automatic evaluator, only 16 were human-supported, giving a precision of 0.67 (95% CI 0.47–0.83). This suggests that the automatic SPD estimates may be inflated, although the small validation subset does not support reliable condition-specific corrections.

The "overstated" label was never emitted by either automatic evaluator on any of the 360 claims, though the human annotator used it, making one pre-specified secondary metric unmeasurable with this instrument (§3.6).

Human-labelled coverage is also thin and unevenly spread, limiting what the subset can confirm (Table 7).

Table 7. Human-labelled subset coverage and human-labelled SER per condition. Intervals are exact 95% intervals.

Condition	Human Claims	Judged Factual	Severe	Emails	Human SER	95% CI
A (closed-book)	16	11	7	6	0.636	0.308–0.891
B (grounded)	10	10	0	4	0.000	0.000–0.308
C (verify-only)	7	3	0	3	0.000	0.000–0.708
D (grounded + verified)	3	3	0	1	0.000	0.000–0.708

On the human-labelled subset, the closed-book SER is 0.636 (7/11), well above the automatic 0.344, consistent with the leniency above. But C and D rest on three human-labelled factual claims each - D's from a single email - and their human intervals extend to 0.71. The human subset confirms the closed-book baseline is bad and is consistent with grounding being better; it does not confirm the per-condition estimates for B, C, or D, and we do not claim it does.

4.3 The closed-book effect is concentrated in one field

Pooling across fields conceals heterogeneity that bears directly on how far the headline number generalizes (Table 8).

Table 8. Severe Error Rate by Field and Condition (automatic labels).

Field	A (closed-book)	B (grounded)	C (verify-only)	D (grounded + verified)
Psychology	**17/19 = 0.89**	0/30 = 0.00	0/15 = 0.00	0/31 = 0.00
Computer Science	3/23 = 0.13	0/33 = 0.00	1/20 = 0.05	0/26 = 0.00
Biomedicine	1/19 = 0.05	0/27 = 0.00	0/15 = 0.00	0/30 = 0.00

The pooled closed-book rate of 0.344 is not a homogeneous effect. The four psychology professors account for 17 of the 21 closed-book severe errors (81%). Two professors alone account for 11 of the 21, and 5 of the 12 professors produced no severe errors at all, even closed-book. Excluding psychology, the closed-book SER falls to 0.095 (4/42, 95% CI 0.027–0.226); excluding computer science, it is 0.474, and excluding biomedicine, 0.476.

Figure 3 shows the per-field rates against the pooled estimate. Under the primary automatic evaluator, every sampled field contained at least one severe closed-book error, while neither grounded condition produced a severe error. These field-level findings are descriptive, and the small number of professors per field limits their generalizability. The magnitude of the pooled baseline error rate is dominated by one field containing four professors and should not be read as a general closed-book hallucination rate for academic outreach.

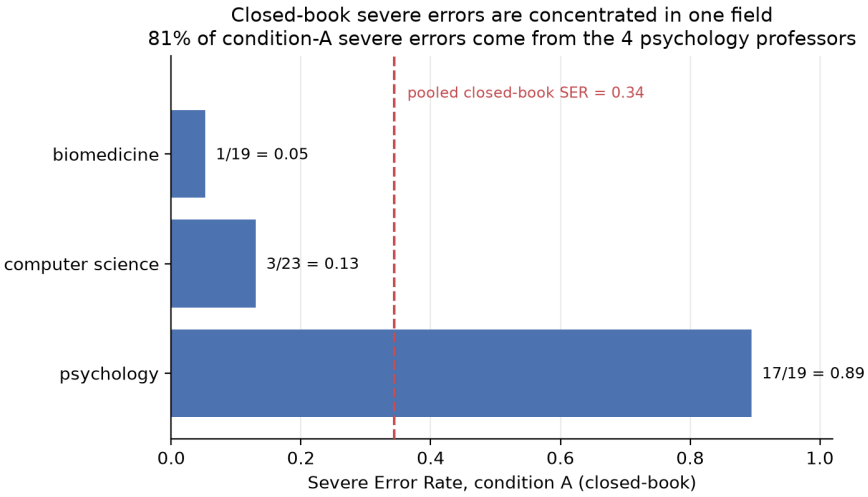


Figure 3. Closed-book (condition A) Severe Error Rate by Field, against the pooled rate of 0.34. The pooled estimate is driven by the psychology subsample; the other two fields fall well below it.

4.4 Verifier edit behaviour

The verifier's edit log distinguishes factuality achieved by preserving supported content from factuality achieved by deletion (Table 9).

Table 9. Verifier edit decisions. Intervals are exact 95% intervals.

Condition	Edit decisions	Delete	Soften/rewrite	Keep	Deletion rate	95% CI
C (verify-only)	**49**	27	14	8	0.551	0.402–0.693
D (grounded + verified)	56	14	14	28	0.250	0.144–0.384

The verifier deleted 55.1% of edit decisions when repairing ungrounded drafts, compared with 25.0% when the draft was already grounded (exploratory Fisher's exact test, $p = 0.0025$). This test does not account for clustering of edit decisions within emails and professors; the p-value should therefore be interpreted cautiously. Correspondingly it kept 28 of 56 decisions in D but only 8 of 49 in C.

5. Discussion

5.1 Principal findings

Grounding at write time (B) cut severe errors sharply while raising supported specificity: factuality and usefulness moved together rather than trading off. Absent evidence, a capable model readily produced plausible but unsupported research descriptions for a real professor - 0.344 under the primary evaluator, 0.455 under the second, 0.636 on the human subset. Supplying the professor's real publications beforehand reduced that to zero under the primary evaluator and 0.106 under the second, while making emails more specific, not less.

Three boundaries apply. First, zero observed errors in 90 claims is not a demonstrated zero: the data are consistent with a true rate up to roughly 4% at the claim level and 14% at the email level. Grounding reduced misrepresentation below this instrument's detection threshold at this sample size; it is not shown to eliminate it, and a deployment claim of "no misrepresentation" would not be supported. Second, that zero belongs to the primary evaluator, not to the emails (§4.2). The defensible form of the central finding is that grounding produced a large reduction in severe misrepresentation while increasing specificity, with the reduction's size judge-dependent and its floor not established. Third, the baseline problem is concentrated in one of three fields (§4.3), so the improvement's size is far less well established than its direction.

5.2 Grounding versus correction

The key comparison, B versus C, favours grounding on both axes (B: SER 0.000, upper bound 0.040, SPD 1.71; C: 0.020, SPD 1.04). The verifier's edits suggest a possible mechanism: the deletion rate was higher when repairing an ungrounded draft than when the draft was already grounded (0.551 versus 0.250). The exploratory Fisher's exact test yielded $p=0.0025$, although this test does not account for clustering within emails and professors. A verifier fixing an ungrounded draft achieves factuality partly by removing risky but specific content, lowering specificity; when the draft was grounded, it finds less to delete and specificity survives. These findings suggest that grounding during drafting may preserve more supported personalization than post-generation verification alone. However, the observed results do not establish a general factuality–specificity trade-off, and the magnitude of this effect remains uncertain.

The observed ranking of grounding alone versus verification alone is consistent across both automatic evaluators, although the magnitude and uncertainty of the difference remain evaluator-dependent. On the primary evaluator the factuality comparison is thin, resting on a single severe error in C, so the ranking would rest mainly on the specificity gap and the deletion mechanism. The second evaluator separates the conditions far more clearly - 0.106 for B against 0.204 for C, on 94 and 54 judged claims - and also ranks both grounded conditions above verify-only. Both automatic evaluators reported a lower pooled SER for grounding than for post-generation verification alone. Grounding also produced a higher observed SPD under the primary evaluator. However, some field-specific SER comparisons were tied, and the uncertainty around the differences remains substantial. This extends grounded-generation findings that evidence during drafting beats post hoc citation[11] into a recipient-focused outreach setting with an explicit specificity measure and operationalizes the personalization-induced-hallucination concern[4] concretely.

5.3 Threats to validity

The primary instrument is not validated. Agreement with the human annotator was fair at best ($\varkappa = 0.33$, 95% CI 0.16–0.49), which is why we report pilot estimates. Disagreement was predominantly lenient in the human-labelled subset (14 of 16 disagreements), suggesting potential underestimation of severe errors, although the small sample does not establish the extent of this bias across the full dataset. But two structural defects mean the bias is not a simple downward scaling: severe claims were reclassified into a category excluded from the denominator (7 of 36 claims), and generic praise was counted as supported (6 of 36), suggesting that the automatic SPD estimates may be inflated. The second evaluator agreed with the human better than the primary did, suggesting that evaluator choice influences reliability estimates, although the small human-labelled subset limits conclusions about their relative performance.

The headline zeros are judge-dependent. The most serious constraint is that the study's most quotable number does not replicate across instruments (§4.2, Table 5). Absolute rates should be interpreted as evaluator-dependent estimates. Although both automatic evaluators reported consistent directional differences between the principal conditions, these comparisons remain preliminary because of limited human validation, evaluator disagreement, and sample heterogeneity.

Verifier and evaluator share a model family. Writer (OpenAI) and evaluator (Anthropic) are deliberately different families, but in C and D the verifier (claude-haiku-4.5) and evaluator (claude-3-haiku) are both Anthropic. The near-zero rates in C and D may partly reflect family agreement between the model that edited the text and the one that graded it. Condition B, which carries the central claim, uses no verifier and is unexposed; the second evaluator, from a third family, still ranks both grounded conditions above both ungrounded ones, and its higher rate for C (0.204) than for either grounded condition is consistent with post hoc repair being genuinely weaker rather than with a family artifact alone. Still, C and D should be treated as more fragile than B.

Effect concentrated in one field. As reported in §4.3, 81% of closed-book severe errors came from four psychology professors and five of twelve professors produced none. The pooled baseline is not a general rate.

Zero-event conditions. Under the primary evaluator two of four conditions contain no events, so the design has little power to distinguish B from D. That D offers no factuality gain over B is a null result at this sample size, not evidence of equivalence - and the second evaluator ranks D slightly ahead of B, a further reason to treat that ordering as undetermined.

5.4 Practical implications

For anyone deploying LLMs to draft academic or professional outreach, retrieval grounding is the more promising default of the two safeguards: grounding was associated with lower observed severe-error rates and higher supported-personalization density than the closed-book baseline. Verification alone also improved these observed metrics relative to baseline, but grounding alone showed more favourable results than verification alone in the comparisons reported here. Post-generation verification remains valuable as an additional safeguard, as it reduced observed severe-error rates in ungrounded drafts. However, verification alone produced lower supported-personalization density than grounding alone, and both automatic evaluators reported less favourable severe-error rates for verification alone.

Two cautions follow. This favours grounding in system design but is not grounds for advertising a factuality guarantee: a deployed assistant should expect residual misrepresentation even when grounded, at a rate that remains uncertain. The primary evaluator observed zero severe errors in B, with an upper 95% confidence limit of approximately 4% under an independence assumption, whereas the second evaluator observed a rate of approximately 11%. And our experience cautions about evaluation practice as much as generation: had we relied on a single automatic judge, as many pipelines do, we would have reported elimination of an error mode a second judge still detects. Teams evaluating factuality should run at least two independent judges before believing a zero.

5.5 Limitations and future work

Human validation covers 36 of 360 claims (10%), with only 3 judged factual claims each in conditions C and D. Although per-condition human estimates are reported in Table 7, their small denominators and wide confidence intervals prevent reliable comparisons between conditions. Scaling human annotation is the highest-value next step; we would prioritise a stratified expansion - every claim the automatic evaluator flagged severe, plus roughly 30 per condition - yielding per-condition human rates with usable intervals at tractable cost. We used a single writer model. Evidence packets are provisional and rely primarily on abstracts and selected passages, so some claims that are true in the full publications may not be adequately assessed. This introduces additional uncertainty into the severe-error estimates, particularly when claims are classified as unsupported rather than outside evidence scope. The sample is small (12 professors, English only) and mid-visibility by construction, so results may not transfer to very famous researchers a model may have memorised. Future

work should scale human labelling, adopt a stronger and validated automatic judge drawn from a family distinct from both writer and verifier, add cross-model writers, incorporate full-text evidence, and broaden field and language coverage.

6. Conclusion

This offline pilot evaluated whether supplying publication evidence during the generation of personalized academic outreach emails reduces factual misrepresentation while preserving research-specific content. Across 96 emails and 360 extracted claims, the observed results suggest that grounding during drafting may provide greater factual reliability and supported personalization than relying on post-generation verification alone.

Both automatic evaluators reported lower severe-error rates for grounding alone than for verification alone, while the primary evaluator also reported higher supported-personalization density. However, the magnitude of these differences remains uncertain. The primary evaluator achieved only fair agreement with the human annotator (Cohen's kappa = 0.33), and its zero-error estimates for grounded conditions were not reproduced by the second evaluator, which recorded severe-error rates of 0.106 for grounding alone and 0.044 for grounding combined with verification.

The study was further limited by a small professor sample, limited human annotation, provisional evidence packets, and substantial variation across sampled disciplines. Consequently, the findings should be interpreted as preliminary evidence favouring grounding during drafting rather than proof of a generalizable effect or elimination of hallucinations.

Future research should prioritize expanded human annotation, stronger and independently validated evaluators, larger and more diverse professor samples, and improved evidence verification.

DECLARATIONS

Abbreviations

LLM—large language model; SER—severe error rate; SPD—supported personalization density; RAG—retrieval-augmented generation; κ—Cohen's kappa; CI—confidence interval.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no competing financial interests or personal relationships that could have influenced this work.

Ethics Statement

This study analysed publicly available scholarly information, curated publication evidence, and AI-generated research outreach emails. No generated emails were sent to the professors, and no professors were recruited as research participants. Human annotation was conducted as part of the research team's evaluation process. Professor identities were replaced with anonymized identifiers in the released analytical materials, and the identity mapping was withheld to reduce the risk of associating potentially fabricated claims with identifiable researchers.

Data Availability Statement

Code, de-identified claim annotations, analysis scripts, and figures are available at:

https://github.com/PhongCT1105/Hack-Research.

Generated email texts and the professor identity mapping are withheld to protect the identities of researchers and avoid disseminating potentially fabricated claims. The email corpus may be made available to editors and reviewers upon request for verification.

AI Use Statement

The authors used Anthropic Claude (via Claude Code) and OpenAI Codex to help implement and test the evaluation pipeline. All AI-generated code and text were reviewed, edited, and verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript. No citation, datum, method, or result was generated without author verification; every statistic was regenerated from the released data using the analysis scripts provided in the accompanying repository, and every reference was checked against the arXiv record. LLMs were also the object of study: the writer (OpenAI gpt-4o-mini), verifier (Anthropic claude-haiku-4.5), claim-extractor and primary evaluator (Anthropic claude-3-haiku), and second evaluator (Google gemini-2.5-flash-lite) are named in the Methods.

Author Contributions

Abhinav Singh - generation/verification pipeline, evaluation infrastructure, analysis. Phong Cao - dataset and evidence packets. Kanan Gasimov - annotation framework and human validation. Azhdar Mammadov - research paper author.

Acknowledgements

The authors acknowledge the collaborative efforts and contributions of all individuals involved in the development and evaluation of this research. We also express our appreciation for the constructive discussions and feedback that helped refine the study's methodology and presentation.

REFERENCES

1. Nicolicioiu A, Iofinova E, Jovanovic A, Kurtic E, Nikdan M, Panferov A, et al. Panza: design and analysis of a fully-local personalized text writing assistant. arXiv:2407.10994. 2024.

2. Wang Y, Jiang J, Zhang M, Li C, Liang Y, Mei Q, et al. Automated evaluation of personalized text generation using large language models. arXiv:2310.11593. 2023.

3. Kong R, Zhang C, Sun R, Chhabra V, Nadimpalli T, Konstan JA. Multi-objective personalization in multi-stakeholder organizational bulk e-mail: a field experiment. arXiv:2302.11156. 2023.

4. Sun Z, Zhan Y, Shen C, Yu W, Zhang X, He M, et al. When personalization misleads: understanding and mitigating hallucinations in personalized LLMs. arXiv:2601.11000. 2026.

5. Min S, Krishna K, Lyu X, Lewis M, Yih W, Koh PW, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251. 2023.

6. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. arXiv:2004.14974. 2020.

7. Wadden D, Lo K, Kuehl B, Cohan A, Beltagy I, Wang LL, et al. SciFact-Open: towards open-domain scientific claim verification. arXiv:2210.13777. 2022.

8. Gao L, Dai Z, Pasupat P, Chen A, Chaganty AT, Fan Y, et al. RARR: researching and revising what language models say, using language models. arXiv:2210.08726. 2022.

9. Fatahi Bayat F, Qian K, Han B, Sang Y, Belyi A, Khorshidi S, et al. FLEEK: factual error detection and correction with evidence retrieved from external knowledge. arXiv:2310.17119. 2023.

10. Ashok D, Kulkarni A, Pham H, Póczos B. SciFix: outperforming GPT3 on scientific factual error correction. arXiv:2305.14707. 2023.

11. Ye X, Sun R, Arik SÖ, Pfister T. Effective large language model adaptation for improved grounding and citation generation (AGREE). arXiv:2311.09533. 2023.

12. Wang Y, Ma X, Nie P, Zeng H, Lyu Z, Zhang Y, et al. ScholarCopilot: training large language models for academic writing with accurate citations. arXiv:2504.00824. 2025.

13. Priem J, Piwowar H, Orr R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833. 2022.

SUPPLEMENTARY MATERIAL

Available in the GitHub repository.