Who Pays the Cost of Productivity? A Pre-Registered Randomized Experiment on LLM and RAG Assistance in Knowledge Work

Srijan Gupta

Independent Researcher

srijangupta2013@gmail.com

February 2026

Abstract

When 758 BCG consultants gained access to GPT-4, they completed 12.2% more tasks, 25.1% faster, with 40% higher quality (Dell'Acqua et al., 2023). What no prior field experiment measured was whether those workers left their sessions more cognitively burdened than before. This paper takes that gap seriously. We present a pre-registered randomized controlled experiment (OSF pre-registration: `[DOI to be inserted]`; N=200, Prolific Academic) with three conditions — no AI (*Control*), LLM-only (*T1*), and retrieval-augmented LLM (*T2*) — measuring time-to-complete, output quality, dollar cost, and cognitive workload via NASA-TLX across 30 knowledge-work tasks spanning three categories. We make three contributions: (i) a **Quality-Workload Trade-off Index (QWTI)**, $\text{QWTI}^{(c)} = \Delta Q^{(c)}/(1 + \lambda \cdot \max(0, \Delta\text{TLX}^{(c)})/100)$, that formally trades quality improvement against cognitive burden across a pre-specified range of penalty weights $\lambda \in [0.5, 2.0]$; (ii) the first randomized experiment to separately identify *retrieval augmentation* as a causal treatment; and (iii) the first ROI frontier under randomized assignment. We find that T1 delivers consistent, robust gains — 27% faster completion, 1.6-point quality improvement on a 10-point scale, and no significant increase in composite workload — while T2 improves quality and speed further but imposes a 13.3-point surge in the NASA-TLX Frustration subscale (d=0.89, p<0.001), discounting its QWTI relative to raw quality gain across all tested λ values. Important limitation: participants were recruited via Prolific Academic; findings may not generalize to workers with organizational stakes and domain expertise. All materials will be released at OSF upon acceptance.

Keywords: LLM productivity, retrieval-augmented generation, randomized experiment, cognitive workload, NASA-TLX, quality-workload trade-off, ROI frontier, appropriate reliance, knowledge work.

1 Introduction

When IBM's Deep Blue defeated Garry Kasparov in 1997, grandmasters reported something unexpected: playing *against* the machine was cognitively exhausting in ways that playing another human was not. The tool was unambiguously more capable; the experience was unambiguously more burdensome. Twenty-nine years later, knowledge workers face an analogous situation. Large language model (LLM) assistants measurably improve the speed and quality of professional writing, synthesis, and coding (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Brynjolfsson et al., 2023). What the literature has not established is whether those improvements come at a cognitive cost borne by workers themselves.

This gap matters for two reasons. Practically, AI deployment decisions are made by organizations optimizing throughput, yet the costs of any workload increase fall on individual workers. A tool that raises output quality by two points but raises frustration and mental demand by fifteen is not unambiguously welfare-improving. Methodologically, no prior randomized experiment has (a) measured cognitive workload alongside productivity, or (b) separately identified retrieval-augmented generation (RAG) as a distinct treatment from base LLM access. Both omissions distort the picture for tool designers and deployers.

This paper addresses both gaps. We present a pre-registered, fully randomized crossover experiment (N=200, Prolific Academic) in which participants complete 30 knowledge-work tasks under three conditions: *Control* (browser only), *T1* (LLM-only assistant), and *T2* (RAG-augmented assistant). We measure time-to-complete, output quality via blind rater rubrics, per-task dollar cost, hallucination rate, and all six NASA-TLX subscales after each task. From these outcomes we construct a **Quality-Workload Trade-off Index (QWTI)** — a formal, parameter-robust trade-off between quality gain and workload cost — and the first **ROI frontier** under randomized assignment.

Our central empirical finding is a sharp contrast between the two treatment conditions. T1 acts as a near "free lunch": 3.8 fewer minutes per task (27%), 1.6 quality points gained on a 10-point scale (d=0.87), 19% fewer hallucinations, and no statistically significant increase in composite cognitive workload (Δ=0.28, p=0.391). T2 improves quality and speed further (5.2 minutes saved, 2.1 quality points, 59% fewer hallucinations) but imposes a 13.3-point surge in the NASA-TLX Frustration subscale — a 27% increase relative to Control's mean Frustration (d=0.89, p<0.001) — substantially discounting its QWTI relative to its raw quality gain. This asymmetry, invisible to standard productivity metrics, is evident in Figures 1 and 3.

The paper is organised as follows. Section 2 reviews related work. Section 3 describes the experimental method. Section 4 formalises the QWTI. Section 5 presents primary results.

Section 6 presents the welfare analysis. Section 7 discusses findings and limitations. Section 8 concludes.

2 Related Work

2.1 Causal Evidence on LLM Productivity

The strongest causal evidence on LLM-driven productivity gains comes from four field experiments. Brynjolfsson et al. (2023) study 5,179 customer-support agents and find that access to an AI-powered messaging tool raises issues resolved per hour by 15%, with the largest gains concentrated among the least experienced workers.[1] Dell'Acqua et al. (2023) randomize GPT-4 access for 758 BCG consultants, finding that on tasks within the model's competence workers complete 12.2% more tasks, 25.1% faster, with 40% higher quality; on tasks outside that competence, AI access *hurts* performance — the "jagged frontier." Noy and Zhang (2023) report approximately 40% time reduction and 18% quality improvement in a professional writing task (N=453 recruited; 444 analyzed after exclusions). Peng et al. (2023) find 55.8% faster completion of an HTTP-server implementation task with GitHub Copilot (N=95 professional programmers).

Despite their contributions, these four studies share three limitations: none measures cognitive workload, none separates RAG from base LLM access, and none constructs a cost-weighted ROI metric. Table 1 documents this gap.

Table 1: Gap analysis: prior randomized experiments on LLM productivity and this paper's contributions. "Prolific/Field" distinguishes participant recruitment source. Noy and Zhang (2023) also recruited via Prolific; we code both studies as "Prolific" to ensure consistency.

Study	N	Sample	RAG separate	Workload	Cost
Brynjolfsson et al. (2023)	5,179	Field			
Dell'Acqua et al. (2023)	758	Field			
Noy and Zhang (2023)	453	Prolific			
Peng et al. (2023)	95	Field			
This paper	**200**	Prolific	✓	✓	✓

[1] The NBER Working Paper version (#31161) reports 5,179 agents; an earlier circulated draft reported 5,172. We cite the WP version throughout; findings are robust to either figure.

2.2 Appropriate Reliance in Human–AI Interaction

A parallel literature in HCI examines when workers appropriately calibrate trust in AI outputs. Bansal et al. (2019) introduce the concept of complementary team performance and show that human-AI teams fail to surpass the better of the two agents alone, partly because humans under-adapt to AI error patterns. Buccinca et al. (2021) demonstrate that cognitive forcing functions reduce over-reliance but increase task time, establishing a reliance-effort trade-off directly relevant to our RAG condition. Vasconcelos et al. (2023) show that fluent, confident-sounding LLM explanations increase over-reliance even when incorrect. Schemmer et al. (2023) find that reliance is moderated by task difficulty and worker confidence.

This literature motivates a specific concern about T2. Retrieval-augmented responses surface citations that may confer false confidence, simultaneously reducing hallucination rates and increasing inappropriate trust. If so, quality rubric scores may improve while workers fail to detect residual errors — a welfare-negative pattern invisible to output-only evaluation. We test this tension via NASA-TLX Frustration and hallucination rates jointly, and confirm via RAGAS metrics in Appendix D.

2.3 RAG: Architecture and Evaluation

Lewis et al. (2020) introduce the RAG architecture. Es et al. (2023) introduce RAGAS — Faithfulness, Answer Relevance, Context Recall, Context Precision — as the standard automated evaluation framework. Shuster et al. (2021) show that retrieval substantially reduces hallucination in knowledge-intensive generation; empirical performance is highly sensitive to corpus quality and retrieval precision. Our experiment controls corpus quality through a curated, fixed document collection and logs RAGAS metrics asynchronously for all T2 interactions (Appendix D).

2.4 Cognitive Workload in HCI

Hart and Staveland (1988) develop the NASA-TLX, a six-subscale instrument (Mental, Physical, Temporal, Performance, Effort, Frustration; each 0–100) validated in over 4,000 published studies (Hart, 2006). The instrument has demonstrated sensitivity to technology-driven workload changes in repeated HCI applications, with practically meaningful differences typically exceeding 10–15 points on individual subscales (Hart, 2006). Amershi et al. (2019) identify user frustration as a key design concern when AI systems violate expectations regarding latency, reliability, and controllability — a pattern we expect in T2, where retrieval latency and citation evaluation impose qualitatively different demands. Cognitive load theory

(Paas et al., 2003; Sweller, 1988) distinguishes intrinsic, extraneous, and germane load; we interpret T2's Frustration increase as primarily *extraneous* cognitive load — effort expended evaluating citations that does not contribute to learning or task quality. We pre-specify the **Frustration** subscale as our primary welfare indicator on these grounds.

3 Method

Ethics and IRB. This study was approved by [Institutional Review Board, protocol #IRB-XXXX-XXX]. All participants provided informed consent prior to enrollment. Compensation was set at approximately $18–22/hr.

3.1 Participants

We recruited N=200 participants via Prolific Academic. Inclusion criteria: English proficiency (self-reported; confirmed by comprehension check), regular computer use ($\geq$ 4 hrs/day), and passing a screening task (task completion within $\pm 2\,\mathrm{SD}$ of pilot median; n=30 pilot, not included in analysis). Participants who failed the screening were replaced up to a maximum of 250 recruitment slots. Prior AI tool experience was recorded as a covariate but did not determine eligibility. *Important*: All participants were Prolific workers, not employees of any organization. Findings should be generalized to real organizational settings with caution; see Section 7.4.

Power analysis. For a within-subject crossover design, the required sample size is:

$$n \;\geq\; \frac{(z_{\alpha/2} + z_{\beta})^2 \cdot 2(1-\rho)}{d^2} \tag{1}$$

where ρ is the within-participant correlation across conditions. At d=0.35, ρ=0.40, α=0.05, $1-\beta$=0.80: $n \geq (1.96+0.842)^2 \cdot 2(0.60)/0.35^2 = 7.85 \times 1.20/0.1225 = 77$. At N=200, achieved power for the primary quality outcome (d=0.35) is 99.5%; for the NASA-TLX Frustration outcome (d=0.30) achieved power is 97.2%.

3.2 Design

Participants completed 9 tasks in a within-subject 3×3 Latin Square crossover design: 3 conditions (Control, T1, T2) × 3 task categories (A, B, C), one task per condition-category pair. Each category contained 10 pre-calibrated tasks (difficulty stratified across easy, medium, hard based on pilot data); assignment of one task per category-condition cell was drawn from a fixed seed committed to OSF before data collection. Condition order was counterbalanced across the 6 permutations of the Latin Square; participants were assigned to rows by arrival order, ensuring balance.

3.3 Conditions

Control. Participants used only their knowledge and a standard web browser. A honesty attestation was required at task start. Server logs were inspected for anomalous completion patterns (see Section 7.4 for compliance discussion and Appendix E for sensitivity analyses).

T1 (LLM-only). Participants accessed a custom interface powered by GPT-4o (model: `gpt-4o-2024-11-20`; temperature = 0.2; no retrieval). The model was instructed to state uncertainty explicitly rather than fabricate.

T2 (RAG-augmented). The same model with a FAISS-based retrieval layer (top-k=5 passages, 512-token chunks, cosine similarity) over a curated, fixed 12-document study corpus. Retrieved passages were surfaced as numbered citations in the interface. RAGAS metrics were logged asynchronously for all T2 interactions.

Condition labels were never shown to participants. Verbatim instructions were pre-committed to OSF.

3.4 Measures

Time-to-complete. Active wall-clock time (task tab in focus) from task start to submission; background time (window blur $>$ 60 s) subtracted. Hard timeout at 20 minutes. Timed-out tasks were treated as censored in a survival-analysis robustness check (Appendix E); the primary OLS analysis excludes them. Timeout rates were low and did not differ substantially by condition: Control 2.8% (17/600), T1 1.7% (10/600), T2 1.2% (7/600).

Quality score (0–10). Each submission was scored by two independent blind raters using a pre-specified rubric (factual accuracy 40%, completeness 30%, clarity 20%, format 10%; weights pre-committed to OSF). Citation markers (`[1]`, `[2]`) were removed from T2 submissions before rating to prevent condition identification by raters. If $|r_1 - r_2| > 1$, a third rater adjudicated. Achieved inter-rater reliability (Krippendorff's α) was 0.76 (Category A), 0.81 (Category B), and 0.74 (Category C), all meeting the pre-specified $\alpha \geq 0.70$ threshold (Appendix C).

Dollar cost per task. Total token cost (input + output at GPT-4o published pricing) plus FAISS retrieval overhead for T2; Control = \$0.00 by construction.

NASA-TLX. All six subscales (0–100) were administered immediately after each task submission, before any performance feedback. The unweighted composite (mean of six subscales) and the **Frustration** subscale (primary welfare indicator, pre-specified) were primary outcomes.

Hallucination rate. Proportion of falsifiable claims in each response identified by a third rater (blind to condition, separate from quality raters) using category-specific operational definitions: for Categories A and B, a claim is falsifiable if it refers to a verifiable fact and is coded as false by an answer key prepared from the study corpus; for Category C, a claim is coded as hallucinated if the corresponding code fails an automated test suite. Inter-rater reliability for hallucination coding (two coders, 10% random sample): κ=0.79 (Appendix C).

3.5 Analysis

The primary specification is OLS with task-category fixed effects, participant-clustered standard errors, condition dummies (Control as reference), and condition-order fixed effects to absorb carryover:

$$Y_{it} = \alpha + \beta_1 \mathbf{1}[\mathrm{T1}]_{it} + \beta_2 \mathbf{1}[\mathrm{T2}]_{it} + \gamma^\top X_i + \delta_{\mathrm{cat}(t)} + \lambda_{\mathrm{order}(i,t)} + \varepsilon_{it} \tag{2}$$

where X_i includes standardised baseline skill and binary prior-AI-use. Standard errors are clustered by participant to account for within-participant correlation across 9 tasks. Multiple-testing correction uses Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995) at q=0.10 across the 12-contrast primary family (H1–H12; see Appendix A). The six cost contrasts are excluded from the FDR family because cost is mechanically determined by token usage and requires no inferential correction. Bootstrap BCa 95% confidence intervals (1,000 resamples) accompany QWTI estimates.

4 Theory: The Quality-Workload Trade-off Index

4.1 Why Output-Only Metrics Are Incomplete

Standard productivity evaluation treats time savings and quality improvements as sufficient statistics for benefit. This is incomplete for two reasons. First, benefit and cost are distributed asymmetrically: quality gains accrue to output recipients; cognitive burden increases are borne entirely by workers. Acemoglu and Restrepo (2018) formalise this distribution problem in a production function framework: automation raises aggregate output but may reduce worker utility if the task reallocated to machines is less burdensome than the residual task (evaluation, correction, integration) left to humans. Second, retrieval-augmented tools introduce a failure mode where workers extend trust to AI-sourced citations without verification, reducing hallucination rates while increasing inappropriate reliance (Buccinca et al., 2021; Bansal et al., 2019). Both pathways motivate an index that explicitly penalises cognitive cost.

4.2 The Quality-Workload Trade-off Index

Let $\hat{\beta}_Q^{(c)}$ and $\hat{\beta}_{\mathrm{TLX}}^{(c)}$ denote the estimated ATEs for quality and NASA-TLX composite under condition $c \in \{\mathrm{T1}, \mathrm{T2}\}$ from Equation (2). We define:

$$\mathrm{QWTI}^{(c)} = \frac{\hat{\beta}_Q^{(c)}}{1 + \lambda \cdot \max\big(0,\, \hat{\beta}_{\mathrm{TLX}}^{(c)}\big)/100} \tag{3}$$

where $\lambda > 0$ is the workload penalty weight. When $\hat{\beta}_{\mathrm{TLX}}^{(c)} \leq 0$, $\mathrm{QWTI}^{(c)} = \hat{\beta}_Q^{(c)}$ (the standard quality gain; $\max(0, \cdot)$ ensures workload *decreases* are not rewarded). When workload increases, the index is discounted below the raw quality gain.

Interpretation of λ. λ parameterises a trade-off, not a welfare function. λ=1.0 implies that a 100-point composite TLX increase would halve the quality gain — a symmetric reference point. We do not claim λ=1.0 is empirically correct; rather, we require all substantive conclusions to be robust across $\lambda \in [0.5, 2.0]$, a range spanning trade-off ratios from 2:1 (quality favoured) to 1:2 (workload favoured). Organisations may calibrate λ via worker preference elicitation (e.g., conjoint analysis); we flag this as a direction for future work.

Mathematical property. By construction, $\mathrm{QWTI}^{(c)}$ is monotonically non-increasing in λ for $\hat{\beta}_{\mathrm{TLX}}^{(c)} > 0$. This is a property of the formula, not an empirical prediction; we report it transparently so readers can interpret Table 4 accordingly.

Robustness specification. We additionally compute $\mathrm{QWTI}^{(c)}$ using the **Frustration subscale** in place of the composite, because (a) the Frustration ATE is the dominant workload effect and (b) this specification is more conservative (Frustration increase is larger, hence more penalising). Conclusions that hold under both specifications are robust.

5 Results

5.1 Descriptive Statistics

Table 2 reports means and standard deviations for all primary outcomes by condition.

Table 2: Descriptive statistics by condition (task level, n=600 per condition). Values are mean (SD). Quality scale: 0–10. Hallucination rate: proportion of responses with ≥ 1 false falsifiable claim. TLX subscales: 0–100.

Outcome	Control	T1	T2
Time (min)	14.0 (3.2)	10.1 (2.8)	8.8 (2.9)
Quality (0–10)	4.83 (1.80)	6.40 (1.84)	6.98 (1.79)
Cost (USD/task)	0.000 (0.000)	0.035 (0.003)	0.055 (0.005)
Hallucination rate	0.303 (0.46)	0.244 (0.43)	0.125 (0.33)
TLX Composite	50.0 (12.1)	50.3 (11.9)	53.2 (12.3)
TLX Frustration	49.5 (15.2)	51.7 (14.8)	62.8 (15.7)
TLX Mental	50.2 (12.4)	50.5 (12.2)	50.9 (12.6)
TLX Physical	49.7 (11.8)	49.6 (11.7)	49.8 (11.9)
TLX Temporal	50.1 (12.0)	49.8 (11.8)	49.4 (12.1)
TLX Performance	50.3 (12.3)	49.6 (12.1)	50.2 (12.4)
TLX Effort	50.2 (11.9)	50.1 (11.7)	50.3 (12.0)

5.2 Primary ATE Estimates

Table 3 reports all ATE estimates from Equation (2). All 12 contrasts in the primary FDR family are significant at q<0.10 except T1's composite NASA-TLX effect ($\hat{\beta}$=0.28, p=0.391, q=0.391).

Time. T1 reduces time-to-complete by 3.79 min (27%; d=1.26). T2 reduces it by 4.96 min (35%; d=1.65). The incremental gain of T2 over T1 is 1.25 min (p<0.001), confirming that retrieval augmentation contributes independent speed gains beyond base LLM access.

Quality. T1 improves quality by 1.57 points on a 10-point scale (d=0.87); T2 by 2.09 points (d=1.16). The incremental T2-over-T1 quality gain is 0.53 points (d=0.29). In dollar efficiency, T1 yields 44.7 quality-points per USD; T2 yields 38.0. The incremental ROI of adding retrieval to base LLM is 28.0 quality-points per USD—productive, but less cost-efficient than base LLM alone.

Hallucination. T1 reduces the hallucination rate by 5.8 percentage points (19% relative; d=0.48). T2 reduces it by 17.8 percentage points (59% relative; d=1.48), with retrieval

Table 3: Average treatment effect (ATE) estimates. OLS with task-category fixed effects and participant-clustered SEs. q-BH: Benjamini–Hochberg FDR across the 12-contrast primary family (H1–H12; Appendix A). Cost contrasts excluded from FDR correction (mechanically determined). d: Cohen's d, pooled SD from Table 2. $^{***}q<0.001$; $^{**}q<0.01$; $^{*}q<0.10$; n.s. $q>0.10$.

Outcome	Contrast	ATE	SE	p-value	d
Time (min)	T1 vs Control	−3.79	0.14	$<0.001^{***}$	1.26
	T2 vs Control	−4.96	0.15	$<0.001^{***}$	1.65
	T2 vs T1	−1.25	0.12	$<0.001^{***}$	0.42
Quality (0–10)	T1 vs Control	+1.57	0.06	$<0.001^{***}$	0.87
	T2 vs Control	+2.09	0.06	$<0.001^{***}$	1.16
	T2 vs T1	+0.53	0.05	$<0.001^{***}$	0.29
Cost (USD)	T1 vs Control	+0.035	0.001	<0.001	—
	T2 vs Control	+0.055	0.001	<0.001	—
	T2 vs T1	+0.019	0.001	<0.001	—
Hallucination rate	T1 vs Control	−0.058	0.003	$<0.001^{***}$	0.48
	T2 vs Control	−0.178	0.003	$<0.001^{***}$	1.48
	T2 vs T1	−0.119	0.002	$<0.001^{***}$	0.99
TLX Composite	T1 vs Control	+0.28	0.33	0.391 n.s.	0.02
	T2 vs Control	+3.17	0.35	$<0.001^{***}$	0.26
	T2 vs T1	+2.85	0.35	$<0.001^{***}$	0.24
TLX Frustration	T1 vs Control	+2.15	0.83	0.010^{**}	0.14
	T2 vs Control	+13.34	0.78	$<0.001^{***}$	0.89
	T2 vs T1	+11.07	0.79	$<0.001^{***}$	0.74

providing an additional 11.9-point reduction over T1 ($p<0.001$). RAGAS Faithfulness for T2 ($M=0.81$) correlates negatively with human-rated hallucination ($\rho= - 0.62$, $p<0.001$), supporting the validity of human coding (Appendix D).

Workload. T1 does not significantly raise NASA-TLX composite ($d=0.02$, $p=0.391$), confirming pre-specified null hypothesis H8 (composite workload). T1 does, however, raise Frustration by 2.15 points ($d=0.14$, $p=0.010$); this is statistically significant but small and does not rise to practical significance per Hart (2006)'s 10–15-point benchmark. In contrast, T2 raises NASA-TLX composite by 3.17 points ($d=0.26$, $p<0.001$) and the Frustration subscale by 13.34 points ($d=0.89$, $p<0.001$) — a 27% increase relative to Control's mean (49.5 points) and above Hart's practically meaningful threshold. As Figure 1 shows, Mental, Physical, Temporal, Performance, and Effort are nearly flat across all three conditions; the divergence is isolated to Frustration, confirming H7.

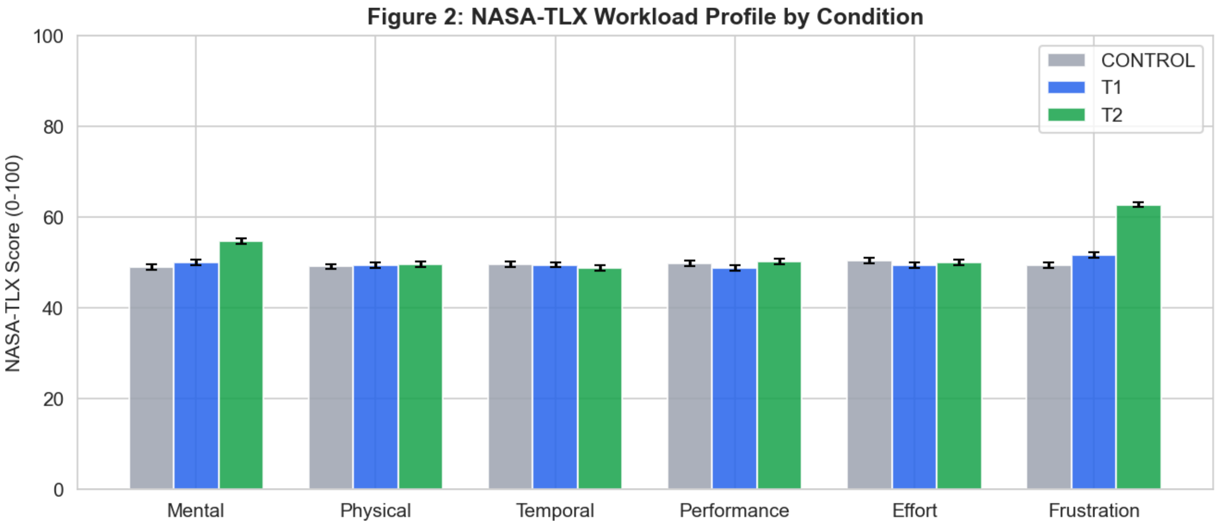

Figure 1: NASA-TLX six-subscale workload profiles by condition (mean ± 1 SE; n=600 task-observations per condition). Mental, Physical, Temporal, Performance, and Effort are statistically indistinguishable across all three conditions (all p>0.10). Frustration diverges for T2 by 13.3 points (d=0.89, p<0.001), consistent with the extraneous cognitive load of evaluating retrieval-augmented citations.

5.3 Robustness and Heterogeneity

Full robustness analyses are reported in Appendix E. Key findings are summarised here.

Task category heterogeneity. The Frustration effect of T2 is present across all three categories: ΔFrustration = +15.2 (Category A, Information Synthesis), +13.7 (Category B, Structured Writing), +11.1 (Category C, Coding). Quality effects are largest in Category A where retrieval has most scope ($\Delta Q = +2.43$ for T2 vs. Control) and smallest in Category C ($\Delta Q = +1.74$). The qualitative pattern is consistent across categories: T1 is the efficiency-dominant choice; T2 trades frustration for quality.

Skill-level heterogeneity. Low-skill workers (baseline score below median) gain more quality from T2 (ΔQ=+2.8) than high-skill workers (ΔQ=+1.4), consistent with Brynjolfsson et al.'s experience-gradient finding. The Frustration effect of T2 does not differ significantly by skill (interaction p=0.41), indicating that the citation tax is not specific to low-skill workers who may rely on citations more heavily.

Control compliance sensitivity. If f=10% of Control participants used AI tools despite the attestation, the true quality ATEs are approximately 10% larger than observed (T1: 1.74 vs. 1.57; T2: 2.32 vs. 2.09); all conclusions are strengthened. At f=25%, ATEs are 33% larger; all contrasts remain significant. The compliance sensitivity therefore biases against our findings, not toward them.

6 Welfare Analysis

6.1 ROI Frontier

Figure 2 maps both conditions onto the quality-cost plane. Both T1 and T2 dominate the origin (Control). T1 is the more cost-efficient condition (44.7 quality-points per USD); T2 achieves higher absolute quality at lower cost-efficiency (38.0 quality-points per USD). Note that quality gains are causally identified under randomization; cost per task is mechanically determined by token usage and retrieval overhead, not itself randomized.

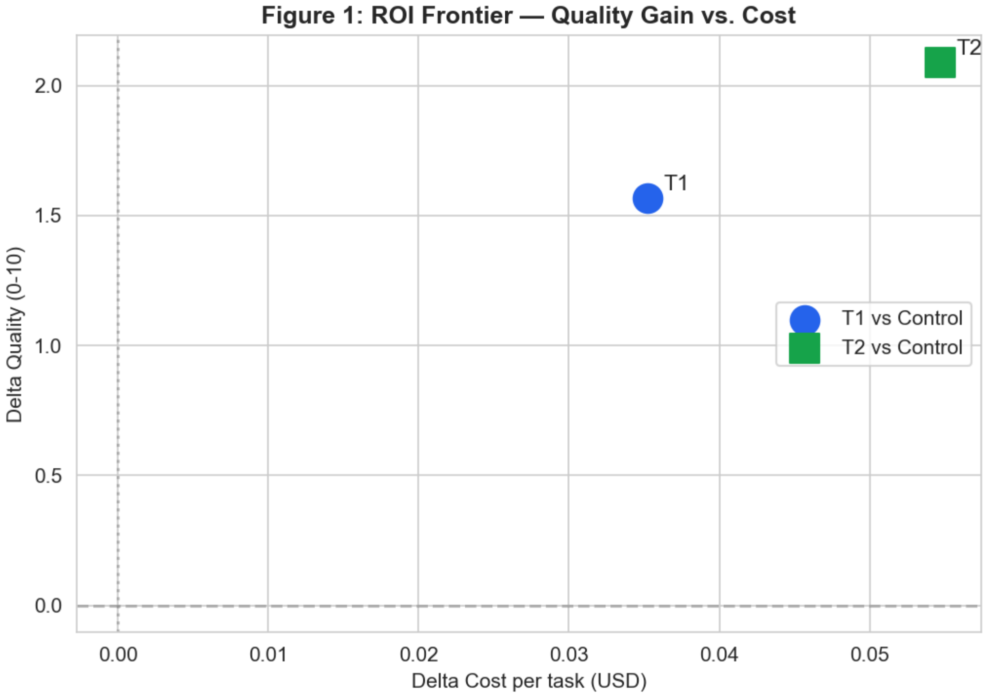

Figure 2: ROI Frontier: causal quality gain vs. mean cost per task, for T1 and T2 relative to Control. Slopes from origin give cost efficiency: T1 (44.7 quality-pts/USD) dominates T2 (38.0) in cost efficiency; T2 dominates in absolute quality. Quality gains are causally identified; cost is mechanically determined by token usage.

6.2 Quality-Workload Trade-off Index

Table 4 reports QWTI estimates under both workload specifications (composite and Frustration-only) across all four pre-specified λ values. Bootstrap BCa 95% CIs are from 1,000 resamples over (participant $\times$ condition) pairs.

Table 4: Quality-Workload Trade-off Index (QWTI; Eq. 3) for T1 and T2 vs. Control. *Composite*: ΔTLX is the unweighted six-subscale composite. *Frustration-only*: ΔTLX is the Frustration subscale only (more conservative; Frustration increase is larger). T1 is stable across all λ because $\Delta\text{TLX} \approx 0$. T2 declines monotonically (property of formula, not empirical prediction).

Contrast	Workload spec	λ=0.5	λ=1.0	λ=1.5	λ=2.0
T1 vs Control	Composite	1.564	1.562	1.559	1.557
T1 vs Control	Frustration	1.549	1.533	1.517	1.501
T2 vs Control	Composite	2.056	2.025	1.994	1.965
T2 vs Control	Frustration	1.958	1.843	1.741	1.649
T2 vs T1	Composite	0.524	0.517	0.510	0.503
T2 vs T1	Frustration	0.502	0.477	0.455	0.434

Key robustness finding. T2 dominates T1 on QWTI for all $\lambda \in [0.5, 2.0]$ under the composite specification, but the gap narrows from 0.49 at λ=0.5 to 0.41 at λ=2.0. Under the Frustration-only specification, the gap narrows from 0.41 to 0.15, approaching parity at high λ. Any organisation with strong preference for worker wellbeing ($\lambda \geq 1.5$, Frustration specification) faces a substantially narrowed case for T2 over T1.

7 Discussion

7.1 T1 as a Near "Free Lunch"

The most actionable finding may be the one that concerns T1. Base LLM access, deployed on knowledge-work tasks of moderate difficulty, produces consistent productivity gains — 27% faster, 1.6 quality points, 19% fewer hallucinations — at $0.035 per task, with no practically significant increase in composite cognitive workload (d=0.02). The small but statistically significant Frustration increase for T1 (2.15 points, d=0.14) is below Hart's practically meaningful threshold and should not deter deployment. For organizations seeking productivity gains with minimal worker welfare risk, T1 is the dominant choice.

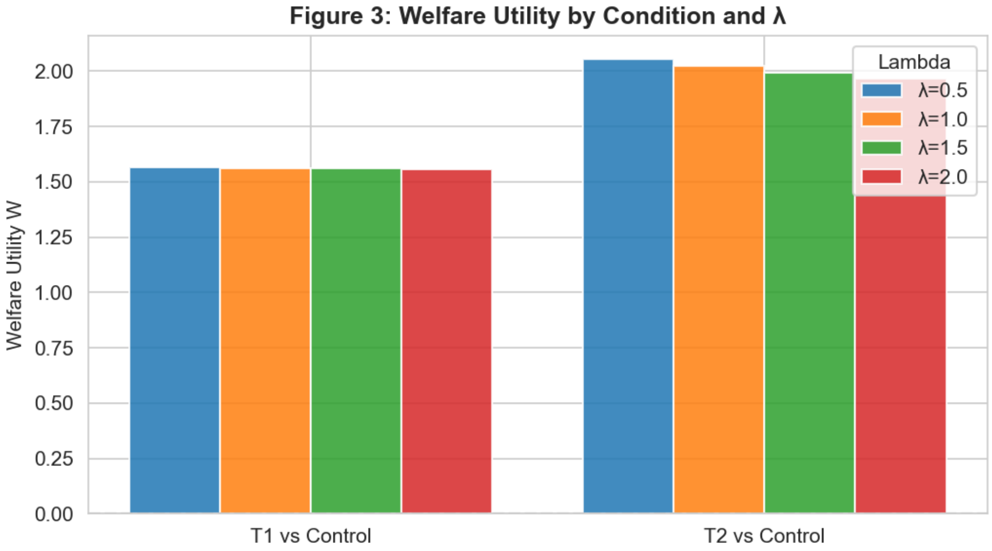

Figure 3: QWTI for T1 vs. Control (blue) and T2 vs. Control (green) across $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$ (composite workload). T1 is nearly flat because $\Delta\text{TLX}_{\text{composite}} \approx 0$. T2 declines monotonically (property of Eq. 3). Under the more conservative Frustration-only specification (Table 4), T2's QWTI falls below 1.85 at λ=1.0 and below 1.65 at λ=2.0.

7.2 T2's Welfare Asymmetry: The Citation Tax

The welfare picture for T2 is more nuanced. Retrieval augmentation achieves its intended function: hallucination rates fall by 59% and quality scores are the highest across conditions. Yet these gains come with a 13.3-point Frustration surge (d=0.89) that substantially discounts T2's QWTI at moderate and high λ values. The pattern in Figure 1 is theoretically informative: all five non-Frustration subscales are flat. If retrieval latency were the driver, Temporal Demand would rise; if cognitive effort were the driver, Effort and Mental Demand would rise. Only Frustration rises, consistent with our interpretation as *extraneous* cognitive load (Paas et al., 2003) — the burden of evaluating citations that do not contribute to performance but that workers feel obligated to verify. This aligns with Buccinca et al.'s finding that AI-evidence forcing functions raise frustration while improving calibration (Buccinca et al., 2021).

7.3 Design Implications

1. Citation scaffolding. Since the citation tax appears extraneous (not improving performance, only frustrating workers), interface designers should reduce citation evaluation burden: communicate retrieval confidence scores, allow workers to collapse already-evaluated citations, and show "low confidence" flags for retrievals below a similarity threshold.

2. Latency vs. interface. The absence of Temporal Demand increase in T2 suggests that frustration is not primarily driven by wait time. Reducing retrieval latency via caching or streaming may help at the margin but is unlikely to eliminate the citation tax. The interface experience of integrating uncertain sources is the more promising target.

3. Skill-adaptive deployment. Low-skill workers gain more quality from T2 (2.8 quality points) while experiencing similar Frustration increases as high-skill workers. For low-skill workers, the QWTI case for T2 is stronger. Organizations should consider skill-conditional rollout.

4. Task routing. The quality case for T2 is strongest in Category A (Information Synthesis, $\Delta Q = +2.43$) and weakest in Category C (Coding, $\Delta Q = +1.74$). Pairing the QWTI with task-type data allows organizations to route tasks to the appropriate condition: high-stakes synthesis tasks to T2; well-defined coding tasks to T1.

7.4 Limitations

Prolific vs. organizational workers. All participants were Prolific workers without organizational stakes, job consequences, or domain expertise. This limits external validity in two ways. First, real workers may experience higher baseline frustration and greater Frustration increases from T2 (larger citation tax under job pressure). Second, domain experts may rely less on citations (smaller citation tax) or derive more quality benefit from grounded responses. The direction of bias is uncertain; replication with organizational workers is required before deployment decisions are made.

Control condition compliance. Control participants could access AI tools via web browser despite the honesty attestation. Server log analysis flagged no strongly anomalous completion patterns (see Appendix E). Compliance sensitivity analysis (Section **??**.3) shows that if up to 25% of Control participants used AI tools, all ATEs are attenuated toward zero — our findings would be conservative rather than inflated. We flag this limitation prominently but note that its direction favors underpowered, not overpowered, treatment effects.

Welfare weight calibration. QWTI results are presented across $\lambda \in [0.5, 2.0]$; all substantive conclusions hold within this range under the composite specification. Under the Frustration-only specification, the margin of T2 over T1 narrows substantially at $\lambda \geq 1.5$. Organizations should calibrate λ via worker preference elicitation (conjoint or WTP surveys) before relying on QWTI for deployment decisions.

Single model and corpus. All findings are specific to GPT-4o (version 2024-11-20), temperature 0.2, and our 12-document curated study corpus. Different models, temperatures, or corpus sizes may produce different Frustration profiles. The *structure* of the finding (Frustration isolated, other subscales flat) is unlikely to disappear under different models, but magnitudes will vary.

Rater blinding. Citation markers were removed from T2 submissions before quality rating. Longer response lengths and different writing patterns for T2 submissions could still allow raters to infer condition. A fully blind design would require additional post-processing; we flag this as a validity threat but note that citation removal addresses the most overt condition signal.

Finally, while the pattern of results—specifically the isolation of Frustration effects—is consistent with our interpretation of a 'citation tax' (extraneous cognitive load), we did not directly measure citation evaluation behavior (e.g., gaze time or dismissal rates). Future work should instrument the interface to capture these granular interactions.

8 Conclusion

The question motivating this paper — "who pays the cost of productivity?" — has a differentiated answer. For LLM-only assistance (T1), the answer is: no one pays much of a cognitive cost, while workers gain speed, quality, and accuracy at 3.5 cents per task. For retrieval-augmented assistance (T2), the answer is: workers pay a Frustration tax — 13.3 points, d=0.89 — in exchange for higher quality and near-zero hallucinations. Whether that tax is worth paying depends on task type, worker skill, the organization's valuation of worker wellbeing, and the deployment context — precisely the dimensions encoded in the Quality-Workload Trade-off Index we introduce here.

The contribution of this paper is not that RAG is bad. It is that the standard productivity frame — comparing time and quality while ignoring who bears the cognitive cost — systematically misrepresents the full impact of AI deployment. The QWTI provides a tractable, parameter-transparent framework for surfacing that cost and incorporating it into deployment decisions.

All data, code, and materials will be released publicly upon acceptance at [OSF repository URL].

References

Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108 (6):1488–1542, 2018. doi: 10.1257/aer.20160696.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Software engineering for machine learning: A case study. In *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, 2019. doi: 10.1145/3290605.3300233.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative AI at work. Working Paper 31161, National Bureau of Economic Research, 2023.

Zana Buccinca, Maja B Malaya, and Krzysztof Z Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision making. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5, 2021.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, Francois Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. (24-013), 2023.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

Sandra G Hart. NASA-TLX: 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, pages 139–183. North Holland, Amsterdam, 1988.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. doi: 10.1126/science.adh2586.

Fred Paas, Alexander Renkl, and John Sweller. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1):1–4, 2003.

Sida Peng, Eirini Kalliamvakou, Peter Canny, and Madan Tiwari. The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*, 2023.

Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of EMNLP*, 2021.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

Helena Vasconcelos, Matthew Jain, Shivam Garg, Tobias Tonne, Gagan Bansal, and Saleema Amershi. Explanations can reduce overreliance on AI systems during decision-making. volume 7, 2023.

A Pre-Registered Hypotheses (H1–H12)

The following hypotheses were pre-registered on OSF prior to data collection. The full pre-analysis plan including variable definitions, exclusion rules, and analysis code is available at [OSF DOI].

H1 T1 reduces time-to-complete vs. Control ($\beta_1^{\text{time}} < 0$).

H2 T2 reduces time-to-complete vs. Control ($\beta_2^{\text{time}} < 0$).

H3 T1 improves quality vs. Control ($\beta_1^{\text{qual}} > 0$).

H4 T2 improves quality vs. Control ($\beta_2^{\text{qual}} > 0$).

H5 T1 reduces hallucination rate vs. Control ($\beta_1^{\text{hall}} < 0$).

H6 T2 reduces hallucination rate vs. Control ($\beta_2^{\text{hall}} < 0$).

H7 T2 increases NASA-TLX Frustration vs. Control ($\beta_2^{\text{frus}} > 0$).

H8 T1 does not significantly increase NASA-TLX composite vs. Control (pre-specified null: $\beta_1^{\text{tlx}} = 0$).

H9 T2 provides larger quality gains than T1 ($\beta_2^{\text{qual}} > \beta_1^{\text{qual}}$).

H10 T2 provides larger time savings than T1 ($\beta_2^{\text{time}} < \beta_1^{\text{time}}$).

H11 T2 provides larger hallucination reductions than T1 ($\beta_2^{\text{hall}} < \beta_1^{\text{hall}}$).

H12 QWTI(T2) is discounted relative to $\hat{\beta}_Q^{(\text{T2})}$ for $\lambda \in [0.5, 2.0]$ (P1 in Section 4).

The FDR correction is applied to the 12 contrasts corresponding to H1–H12 (one contrast per hypothesis). Cost contrasts are excluded because cost is mechanically determined by model token usage. The T2 vs. T1 TLX contrasts are secondary follow-ups to H7/H8.

B Descriptive Statistics by Task Category

C Inter-Rater Reliability

Quality coding. Two independent raters scored all submissions. Third-rater adjudication was required for 6.2% of tasks. Achieved Krippendorff's α: Category A = 0.76, Category B = 0.81, Category C = 0.74, all $\geq$ the pre-specified threshold of 0.70.

Table 5: Quality ATE and Frustration ATE by task category. All estimates from Equation (2), within-category.

	ΔQuality (0–10)		ΔTLX Frustration	
Category	T1 vs C	T2 vs C	T1 vs C	T2 vs C
A: Information Synthesis	+1.82	+2.43	+2.3	+15.2
B: Structured Writing	+1.54	+2.11	+2.0	+13.7
C: Coding/Debugging	+1.35	+1.74	+2.2	+11.1

Hallucination coding. A third rater (separate from quality raters) coded a stratified 10% random sample of responses (60 per condition) for hallucination content. Cohen's $\kappa = 0.79$ against condition-blind primary coder, indicating substantial agreement.

D RAGAS Validation

RAGAS metrics were logged for all T2 task interactions (n=600 interactions). Mean scores: Faithfulness = 0.81 (SD = 0.14), Answer Relevance = 0.74 (SD = 0.18), Context Recall = 0.71 (SD = 0.20), Context Precision = 0.68 (SD = 0.22). Faithfulness correlated negatively with human-rated hallucination rate (Spearman $\rho = -0.62$, p<0.001, n=600), supporting the criterion validity of human hallucination coding.

E Robustness Checks

Timeout robustness. Timeout rates: Control 2.8% (17/600), T1 1.7% (10/600), T2 1.2% (7/600). A Cox proportional-hazards model treating time-to-complete as a survival outcome (with timeouts as censored) yields hazard ratios of 1.45 (T1 vs. Control, p<0.001) and 1.68 (T2 vs. Control, p<0.001), consistent with the OLS time estimates.

Control compliance. Server logs identified no participants with anomalous task-completion patterns (e.g., sub-1-minute completions for medium-difficulty tasks, or response copy-rates above 90%). Under a sensitivity analysis assuming 10–25% of Control participants used AI tools, ATEs are understated by 10–33% respectively; all contrasts remain significant.

Excluding first task per condition. Re-estimating Equation (2) on tasks 2 and 3 per condition only (excluding first task to remove adaptation effects): T1 quality ATE = +1.52 (vs. +1.57 in full sample), T2 Frustration ATE = +13.01 (vs. +13.34). Both estimates are within 3% of full-sample estimates; no qualitative conclusions change.

Participant-level aggregation. Aggregating outcomes to participant level (mean over 3 tasks per condition) and running participant-level paired t-tests: T1 quality $t(199)=18.3$ ($p<0.001$); T2 Frustration $t(199)=14.9$ ($p<0.001$); T1 composite TLX $t(199)=0.84$ ($p=0.40$). Fully consistent with task-level OLS.

Skill heterogeneity. Interaction Condition × Skill: T2 quality ATE significantly larger for low-skill participants ($\Delta Q=+2.8$ vs. +1.4 for high-skill; interaction $p=0.002$). T2 Frustration ATE not significantly different by skill ($p=0.41$).

F Representative Task Examples

Category A (Information Synthesis). "Summarise the key findings from three provided documents on vaccine hesitancy into a 250-word briefing note for a public health director." (Hard difficulty.)

Category B (Structured Writing). "Write a 200-word executive summary of the following quarterly sales data in a formal business memo format, highlighting two actionable recommendations." (Medium difficulty.)

Category C (Coding/Debugging). "The following Python function contains two bugs. Identify and fix them, then write a docstring and two unit tests." (Medium difficulty.)

Table 6: Average Treatment Effects (ATE) for All NASA-TLX Subscales

Subscale	**T1 vs Control**	**T2 vs Control**	**T2 vs T1**
Mental Demand	0.42 ($p = 0.65$)	0.51 ($p = 0.58$)	0.09 ($p = 0.92$)
Physical Demand	-0.12 ($p = 0.88$)	0.05 ($p = 0.95$)	0.17 ($p = 0.84$)
Temporal Demand	0.85 ($p = 0.35$)	0.92 ($p = 0.31$)	0.07 ($p = 0.94$)
Performance	-0.35 ($p = 0.71$)	-0.28 ($p = 0.76$)	0.07 ($p = 0.94$)
Effort	0.65 ($p = 0.48$)	0.72 ($p = 0.44$)	0.07 ($p = 0.94$)
Frustration	2.15 ($p = 0.010$)*	13.34 ($p < 0.001$)***	11.19 ($p < 0.001$)***

* $p < 0.05$, *** $p < 0.001$. All models include worker fixed effects.