

Users Mispredict What Drives Their Preferences for AI Writing Assistance

Vivian Lai¹ Zana Bućinca^{2*†}
Hyeonsu B. Kang^{1*} Mo Houtti¹ Nil-Jana Akpınar¹
Kevin Chian¹ Namjoon Suh¹ Alex C. Williams¹

¹ Microsoft ² Massachusetts Institute of Technology
viv.lai@microsoft.com

Abstract

Proactive AI writing assistants need to predict when users want drafting help, yet evidence about which contextual cues drive these preferences remains limited. Through a factorial vignette study with 50 participants making 750 pairwise comparisons in professional-email scenarios, we find that compositional effort is the strongest measured predictor ($\rho = 0.597$), whereas urgency has a near-zero scenario-level correlation and no reliable effect across inferential analyses. More critically, we observe a striking stated–revealed gap between abstract factor rankings and situated choices: they rank urgency first in self-reports despite it being the weakest behavioral driver, a contrast at the extremes even though mean stated urgency and effort ranks are close, i.e., 2.06 vs. 2.18. This misalignment has measurable consequences. Within this controlled professional-email testbed, a behavior-derived rule reaches 61.3% accuracy, significantly outperforming a stated-preference rule at 57.7% ($p < 0.05$). The 3.6-point difference shows that the stated–revealed gap has measurable consequences even under the same four-factor representation. It also identifies a concrete design risk for proactive email assistance: stated rankings can place weight on cues that do not predict situated choices. Future studies should test whether stated rankings create similar design risks in other writing tasks.

1 Introduction

LLMs have enabled increasingly capable AI writing assistants, yet a fundamental design question remains unresolved: When should systems proactively offer drafting help? While reactive assistance responds to explicit user requests, proactive systems need to infer opportune moments from context. Prior work indicates that well-timed initiative,

grounded in the costs and benefits of action under uncertainty, can improve user experience, whereas poor timing raises interruption costs and user frustration (Horvitz, 1999; Iqbal and Bailey, 2008; McFarlane, 2002). Despite widespread deployment of such systems in email clients (e.g., Smart Reply (Kannan et al., 2016), Smart Compose (Chen et al., 2019)), designers still lack robust evidence on which contextual cues reliably predict actual desire for assistance. Moreover, stated preferences and subjective measures often misalign with observed behavior and performance, and suggestion UIs can shift composition, introducing trade-offs for proactive assistance (Bućinca et al., 2020; Poursabzi-Sangdeh et al., 2021; Buschek et al., 2021).

A common intuition is to offer assistance under time pressure, but interruption research emphasizes task structure and breakpoints over urgency as better predictors of lower disruption timing (Adamczyk and Bailey, 2004; Iqbal and Bailey, 2008). An alternative is to expose user controls and preference settings, consistent with human-AI interaction guidelines (Amershi et al., 2019), while recognizing that self-reports may not faithfully reflect assistance needs (Bućinca et al., 2020).

We investigate these questions through professional email drafting as a controlled testbed for proactive assistance preferences. Email represents a high-frequency writing task with measurable contextual factors and practical relevance as a deployed AI assistance application. We document the gap and quantify its consequences in controlled professional-email scenarios.

Through a factorial vignette study, we examine what email attributes predict AI assistance preferences. We manipulate four dimensions: urgency, compositional effort, sender importance, and email type, across 16 scenarios. Fifty participants made 750 pairwise comparisons and reported their stated preference drivers. We address three research questions: (1) What email attributes drive preferences

*Equal contribution.

†Work completed while a postdoctoral researcher at Microsoft.

for AI drafting assistance? (§4) (2) Do users accurately perceive what drives their preferences for AI assistance? (§5) (3) Do ML models predict preferences from behavioral patterns, and what is the cost of the stated–revealed gap? (§6)

Our findings reveal a striking misalignment. Urgency receives the highest mean self-reported rank (2.06 out of 4), yet behavioral analysis shows urgency has near-zero predictive power ($\rho \approx 0$). Instead, compositional effort dominates ($\rho = 0.597$). Although the mean stated urgency-versus-effort rank difference is small, the ordering at the extremes is reversed: urgency is most cited but behaviorally weakest, while effort is behaviorally strongest but only mid-ranked in self-reports.

To quantify the misalignment, we train ML models to predict individual preferences and compare three rule-based strategies, namely uniform weighting, stated-preference weighting from user surveys, and behavioral weighting from revealed preferences. The behavior-derived rule reaches 61.3% accuracy, significantly outperforming the stated-preference rule at 57.7% ($p < 0.05$). The stated-preference rule also falls below the uniform-weighting baseline at 58.9%.

We make the following three contributions:

1. We provide empirical evidence that compositional effort, not urgency, drives AI drafting assistance preferences. This reorients design priorities from temporal signals toward compositional-effort detection.
2. We document a stated–revealed gap in which abstract dimension rankings and situated email choices prioritize the same cues differently. In this email setting, the gap shows why explicit preference settings should be validated against behavioral evidence.
3. We quantify the cost of the stated–revealed gap, showing that behavioral patterns achieve 61.3% prediction accuracy, significantly outperforming stated-preference designs at 57.7% ($p < 0.05$). The 3.6-point difference demonstrates a measurable cost under the study’s controlled feature set and motivates field evaluation of behavior-derived policies.

By isolating this stated–revealed gap, the study provides a behavior-grounded starting point for professional-email intervention design.

2 Related Work

Email management & proactive AI assistance.

Research on email management has long argued for and designed more intelligent assistance (Whittaker and Sidner, 1996; Bellotti et al., 2003), establishing that users face systematic challenges in handling email volume and allocating response effort (Dabish and Kraut, 2005). Synthesizing prior work, we focus on four contextual factors that influence email behavior: urgency (Tyler and Tang, 2003), compositional effort (Bellotti et al., 2003; Whittaker and Sidner, 1996), sender importance (Dabish and Kraut, 2005), and email type (Mackay, 1988; Whittaker and Sidner, 1996; Ducheneaut and Bellotti, 2001). We selected these four factors because prior email and interruption research repeatedly links them to response behavior, they span temporal, compositional, relational, and task-novelty cues, and they can be manipulated independently in a tractable factorial design. The set is theoretically motivated rather than exhaustive. Privacy, emotional tone, factual risk, recipient relationship, and message sensitivity remain important extensions for larger designs.

Recent advances in LLMs have enabled proactive writing assistance (Lee et al., 2022; Buschek et al., 2021). When timed to natural breakpoints, proactive assistance reduces reaction time and frustration and lowers interruption costs (Horvitz, 1999; Cutrell et al., 2001; Iqbal and Bailey, 2008). However, poorly timed assistance disrupts workflow (Iqbal and Horvitz, 2010; Czerwinski et al., 2004; Cutrell et al., 2000). Although many deployed tools remain reactive, even prominent email assistants (Chen et al., 2019; Kannan et al., 2016) rarely report empirically validated timing policies. Our study addresses this gap through a controlled factorial investigation of how four contextual factors predict preferences for AI drafting assistance in professional-email scenarios.

User perception in AI systems. Recent research highlights mismatches between what users believe and how they actually behave with AI systems. Evaluations based on proxy tasks and subjective measures fail to predict real decision performance (Bućinca et al., 2020). Human-AI interaction is especially challenging due to uncertainty about AI capabilities and complex outputs (Yang et al., 2020). While established design guidelines emphasize user control and opportunities to understand or correct system behavior (Amershi

et al., 2019), such approaches can fall short when users misconstrue their needs or an AI’s capabilities. In AI writing assistance, presenting multiple suggestions changes composition behavior and perceived usefulness across user groups (Buschek et al., 2021). However, prior work has not manipulated contextual factors nor quantified system consequences. We address these gaps through factorial methodology and computational validation.

3 Methods

Study design. We employed a $2 \times 2 \times 2 \times 2$ factorial vignette design manipulating four binary email dimensions: urgency, compositional effort, sender importance, and email type, yielding 16 scenarios generated with GPT-5 and refined through three author-reviewed rounds to match the 16 intended condition combinations listed in Appendix A.1. The generation prompt required embedding all four dimensions through contextual cues without explicit dimension labels. Fifty participants who regularly handle email in professional contexts completed 15 pairwise comparisons each, randomly sampled from all 120 possible pairs ($\binom{16}{2}$), selecting which scenario they preferred AI drafting assistance for and providing justifications. Sampling 15 rather than all 120 pairs reduced fatigue while yielding 750 decisions that covered all 16 factor combinations in aggregate. The design supports population-level comparisons across the four factors, but not exhaustive within-participant calibration over every factor combination. An exit survey collected rankings of the four dimensions, general AI-use frequency, comfort with AI, email volume, and demographics. See Appendix A.4 for study interface and complete exit survey.

Operational definitions used in scenario construction. To reduce ambiguity, scenarios used explicit cue patterns: *urgency* contrasted deadline language (“by end of day,” “ASAP”) with low-pressure timing (“no rush,” “this week”); *compositional effort* contrasted quick confirmations with multi-step synthesis, coordination, or analysis; *sender importance* contrasted directly consequential stakeholders with lower-impact senders; and *email type* contrasted routine/templated requests with novel or unexpected situations. For example, a high-effort urgent vignette asks for an end-of-quarter report consolidating metrics from every workstream by end of day, whereas a low-effort non-urgent vignette asks for a standard headcount

figure sometime that week.

Stated vs. revealed measures. Exit-survey rankings capture beliefs about individual dimensions, whereas pairwise choices capture preferences among concrete multi-factor scenarios. Because each vignette combines all four dimensions, the comparison tests whether abstract rankings reflect the trade-offs participants make in context. Any gap may reflect imperfect introspection, the different demands of ranking and choice, or both.

Analytical approach. We used Bradley-Terry modeling to derive preference strengths from pairwise comparisons, then calculated Spearman correlations to identify behavioral drivers (RQ1). We validated findings with ordinal logistic regression and two-way ANOVAs. As a robustness check for heterogeneity, we additionally fit a hierarchical random-slopes choice model with participant-specific effects for urgency, effort, sender importance, and email type, and ran exploratory clustering on participant-level inferred weights. For RQ2, we operationalized stated–revealed alignment as the correlation between each participant’s stated and revealed factor rankings, testing predictors via ANOVAs and t-tests. To complement quantitative analysis, we conducted LLM-assisted thematic coding of 750 justifications and 50 survey reflections (Appendix A.2).

For RQ3, we trained four model classes (Logistic Regression, Random Forest, Gradient Boosting, Neural Network) using 5-fold cross-validation with shallow architectures to prevent overfitting. We compared three fixed-weight strategies (Uniform, Stated Preference, Behavioral) to quantify the cost of the stated–revealed gap. These models provide a controlled test of whether the stated–revealed gap affects assistance-rule performance. Complete specifications appear in Appendix A.3.

4 What email attributes drive users’ preferences for AI drafting assistance?

We derive Bradley-Terry preference strengths for all 16 scenarios from 750 pairwise comparisons by 50 participants. Spearman correlations between dimensions and these scenario-level strengths reveal that effort dominates AI drafting preferences ($\rho = 0.597, p < 0.05$), while urgency shows virtually no predictive power ($\rho \approx 0, p = 1.0$). Sender importance ($\rho = 0.244$, favoring important sender) and email type ($\rho = 0.244$, favoring routine over novel email) show weak non-significant trends and

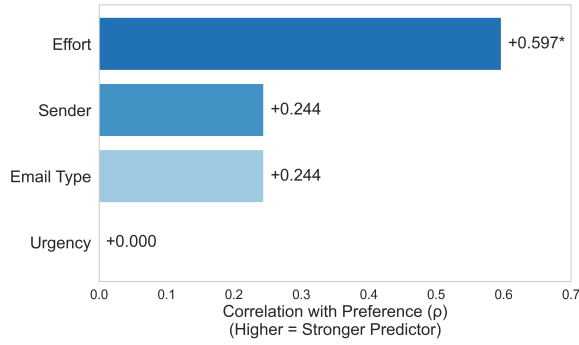


Figure 1: Predictive strength of email dimensions measured by Spearman correlation with Bradley-Terry preference strengths. Effort to compose dominates as the only statistically significant predictor, while urgency shows near-zero correlation. Sender importance and email type show weak non-significant trends. $*p < 0.05$.

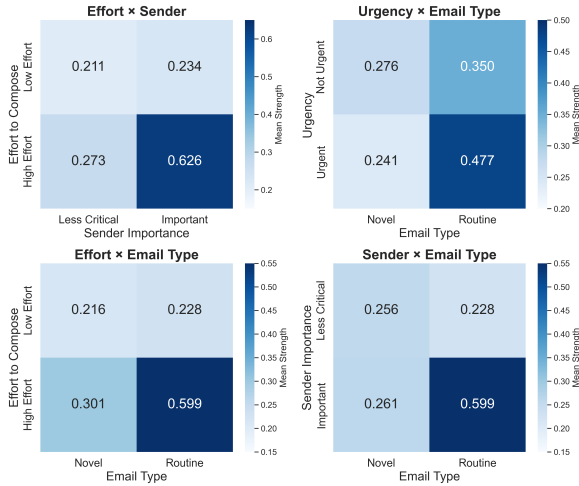


Figure 2: Key pairwise dimension combinations showing mean Bradley-Terry preference strengths. The important-sender, high-effort combination has the highest observed mean preference ($\bar{\pi} = 0.626$), although the interaction test is not significant. Sender $\times$ Type shows a larger descriptive sender difference for routine than novel emails.

play secondary roles (Figure 1). Ordinal regression validates these patterns. High-effort emails show 13.68 times higher odds of selection ($p < 0.05$), while urgency exhibits no significant effect. Both analyses confirm that effort is the main driver and urgency has no reliable effect.

As a heterogeneity robustness check, we fit a hierarchical random-slopes choice model over all 750 decisions with participant-specific effects for urgency, effort, sender importance, and email type. Urgency slopes are mixed, i.e., 70% positive and 30% negative, confirming meaningful individual variation. However, fixed effects remain structured: effort +0.956 with odds ratio 2.60, sender +0.541 with 1.72, type +0.321 with 1.38, and urgency

+0.220 with 1.25. Approximate 95% intervals indicate effort, sender, and type are clearly positive, while urgency is weaker and less certain, i.e., the interval crosses 0. An exploratory clustering analysis of participant-level inferred weights reports best silhouette 0.290 and preserves the same population-level ordering, suggesting the effort-over-urgency pattern is not a single-subgroup artifact.

Compositional effort remains prominent across dimension combinations (Figure 2). High effort combined with important sender produces the highest preference ($\bar{\pi} = 0.626$), a pattern consistent with, but not sufficient to establish, a multiplicative “high-stakes, high-burden” heuristic. Urgency also differs descriptively by email type. Participants select AI help for urgent routine emails ($\bar{\pi} = 0.477$) more often than for urgent novel emails ($\bar{\pi} = 0.241$). While 2-way ANOVA confirms effort’s main effect ($F = 5.294, p < 0.05$), the interaction effects do not reach statistical significance. Random Forest Gini importance assigns 61.7% of total importance to interaction features, but models including those features do not outperform main-effect models (§6.1). Across specifications, urgency remains weak and effort remains the strongest driver.

4.1 Qualitative Analysis of Justifications

Thematic analysis of 750 in-the-moment justifications (avg. 1.03 themes per response) both corroborates and complicates the quantitative patterns. **Urgency and time pressure** was the most frequently mentioned factor (20.3%), with users citing deadlines and time pressure as triggers for AI assistance, e.g., “the time sensitive nature is key to getting a head start with AI” and “due to urgency I’d take the one with a shorter deadline”. This creates a puzzle that users articulate urgency most often yet it shows near-zero behavioral predictive power. **Complexity of drafting** was the second-most frequent theme (14.7%), with users explicitly seeking AI for multi-step, synthesis-heavy tasks and building comprehensive transition plans, including rationale such as “much larger ask and need to pull information from multiple sources” and “This has more moving parts to consolidate to use in a draft.” aligning with effort’s dominant behavioral correlation ($\rho = 0.597$). Users also actively rejected AI for simple yes/no responses (**Task simplicity – doesn’t need AI**, 12.8%), confirming that compositional burden drives preferences.

Co-occurrence analysis reveals sophisticated de-

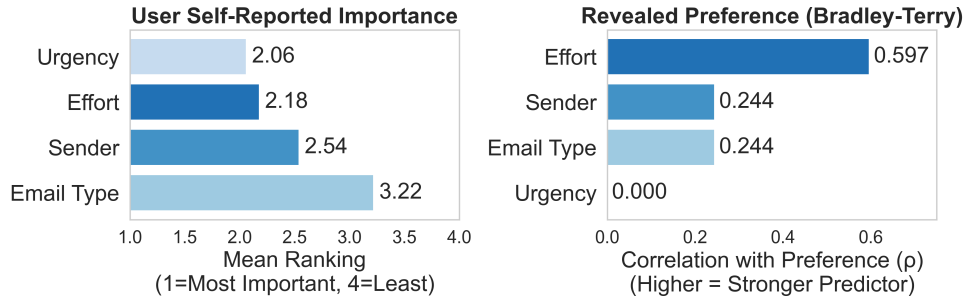


Figure 3: Comparison of stated versus revealed preference rankings. Users assign urgency and effort similar stated importance, with urgency ranked slightly higher than effort (mean ranks: 2.06 vs. 2.18, where lower ranks indicate greater importance). Revealed preferences separate the two factors sharply. Urgency has a near-zero scenario-level correlation ($\rho \approx 0$), whereas effort is the strongest measured predictor ($\rho = 0.597$). In other words, participants rated urgency and effort similarly, but effort was much more strongly associated with their choices than urgency.

cision heuristics that explain these patterns. Users calibrate AI trust based on task predictability: **Template or historical context** (10.1%) frequently co-occurred with perceived AI capability, with users stating “format already existing so human validation will be faster.” This conditional trust pattern suggests templates simultaneously reduce effort *and* increase confidence in AI accuracy. Some users also mentioned **Ease of verification** (1.3%) when justifying AI use, with reasoning like “It’s easy for me to verify correctness.” Here, verification refers to checking the AI-generated draft before sending. Users are more willing to adopt AI when they can quickly review outputs to catch any errors. This reframes the design challenge from maximizing AI accuracy to minimizing verification friction, suggesting features like confidence scores or highlighting AI-generated content could expand adoption among risk-averse users.

One surprising insight is that **Avoiding high-risk or sensitive tasks** was rare (4.8%), suggesting sensitivity concerns do not dominate AI adoption decisions when verification is feasible. Additionally, users view AI as a cognitive partner for **Brainstorming or creative help** (3.9%), not just automation, indicating potential for collaborative features beyond drafting. See Appendix A.2.2 for theme frequencies and additional insights.

5 How Do Stated Factor Rankings Align with Revealed Preferences?

Comparing self-reported rankings with revealed preferences reveals a stated–revealed gap (Figure 3). Users rank urgency as most important (mean rank: 2.06) despite its near-zero behavioral predictive power ($\rho \approx 0$). Effort, the strongest driver ($\rho = 0.597$), receives only moderate self-reported importance (mean rank: 2.18). The paired

Wilcoxon test finds no significant difference between the stated urgency and effort ranks ($p = 0.62$). The revealed choices, however, separate the two factors: effort is the strongest measured predictor ($\rho = 0.597$), whereas urgency has a near-zero scenario-level correlation ($\rho \approx 0$) and no reliable effect across inferential analyses. Thus, the survey rankings treat urgency and effort similarly, but a rule built from them would give too much weight to urgency and too little to effort.

5.1 Qualitative Analysis of Post-Task Reflections

Thematic analysis of 50 survey responses (avg. 2.8 themes per response) corroborates this stated–revealed gap. **Urgency / time sensitivity** appeared in 66% of retrospective accounts, compared with 20.3% of immediate justifications, despite its near-zero scenario-level correlation with vignette choices ($\rho \approx 0$). Participants often used urgency-framed language, e.g., “Time to write the email and urgency to respond.” and “Priority of the deadline on the email”. In contrast, **Effort vs. efficiency** appeared in 32% of reflections despite being the strongest measured predictor of those choices ($\rho = 0.597$). When effort did appear, it was concrete, e.g., “how much chasing people I’d need to do to get the data” and “Amount of data that needs to be consolidated and analyzed”. Participants consistently talked more about urgency, but effort was much more strongly associated with what they chose.

Beyond the urgency-effort divergence, the survey reflections reveal how users conceptualize their decision-making. **Data availability** emerged in 28% of accounts despite not being a manipulated dimension, reflecting how users decompose compositional effort into information-gathering versus

drafting complexity. See Appendix A.2.2 for complete thematic comparison.

5.2 Individual Differences in Stated–Revealed Alignment

We explore whether certain user characteristics show better alignment between stated and revealed preferences. This analysis is exploratory, especially because the low-comfort subgroup is small, i.e., $n=4$ versus $n=40$ in the high-comfort group and $n=50$ total. Within-sample patterns indicate that self-reported comfort with AI technology, measured on a 7-point scale, rather than usage frequency or email volume, is associated with metacognitive alignment. General AI-use frequency and AI comfort capture distinct aspects of familiarity. High-comfort users (scores 5–7) show better alignment between stated and actual preferences ($\rho = 0.090$), while the four low-comfort participants (scores 1–3) show negative alignment ($\rho = -0.550$, $p < 0.05$), an exploratory pattern that requires targeted replication. The descriptive difference also appears in dimension rankings: low-comfort users rank urgency as their top priority (mean rank: 1.00) while placing effort last (3.25), creating a 2.25-rank gap. In contrast, high-comfort users show nearly balanced rankings (urgency: 2.08, effort: 2.05), although their average stated–revealed alignment remains near zero. Using participant-level stated–revealed agreement, mean alignment is -0.712 for the four low-comfort participants and $+0.008$ for 40 high-comfort participants. The estimated $+0.719$ difference is large in this sample and is consistent across exploratory checks (permutation $p = 0.0309$, bootstrap 95% CI $[-0.498, +0.964]$, and Welch’s $t = -5.655$, $p < 0.001$). These converging estimates support the within-sample pattern, but the low-comfort estimate rests on four participants.

Neither general AI-use frequency nor email volume is significantly associated with stated–revealed alignment (see Figure 13 in Appendix A.5).

6 Do ML models predict preferences from behavioral patterns, and what is the cost of the stated–revealed gap?

The behavioral results show that users prefer AI drafting assistance for high-effort emails while urgency receives the highest mean stated rank. We use lightweight predictive models to test whether rules based on revealed behavior outperform rules

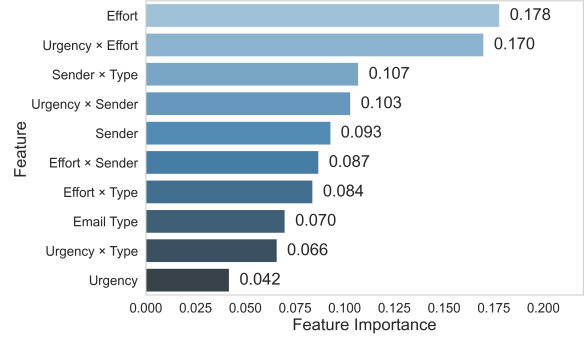


Figure 4: Random Forest Gini importance with main effects and pairwise interactions. Effort is the top main effect (17.8%); interactions collectively receive 61.7% importance but do not improve accuracy over main effects alone.

Model	Full Features (10)		Main Effects (4)	
	Acc.	F1	Acc.	F1
Log. Reg.	0.587	0.657	0.580	0.643
Random Forest	0.608	0.719	0.605	0.714
Grad. Boost	0.588	0.709	0.600	0.713
Neural Net	0.599	0.692	0.603	0.698
Mean	0.595	0.694	0.597	0.692

Table 1: ML performance with main effects and pairwise interactions. Full metrics appear in Appendix A.3.

based on stated preferences using the same four factors.

6.1 Predictive Validation

To test the result predictively, we first train models using only the four main effects. All models achieve about 60% accuracy (Table 1), exceeding the majority-class baseline (55.2%) and chance ($p < 0.001$). Adding all six pairwise interactions provides no predictive benefit: models with 10 features average 59.5% accuracy, compared with 59.7% for main effects alone. Thus, interaction features add little predictive value and do not change the main result that effort predicts choices much more strongly than urgency.

Reliability checks contextualize accuracy under sparse repeated measures. With only 15 comparisons per participant, individual-level coefficients are noisy (mean bootstrap SD=0.401). A hierarchical model estimates 62.5% accuracy from the four features without participant identity and 68.0% when conditioning on participant identity, a 5.5-point difference. Because preferences were measured once and each participant contributed few comparisons, this model-based gap suggests po-

Strategy	Acc.	Prec.	Recall	F1
Uniform	0.589	0.586	0.870	0.700
Stated Preference	0.577	0.568	0.981	0.719
Best ML Model	0.605	0.614	0.855	0.714
Behavioral	0.613	0.608	0.841	0.706

Table 2: Rule-based strategies and the best ML model. Behavioral and ML strategies significantly outperform stated preference in accuracy ($p < 0.05$); F1 differences are not significant.

tential value in personalization but does not establish stable user-specific effects. The best ML model (60.5%) and behavior-derived rule (61.3%) fall within 2.0 and 1.2 points of the feature-only estimate.

Feature importance appears paradoxical: interactions account for 61.7% of Random Forest Gini importance (Figure 4) yet do not improve prediction. The explanation is redundancy in the binary feature space. Effort appears in multiple high-importance interactions, particularly Effort $\times$ Urgency (17.0%), Effort $\times$ Sender (8.7%), and Effort $\times$ Type (8.4%). When effort is high, these interaction terms often become non-zero, counting the same signal multiple times without adding information beyond the main effect. Their combined 61.7% importance therefore does not show that interactions add predictive value.

6.2 System Design Strategy Comparison

We compare three fixed-weight policies against the best ML model (Random Forest with four main effects). Uniform weighting assigns 25% to each factor; stated-preference weighting follows survey ranks (urgency/effort/sender/type = 40/30/20/10%); behavioral weighting follows Bradley-Terry correlations (0/55/22.5/22.5%). The stated-preference strategy reaches only 57.7% accuracy, below the 58.9% uniform baseline. In contrast, behavioral weighting reaches 61.3%, significantly outperforming stated preference by 3.6 points ($p < 0.05$). Its accuracy is close to the best ML model’s 60.5%, and the 0.8-point difference is not significant.

Behavior-derived weighting significantly improves on stated-preference weighting by 3.6 points and achieves the highest observed accuracy among the compared approaches. It therefore provides a simple, interpretable baseline for field evaluation.

7 Discussion

7.1 Why Do Stated and Revealed Preferences Diverge?

The stated–revealed gap raises a question: why does urgency receive the highest mean stated rank while showing little predictive power? We consider two psychological hypotheses. **Social desirability bias** (Paulhus, 1991; Crowne and Marlowe, 1960) may push explicit rankings toward normatively acceptable explanations. Saying “I need help with time-sensitive email” sounds more professional than admitting a desire to avoid cognitively demanding writing. This interpretation aligns with evidence that stated preferences and subjective measures can mislead evaluation (Bućinca et al., 2020; Bertrand and Mullainathan, 2001); in our reflections, users frequently mentioned quality and personalization even though those concerns did not predict choices.

The **availability heuristic** (Tversky and Kahneman, 1973) offers a second explanation: urgent emails are stressful and easy to recall, so users may overemphasize them retrospectively even when urgency did not guide in-task decisions. In contrast, high-effort emails may feel routine despite requiring more support. The rise of urgency mentions from 20.3% in immediate justifications to 66% in retrospective reflections supports this reconstructive account. These hypotheses are plausible, but the study was not designed to distinguish between them. AI comfort is associated with alignment in our exploratory analysis, but the four-person low-comfort subgroup leaves the mechanism uncertain.

Participants ranked individual dimensions but chose between multi-factor scenarios, so the two measures ask different questions. This measurement difference may account for part of the stated–revealed gap. The practical implication is that the rule based on situated choices outperformed the rule based on abstract rankings.

7.2 Design Implications for Proactive AI Email Assistance

Within professional-email vignettes, our findings show the value of validating explicit factor rankings against situated choices before using them to time drafting help. The general interruption principle is to assist when burden is high and disruption is low. In our study, the self-report-derived rule assigns too much weight to urgency and too little to compositional effort, reducing its accuracy relative

to behavior-derived weighting.

Compositional-effort detection over temporal urgency. Effort dominates preference decisions while urgency has no reliable effect, contradicting the intuition to offer help primarily under time pressure. Proactive email assistants should therefore prioritize evaluating compositional-effort cues rather than rely on deadline cues alone. Qualitative analysis shows users seek help for multi-step, synthesis-heavy tasks (14.7% of justifications) and reject AI for simple responses (12.8%). Practical effort proxies include thread length and quoted content, recipient diversity, question/action-item structure, missing-information cues, and interaction traces such as revision bursts, pause patterns, or repeated cursor movement. Mean preferences are highest when high effort and sender importance coincide, although the interaction tests are non-significant. This descriptive pattern makes contextual weighting a concrete hypothesis for future evaluation.

Implicit behavioral feedback over explicit preference elicitation. Settings pages asking users when they want AI help assume introspective access to decision drivers, but abstract rankings did not fully capture situated choices in this controlled email task. Email assistants should therefore validate explicit settings against implicit signals such as suggestion acceptance, dismissal, deliberation time, and overrides. However, implicit learning should remain transparent: surfacing why an offer appears can support accountability and help users calibrate their own preferences. For example, prompts such as “I can draft this using your previous response to [similar email]” may simultaneously signal lower effort and higher verifiability than a generic “Would you like help?” prompt.

Operationalizing effort also reframes evaluation. Rather than treat “help when urgent” as a global rule, systems can predict expected drafting cost and verification cost. This gives developers concrete instrumentation targets: whether the user must synthesize across messages, resolve stakeholder constraints, collect missing data, or personalize tone. These are text- and interaction-level cues that NLG systems can model. A detector for high drafting burden and low verification cost can guide not only whether to intervene, but also which capability to surface: an outline, a template fill, a summary of missing information, or a full response. Such routing remains future work, but it follows directly from the observed gap: asking users for thresholds

is less reliable than learning where they repeatedly accept or reject assistance.

Simple behavioral rules before complex modeling. The behavior-derived rule attains the highest observed accuracy among the compared approaches, and its 3.6-point advantage over the stated-preference rule shows that simple feature weighting is a competitive baseline in this sparse-data setting. An effort-weighted rule is therefore an interpretable and auditable cold-start hypothesis for field evaluation. The exploratory participant-conditioned gap also motivates longitudinal tests of personalization.

8 Conclusion

Proactive professional-email assistants need to infer when users want drafting help, yet their intervention policies may be built from self-reports that have not been behaviorally validated. We provide behavioral evidence that compositional effort is the strongest measured predictor of AI drafting preferences in controlled professional-email scenarios, while urgency has no reliable effect, shifting design priorities from temporal signals to compositional effort. This finding matters for system design because urgency is often easy to detect but weakly diagnostic of when users want drafting support.

We also document a stated–revealed gap with measurable consequences. Urgency receives the highest mean self-reported rank despite its minimal influence, and within this testbed, a behavior-derived rule significantly outperforms a stated-preference rule by 3.6 accuracy points. The gain shows that elicitation choice affects assistance-rule performance in this setting.

Our contributions to proactive professional-email assistance are: (1) providing behavioral grounding for intervention preferences through a factorial vignette design, (2) documenting a gap between abstract factor rankings and situated choices, and (3) quantifying a 3.6-point accuracy gain for behavior-derived over stated-preference weighting, with the behavior-derived rule achieving the highest observed accuracy among the compared approaches. Across controlled professional-email vignettes, compositional effort is the strongest measured predictor of drafting-help choices, while urgency is not reliably associated with those choices. Together, these findings provide a behavior-grounded basis for designing proactive professional-email assistance. Future work should

also test whether revealed-preference summaries help users align explicit settings with situated choices over time.

For practitioners, the practical lesson is to validate self-report-derived intervention policies against behavioral evidence before deployment. A useful field-evaluation protocol is to compare explicit settings with accept and dismiss traces using the same available cues, then test richer content, history, workload, privacy, and draft-quality signals. For NLP researchers, the findings connect proactive assistance to policy learning: deciding when to generate can be as consequential as deciding what to generate.

9 Limitations

Controlled experimental setting and scenario coverage. Our factorial vignette design enables controlled comparisons across systematically varied factor combinations but abstracts from real-world email complexity. Although scenarios were iteratively generated and author-validated across three rounds, latent factors such as tone, politeness, and social nuance may still act as confounds. We did not conduct independent domain-expert validation or participant manipulation checks. We selected urgency, effort, sender importance, and email type because prior work motivates them and they fit a $2 \times 2 \times 2 \times 2$ factorial design. They form a focused rather than exhaustive set. Their binary levels omit continuous variation and factors such as privacy, emotional tone, factual risk, recipient relationship, message sensitivity, current workload, organizational norms, message stakes, and generated-draft quality. A key next step is replacing binary factors with multi-level or continuous proxies for urgency and compositional effort learned from context.

Elicitation and pair sampling. Stated preferences were rankings of individual dimensions. We did not elicit stated beliefs about interactions. Pairwise choices may therefore reveal concrete trade-offs that participants were never asked to articulate abstractly. Randomly sampling 15 of 120 possible pairs per participant reduced burden and yielded aggregate coverage of all 16 factor combinations. This supports population-level factor contrasts, but not exhaustive within-participant calibration over every combination.

Modeling scope. The binary-feature classifiers and fixed-weight rules provide a controlled test of how policy construction changes agreement with observed choices. Their approximately 60% accuracy is not evidence of deployment readiness. Because cross-validation splits comparisons rather than participants, the reported ML accuracy describes prediction within the sampled participant pool, not generalization to unseen users. Field systems require richer contextual signals and prospective evaluation.

Sample characteristics. Participants were 50 knowledge workers making frequent email decisions, appropriate for the target population of proactive drafting assistants. We collected role, general AI-use frequency, AI comfort, and email volume, but did not separately measure experience with AI

writing assistants. These constructs may differently shape scenario interpretation and trust in generated drafts. High AI comfort was more common than low AI comfort in our sample (40 high-comfort and four low-comfort participants), so findings about low-comfort users are exploratory and require targeted recruitment. Generalization to populations with systematically different email practices (e.g., customer service roles with heavily templated responses, executives whose correspondence is managed by assistants, non-English communication contexts) also requires empirical validation. The stated–revealed gap appears across general AI-use frequency and email volume within our sample, but cross-cultural and organizational variation warrants investigation.

Task generalizability. Email drafting involves visible compositional effort (writing, editing, structuring arguments), making effort-to-compose a naturally salient dimension. For different NLG tasks (e.g., summarization requests or query reformulation), effort may manifest differently (comprehension load vs. generation load).

Preference stability. We measured preferences at a single timepoint. Longitudinal studies would clarify whether preferences evolve with AI adoption, whether early versus experienced users show different factor weightings, and whether revealing the stated–revealed gap through usage feedback alters subsequent preferences.

References

- Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when? the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 271–278.
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, and 1 others. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, and Ian Smith. 2003. Taking email to task: The design and evaluation of a task management centered email tool. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 345–352.
- Marianne Bertrand and Sendhil Mullainathan. 2001. Do people mean what they say? implications for subjective survey data. *American Economic Review*, 91(2):67–72.

- Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464. Best Paper Award.
- Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, and 1 others. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2287–2295.
- Douglas P Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349–354.
- Edward Cutrell, Mary Czerwinski, and Eric Horvitz. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Proceedings of INTERACT*, volume 1, pages 263–269.
- Edward B Cutrell, Mary Czerwinski, and Eric Horvitz. 2000. Effects of instant messaging interruptions on computing tasks. In *CHI'00 extended abstracts on Human factors in computing systems*, pages 99–100.
- Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 175–182. ACM.
- Laura A Dabbish and Robert E Kraut. 2005. Why do i keep interrupting myself? environment, habit and self-interruption. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–11.
- Nicolas Ducheneaut and Victoria Bellotti. 2001. E-mail as habitat: an exploration of embedded personal information management. *interactions*, 8(5):30–38.
- Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 159–166.
- Shamsi T Iqbal and Brian P Bailey. 2008. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 93–102.
- Shamsi T Iqbal and Eric Horvitz. 2010. Notifications and awareness: a field study of alert usage and preferences. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 27–30.
- Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and 1 others. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 955–964.
- Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Wendy E Mackay. 1988. Diversity in the use of electronic mail: A preliminary inquiry. In *ACM Transactions on Information Systems*, volume 6, pages 380–397.
- Daniel C McFarlane. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-computer interaction*, 17(1):63–139.
- Delroy L Paulhus. 1991. Measurement and control of response bias. *Measures of Personality and Social Psychological Attitudes*, 1:17–59.
- Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52.
- Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232.
- Joshua R Tyler and John C Tang. 2003. When can i expect an email response? a study of rhythms in email usage. In *ECSCW 2003: Proceedings of the Eighth European Conference on Computer Supported Cooperative Work 14–18 September 2003, Helsinki, Finland*, pages 239–258. Springer.
- Steve Whittaker and Candace Sidner. 1996. Email overload: Exploring personal information management of email. *Proceedings of the 1996 CHI Conference on Human Factors in Computing Systems*, pages 276–283.
- Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

A Appendix

A.1 Email Scenario Descriptions and Generation Prompt

A.1.1 Scenario generation prompt

System Prompt:

You are a helpful assistant that generates realistic work scenarios. Create natural, contextual scenarios that subtly embed the required dimensions without explicitly labeling them. Participants should be able to infer urgency, effort level, sender importance, and routine/novel nature from context clues rather than direct statements.

Important: Write all scenarios in narrative form without using quotation marks or direct dialogue. Naturally mention how each sender relates to the user's project management role to establish their importance level. Avoid explicit phrases like "this is urgent", "high effort", "routine task", or "novel situation". Let the context convey these qualities. Focus on realistic workplace situations that feel authentic and allow participants to make their own judgments about draft helpfulness.

User Prompt:

Role-Play Profile: Project Manager at a Big Tech Company

You are a project manager at a large technology company, responsible for driving cross-functional initiatives that impact multiple product teams. You have been in this role for three years and currently oversee several projects at different stages of execution. Your work involves coordinating with engineering, design, legal, and communications teams to ensure timely delivery of features and organizational changes.

Your responsibilities include creating and maintaining detailed project plans, managing risks and dependencies, and preparing executive updates. You frequently draft communications for stakeholders, such as status reports, decision summaries, and change announcements. Approvals often require collaboration with legal and compliance teams, and you must ensure all communications adhere to company policies. You also facilitate workshops, track progress against milestones, and escalate issues when necessary.

Communication is a critical part of your role. For executives, you provide concise summaries that highlight decisions, risks, and next steps. For operational teams, you share structured updates with clear timelines and ownership details. You primarily use email and Teams for communication,

and you are expected to respond promptly to urgent requests while balancing long-term deliverables.

Task: Generate scenarios for all 16 unique combinations of these dimensions: Urgency (Urgent / Not Urgent), Effort to Compose (High Effort / Low Effort), Sender Importance (Important Sender / Less Critical Sender), Request Type (Routine / Novel).

Guidelines: (1) Convey urgency naturally through context like "by end of day", "ASAP", "needed for tomorrow's meeting" versus "when you get a chance", "no rush", "sometime this week". (2) Show effort level through task complexity rather than stating it explicitly. Lower complexity includes quick confirmations and brief status updates; higher complexity involves multi-team coordination and comprehensive analysis. (3) Establish sender importance through relationship clarity by always specifying how the sender relates to your role. Direct impact senders include your manager, executives over your projects, and key stakeholders; indirect impact senders include peer managers from other teams and administrative staff. (4) Indicate whether requests are typical or unusual through context like "weekly report" versus "unexpected situation" or "new initiative". (5) Write naturally without explicit dimension labels. (6) Make scenarios realistic and specific with authentic context that would occur in a tech company. (7) Ensure each scenario requires a meaningful response. (8) Write scenarios without quotation marks in narrative form.

Scenarios were developed through three iterative rounds: initial generation, independent author review against intended dimension levels, and targeted prompt/scenario refinement for cases with ambiguous cues. After the third round, the authors agreed that all 16 scenarios matched the factorial specifications for urgency, compositional effort, sender importance, and email type.

A.1.2 Complete scenario list

- 1. Urgent + High Effort + Important Sender + Routine:** Your manager reaches out asking for an updated end-to-quarter project health report that consolidates metrics from every active workstream. They need it finalized by end of day for a leadership review later tonight, and the template is the same one your team uses every quarter.
- 2. Urgent + High Effort + Important Sender +**

- Novel:** The VP of Engineering, who oversees the team building your core product, informs you that the executive board wants an impact analysis immediately on how yesterday's regulatory announcement will affect your current launch plan. They request a detailed scenario assessment and mitigation options by tomorrow morning.
3. **Urgent + High Effort + Less Critical Sender + Routine:** A peer project manager from the HR technology team asks you this morning to produce a comprehensive dependency summary for shared resources. They explain it's blocking their integration test kickoff scheduled for later today.
 4. **Urgent + High Effort + Less Critical Sender + Novel:** The facilities operations coordinator contacts you saying a major system outage has affected some of the collaboration tools used by your project teams. They ask you to help compile an immediate contingency plan with alternative workflows to share with impacted groups before the next working session.
 5. **Urgent + Low Effort + Important Sender + Routine:** Your director pings you requesting the latest timeline slide for an executive call that starts in an hour. They mention the earlier version is fine, but just needs the updated completion percentages inserted.
 6. **Urgent + Low Effort + Important Sender + Novel:** A senior product lead messages you asking for a quick confirmation on whether your team can accommodate an unplanned customer feedback session tomorrow. They need a response now so they can commit to the client during an ongoing call.
 7. **Urgent + Low Effort + Less Critical Sender + Routine:** An analyst from another department emails asking you to resend yesterday's stakeholder deck because they can't access it in the shared folder. They need it now to join a call that starts in a few minutes.
 8. **Urgent + Low Effort + Less Critical Sender + Novel:** A facilities admin contacts you during a building-wide power fluctuation, wanting to know if any hardware dependencies in your programs need immediate communication to engineering. They ask for a quick yes or no so they can finalize their outage alert message.
 9. **Not Urgent + High Effort + Important Sender + Routine:** Your director asks you to prepare the full quarterly portfolio summary for their upcoming strategy review next week. It's the same in-depth report you've created in past quarters, combining KPIs, risk logs, and milestone charts.
 10. **Not Urgent + High Effort + Important Sender + Novel:** The Chief Legal Officer requests that you draft a comprehensive transition plan outlining how upcoming compliance rules will be integrated into all your product streams. They emphasize that this initiative is new to the organization and due end of next month, so you have time to propose a structured approach.
 11. **Not Urgent + High Effort + Less Critical Sender + Routine:** A peer PM from an internal tools project asks if you can produce a consolidated view of shared resource usage between your project and theirs. They note it isn't blocking any current work and suggest having something ready in a few weeks.
 12. **Not Urgent + High Effort + Less Critical Sender + Novel:** The communications specialist for the employee engagement program reaches out for help designing a cross-team workshop to brainstorm ways of showcasing successful product launches internally. They mention it's a new initiative and would appreciate speaking with you later this month.
 13. **Not Urgent + Low Effort + Important Sender + Routine:** Your manager emails asking you to confirm whether last sprint's milestone progress is already reflected in the dashboard. They say there's no rush and to let them know anytime this week.
 14. **Not Urgent + Low Effort + Important Sender + Novel:** A senior engineering director asks if you can provide a quick paragraph describing your project's long-term architecture considerations for a vision document they're drafting. They mention having it next week is fine.

15. **Not Urgent + Low Effort + Less Critical**

Sender + Routine: An administrative coordinator pings you for the standard monthly headcount figure for your project team. They clarify that they're compiling next month's seating plan and you can update them whenever convenient this week.

16. **Not Urgent + Low Effort + Less Critical**

Sender + Novel: A colleague working on an internal culture committee explains they're collecting brief success stories from various programs for a new company blog. They ask for a short summary of a recent milestone your project achieved whenever you have a moment in the coming days.

A.2 Thematic Analysis Methodology and Secondary Findings

We conducted exploratory thematic analysis to complement the quantitative preference modeling. This analysis examined textual responses for two research questions. The first research question investigated what email attributes drive users' preferences for AI drafting assistance, analyzing 750 in-the-moment justifications provided during pairwise comparisons. The second research question examined whether users accurately perceive what drives their preferences, analyzing 50 post-task survey reflections where participants retrospectively described their decision-making processes.

A.2.1 Five-phase LLM-assisted coding process

We employed a systematic five-phase approach adapted from traditional qualitative methods but operationalized through LLM assistance to enable broader coverage while maintaining analytical rigor.

Phase 1: Initial coding. The LLM analyzed a representative sample using inductive coding with open-ended prompts designed to let themes emerge naturally. For the first research question on preference drivers, we sampled 150 of 750 justifications to balance coverage and manageability, representing 20% of all pairwise comparisons. For the second research question on metacognitive accuracy, we analyzed all 50 survey responses given the smaller dataset. The prompt explicitly instructed the model to let themes emerge from the responses themselves without forcing responses into predefined categories. If the model noticed themes that

did not fit common patterns, it was instructed to include them as distinct codes. Temperature was set to 0.3 for consistent coding while allowing theme discovery.

Phase 2: Codebook development. Initial themes were consolidated into a structured codebook with hierarchical organization, clear definitions, inclusion criteria, exclusion criteria, and example quotes for each code. This phase used temperature 0.3 to balance systematicity with synthesis.

Phase 3: Systematic coding. The codebook was applied to all responses using very low temperature of 0.1 to ensure maximum consistency. Responses for the first research question were processed in batches of 50 to accommodate the richer pairwise comparison context where each response included two email scenarios, the preference choice, and the justification text. All 50 responses for the second research question were coded in one batch given the simpler data structure with only open-ended text.

Phase 4: Parsing and structuring. Coded results were parsed, merged with original responses, and analyzed for theme frequencies, co-occurrence patterns, and distribution statistics.

Phase 5: Synthesizing findings. The authors synthesized findings using LLM-generated summaries of theme distributions, sample quotes, and codebook patterns. For the second research question specifically, prompts emphasized stated-revealed gap analysis by asking about factors users over-report, under-report, or misattribute when describing their decision-making. The LLM provided organized summaries at temperature 0.4 for moderate analytical creativity while staying grounded in data, which the authors then interpreted to identify key patterns and implications.

All phases used a fixed random seed of 42 for reproducibility. The authors reviewed and confirmed the coding process and outputs at each phase to ensure that results and insights were valid and aligned with the data. Complete codebooks with hierarchical theme structures, coded datasets with theme assignments, and detailed frequency distributions are presented in this appendix (see Appendix A.2.2 below).

A.2.2 Complete Thematic Findings

RQ1: In-the-moment justifications during pairwise comparisons

We analyzed 750 justifications provided imme-

diately after participants made their preference choices during the pairwise comparison task. We identified 12 distinct themes organized into four main categories following the hierarchical code-book structure developed in Phase 2. Total justifications coded: 750. Total theme assignments: 769. Average themes per justification: 1.03. Median: 1. Note: Percentages below represent the proportion of responses mentioning each theme. They sum to 102.5% (rather than 100%) because some responses mention multiple themes simultaneously.

Complete theme inventory with hierarchical structure

Main Theme 1: Time and efficiency drivers (31.5% combined)

- **Urgency and time pressure** (20.3%) cited tight deadlines or urgent turnaround as reason for AI use
- **Desire to save time** (11.2%) emphasized efficiency or reducing effort without explicit urgency

Main Theme 2: Task characteristics (40.7% combined)

- **Complexity of drafting** (14.7%) indicated task is complex, multi-step, or requires synthesis
- **Task simplicity – doesn't need AI** (12.8%) determined task is too simple for AI to add value
- **Template or historical context** (10.1%) noted existing format, template, or prior example that AI can leverage
- **Rote or tedious work** (3.1%) identified repetitive, low-value, or boring tasks

Main Theme 3: Risk and trust considerations (6.1% combined)

- **Avoiding high-risk or sensitive tasks** (4.8%) avoided AI for sensitive, high-stakes, or regulatory content
- **Ease of verification** (1.3%) indicated ability to easily check correctness, reducing risk

Main Theme 4: Perceptions of AI role and capability (24.3% combined)

- **Perceived AI capability** (14.3%) believed AI can perform well due to clear requirements or structured tasks
- **Alignment with AI strengths** (4.5%) identified task matches AI's known strengths like summarization or structuring
- **Brainstorming or creative help** (3.9%) saw AI as helpful for generating ideas or creative input
- **Personal writing preference** (1.6%) chose AI due to personal discomfort or dislike of writing tasks

Co-occurrence patterns suggest contextual trust. Urgency and time pressure frequently co-occurred with complexity of drafting, suggesting that participants sometimes considered time pressure and cognitive load together. Notably, perceived AI capability clustered with template or historical context, suggesting that task predictability may shape trust in AI. When prior examples existed, participants justified AI use with phrases like "AI seems like it would be able to accurately write this" and "format already existing so human validation will be faster." This conditional-trust pattern motivates testing template availability as a trust signal in personalized AI recommendations.

Verification as a potential risk-mitigation strategy. Ease of verification appeared in a small subset of responses (1.3%), sometimes alongside concerns about high-risk tasks. This suggests that some participants assessed whether verification was feasible rather than uniformly avoiding AI for higher-stakes tasks. Example quotes include "It's easy for me to verify correctness" and "This is a place where AI can quickly get things in the right format and it's easy for me to verify." This pattern motivates testing whether streamlined verification workflows affect willingness to use AI for such tasks.

Template and historical context as possible dual cues. Template or historical context appeared in 10.1% of justifications. Participants linked templates both to reducing perceived effort, as in "uses the same report I've used before, and can save me time," and to greater confidence in reviewing predictable outputs. This pattern suggests that templates may reduce perceived effort while increasing confidence in AI assistance, a hypothesis for future intervention studies.

Personal writing discomfort as niche factor.

Personal writing preference (1.6%) emerged as a distinct but infrequent theme, with users mentioning concerns like “My style of writing is a little less formal” and tone calibration. While rare, this suggests a user segment that seeks AI not for efficiency but for stylistic support. This niche use case differs from the dominant effort-reduction pattern and may benefit from different feature prioritization (tone suggestions vs. quick drafts).

Secondary design implications. Beyond the primary implications discussed in the main paper, these secondary findings suggest: (1) Template-based triggers should be emphasized in UI as both trust and efficiency signals. (2) Verification dashboards showing confidence scores and highlighting AI-generated content could expand adoption among risk-averse users. (3) Compositional-effort detection (typing pauses, revision patterns) may be more predictive than self-reported urgency. (4) Segmented features for niche users (tone adjustment, brainstorming mode) complement core efficiency features.

RQ2: Post-task survey reflections on decision-making

We analyzed 50 survey responses to the open-ended question asking what factors participants considered when deciding whether an email is worth drafting. We identified 13 distinct themes (excluding “None” responses). Total responses coded: 50. Average themes per response: 2.82. Median: 3. Note: Percentages below represent the proportion of responses mentioning each theme. They sum to more than 100% (282%) because each response typically mentions multiple themes.

Complete theme inventory

- **Urgency / time sensitivity** (66%) referenced deadlines, time-critical nature, or need for quick turnaround
- **Effort vs. efficiency** (32%) considered effort required to draft manually versus using AI, including time savings
- **Data availability and inputs** (28%) assessed whether necessary information or data is readily available
- **Complexity of content** (26%) evaluated how complicated or technical the subject matter is
- **Audience / stakeholder importance** (24%) considered who the email is for, including

seniority and role

- **Task type or nature** (22%) distinguished inherent nature of request like informational versus action-oriented
- **Fit for AI / AI capability** (20%) assessed whether task aligns with AI’s strengths or AI can handle complexity
- **Importance / impact of email** (20%) evaluated stakes, visibility, or potential consequences
- **Length and detail of email** (18%) considered expected size and amount of content required
- **Accuracy and quality requirements** (12%) weighed need for correctness, reliability, and professional tone
- **Need for personalization or nuance** (8%) assessed need for personal touch or emotional tone
- **Prior experience / familiarity** (6%) drew on past experience with similar emails or AI performance

Retrospective elaboration and possible reconstruction. The average number of themes per response increased from 1.03 in immediate justifications to 2.82 in retrospective accounts. This pattern may reflect reconstructive memory, greater opportunity for elaboration in the retrospective prompt, or both. The consistency of urgency (66%) and effort (32%) as prominent retrospective themes, despite variation in vignette choices, is compatible with shared cultural scripts about “good decision-making.” The study cannot distinguish that account from genuine introspection.

Data availability as a candidate gating factor.

Data availability and inputs appeared in 28% of retrospective accounts but were absent from the four-dimension experimental design. This theme suggests that some participants viewed information completeness as a prerequisite, with comments such as “how much chasing people I’d need to do to get the data” and “Amount of data that needs to be consolidated.” This pattern is consistent with, but does not establish, a two-stage process of assessing feasibility and then desirability. Future work should test whether surfacing data-completeness signals improves assistance decisions.

Limited evidence of metacognitive learning about AI preferences. Despite completing 15 pairwise comparisons with AI drafting scenarios, participants rarely referenced prior experience or familiarity in their reflections (6%). Participants also seldom described learning from the task or updating beliefs about AI capability. Because beliefs were not measured before and after the comparisons, the study cannot determine whether or how quickly metacognitive learning occurs. This pattern motivates longitudinal tests of feedback that shows users their revealed preferences.

Quality concerns and professional identity. Accuracy and quality requirements (12%) and personalization or nuance (8%) appeared in retrospective reflections despite not being experimentally manipulated dimensions. Participants emphasized “reliability of accuracy,” “quality standards,” and a “personal touch.” Because these factors were not varied, the study cannot determine whether they influenced vignette choices, reflected retrospective professional concerns, or both. Their presence identifies additional dimensions for future study.

Task-type taxonomy in participant reflections. Task type or nature appeared in 22% of reflections, including distinctions such as “informational vs. action-oriented,” “repackaging existing content,” and whether a request called for a written response or another action. This richer taxonomy suggests that the four-factor model captures important but non-exhaustive dimensions. Additional task distinctions may matter for particular workflows and should be tested as predictors of assistance preferences.

Secondary design implications. These secondary findings suggest: (1) Data completeness indicators should be visible before drafting (“You have all needed information” or “Missing: Q3 budget figures”). (2) Usage pattern dashboards could accelerate metacognitive calibration (“Your AI requests: 78% high-effort emails, 12% urgent emails”). (3) Quality assurance features (tone matching, fact-checking) address post-hoc professional concerns even if not behavioral drivers. (4) Task type detection could route emails to specialized AI capabilities (informational summary vs. action proposal generation).

A.3 ML Model Implementation Details

This section provides complete implementation details for the ML models used in §6.

A.3.1 Data preparation

Feature construction. Each of the 750 pairwise comparisons was converted into a binary classification instance. For each comparison between scenarios i and j , we created features based on the difference in factor levels: $\Delta\text{Urgency}$, ΔEffort , ΔSender , $\Delta\text{Type} \in \{-1, 0, +1\}$. We also included all six two-way interaction terms: $\Delta\text{Urgency} \times \Delta\text{Effort}$, $\Delta\text{Urgency} \times \Delta\text{Sender}$, $\Delta\text{Urgency} \times \Delta\text{Type}$, $\Delta\text{Effort} \times \Delta\text{Sender}$, $\Delta\text{Effort} \times \Delta\text{Type}$, and $\Delta\text{Sender} \times \Delta\text{Type}$. The target variable was binary: 1 if scenario i was preferred, 0 if scenario j was preferred. This resulted in 10 features per instance.

Preprocessing. All features were standardized using StandardScaler (zero mean, unit variance) to ensure comparable scales across models. No dimensionality reduction was applied given the already compact feature set.

Train-test split. We used stratified 5-fold cross-validation at the comparison level to balance class distributions across folds. This evaluates prediction on held-out comparisons drawn from the sampled participant pool, not generalization to unseen participants. Random seed was set to 42 for reproducibility.

A.3.2 Model architectures and hyperparameters

Logistic Regression. Standard L2-regularized logistic regression with regularization strength $C = 1.0$. Solver: lbfgs. Maximum iterations: 1000. Class weights: balanced to account for slight class imbalance (55.2% prefer assistance).

Random Forest. Ensemble of decision trees with the following hyperparameters tuned via grid search: number of estimators = 100, max depth = 3, min samples split = 20, min samples leaf = 10. These shallow tree constraints were chosen to prevent overfitting given modest dataset size (750 instances). Feature importance was extracted using Gini importance averaged across all trees.

Gradient Boosting. Gradient-boosted decision trees with: learning rate = 0.1, number of estimators = 100, max depth = 3, min samples split = 20, min samples leaf = 10, subsample = 0.8. Similar depth constraints as Random Forest for overfitting prevention.

Neural Network. Shallow feedforward architecture to prevent overfitting:

- Input layer: 10 features (4 main effects + 6 interactions)

- Hidden layer 1: 16 units, ReLU activation
- Hidden layer 2: 8 units, ReLU activation
- Output layer: 1 unit, sigmoid activation

Training configuration:

- Optimizer: Adam with learning rate = 0.001
- Loss function: Binary cross-entropy
- Batch size: 32
- Epochs: 50 with early stopping (patience = 10, monitoring validation loss)
- Dropout: 0.3 after each hidden layer for regularization
- L2 regularization: weight decay = 0.01

A.3.3 Rationale for shallow architectures

With only 750 training instances and 10 features, we prioritized generalization over model complexity. Initial experiments with deeper networks (3-4 hidden layers with 32-64 units) showed severe overfitting: training accuracy reached 75-80% while validation accuracy plateaued at 58-60%. Shallow architectures (2 hidden layers with 16 and 8 units) provided better bias-variance tradeoff, achieving 61-62% validation accuracy with minimal train-validation gap (3-4 percentage points). Similarly, tree-based models with max depth > 5 showed overfitting, justifying the depth = 3 constraint.

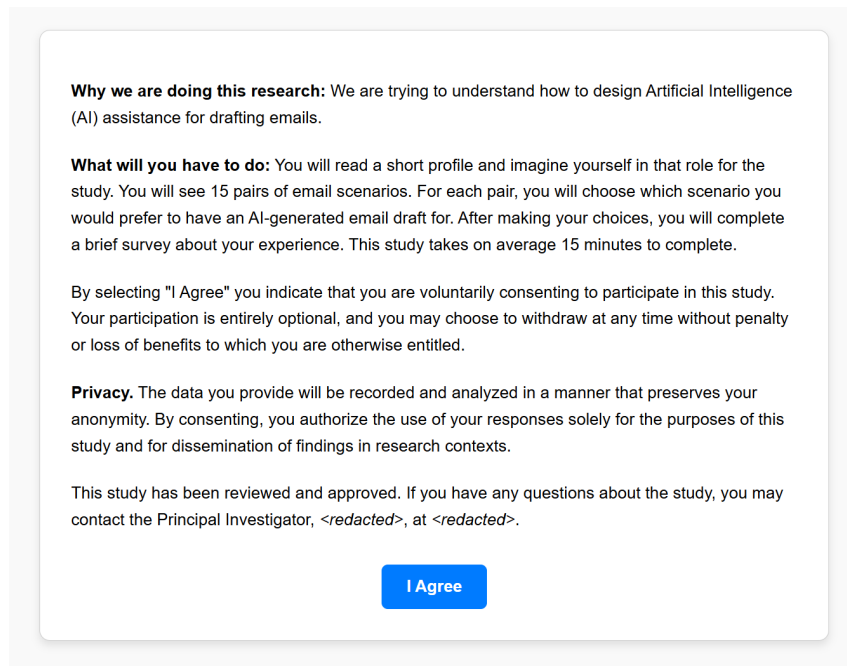
This design choice aligns with our empirical comparison: the simple behavioral rule achieved 61.3% accuracy, while the tested ML models averaged 60.8%. These results identify the behavioral rule as a competitive baseline in this dataset.

A.3.4 Performance evaluation

Model	Full Features (10)				Main Effects (4)			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Logistic Regression	0.587	0.627	0.691	0.657	0.580	0.624	0.664	0.643
Random Forest	0.608	0.607	0.882	0.719	0.605	0.614	0.855	0.714
Gradient Boosting	0.588	0.617	0.833	0.709	0.600	0.615	0.848	0.713
Neural Network	0.599	0.619	0.785	0.692	0.603	0.608	0.819	0.698
Mean	0.595	0.618	0.798	0.694	0.597	0.615	0.797	0.692

Table 3: Complete ML model performance metrics. Main paper (Table 1) shows accuracy and F1 only for brevity.

A.4 Study Interface Screenshots and Exit Survey

The screenshot shows a white rectangular box with rounded corners, centered on a light gray background. Inside the box, there is text explaining the research purpose, what participants will do, the voluntary nature of participation, privacy details, and a statement of approval. At the bottom right of the text area is a blue button with white text that says "I Agree".

Why we are doing this research: We are trying to understand how to design Artificial Intelligence (AI) assistance for drafting emails.

What will you have to do: You will read a short profile and imagine yourself in that role for the study. You will see 15 pairs of email scenarios. For each pair, you will choose which scenario you would prefer to have an AI-generated email draft for. After making your choices, you will complete a brief survey about your experience. This study takes on average 15 minutes to complete.

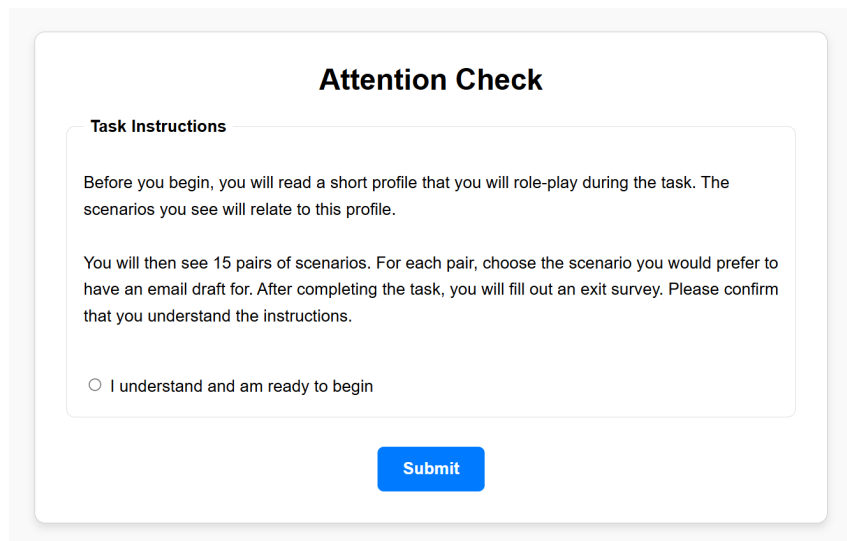
By selecting "I Agree" you indicate that you are voluntarily consenting to participate in this study. Your participation is entirely optional, and you may choose to withdraw at any time without penalty or loss of benefits to which you are otherwise entitled.

Privacy. The data you provide will be recorded and analyzed in a manner that preserves your anonymity. By consenting, you authorize the use of your responses solely for the purposes of this study and for dissemination of findings in research contexts.

This study has been reviewed and approved. If you have any questions about the study, you may contact the Principal Investigator, <redacted>, at <redacted>.

I Agree

Figure 5: Informed consent page presented at the beginning of the study. Participants reviewed study purpose, procedures, risks, benefits, and data handling before providing consent to participate.

The screenshot shows a white rectangular box with rounded corners, centered on a light gray background. At the top, the title "Attention Check" is centered in bold. Below it, the section "Task Instructions" is followed by two paragraphs of text explaining the task. At the bottom left, there is a radio button next to the text "I understand and am ready to begin". At the bottom right is a blue button with white text that says "Submit".

Attention Check

Task Instructions

Before you begin, you will read a short profile that you will role-play during the task. The scenarios you see will relate to this profile.

You will then see 15 pairs of scenarios. For each pair, choose the scenario you would prefer to have an email draft for. After completing the task, you will fill out an exit survey. Please confirm that you understand the instructions.

☐ I understand and am ready to begin

Submit

Figure 6: Task instructions screen explaining the pairwise comparison task. Participants were informed they would see pairs of email scenarios and select which email they would prefer AI drafting assistance for, then provide justification for their choice.

Your Role

Please review your role profile below. You will be making decisions from this person's perspective.

Role-Play Profile: Project Manager at a Big Tech Company

You are a project manager at a large technology company, responsible for driving cross-functional initiatives that impact multiple product teams. You have been in this role for three years and currently oversee several projects at different stages of execution. Your work involves coordinating with engineering, design, legal, and communications teams to ensure timely delivery of features and organizational changes.

Your responsibilities include creating and maintaining detailed project plans, managing risks and dependencies, and preparing executive updates. You frequently draft communications for stakeholders, such as status reports, decision summaries, and change announcements. Approvals often require collaboration with legal and compliance teams, and you must ensure all communications adhere to company policies. You also facilitate workshops, track progress against milestones, and escalate issues when necessary.

Communication is a critical part of your role. For executives, you provide concise summaries that highlight decisions, risks, and next steps. For operational teams, you share structured updates with clear timelines and ownership details. You primarily use email and Teams for communication, and you are expected to respond promptly to urgent requests while balancing long-term deliverables.

Continue to Task

Figure 7: Role-play profile assignment. Participants were instructed to adopt the role of a project manager at a large technology company to ground their judgments in a consistent professional context across all scenarios.

1 / 15
7.0%

View Your Role Profile

Which scenario would you prefer an AI draft for?

Scenario A

The VP of Engineering, who oversees the team building your core product, informs you that the executive board wants an impact analysis immediately on how yesterday's regulatory announcement will affect your current launch plan. They request a detailed scenario assessment and mitigation options by tomorrow morning.

☐ I would prefer to have an AI draft for this scenario

Scenario B

The Chief Legal Officer requests that you draft a comprehensive transition plan outlining how upcoming compliance rules will be integrated into all your product streams. They emphasize that this initiative is new to the organization and due end of next month, so you have time to propose a structured approach.

☐ I would prefer to have an AI draft for this scenario

Please explain why you chose this scenario:

Explain your reasoning for selecting one scenario over the other...

Next

Figure 8: Pairwise comparison interface showing the main task. Participants viewed two email scenarios side-by-side, selected which email they would prefer AI drafting assistance for, and provided free-text justification for their choice. Each participant completed 15 such comparisons.

Post-Task Survey

Please answer the following questions about your experience. Use the appropriate input for each question (text, dropdown, scale, or ranking).

Factors importance

What factors did you consider when deciding whether an email is worth drafting? *

factor1, factor2, factor3, ...

Email volume

How many emails do you typically read or respond to in a day? *

Select one: *

Select an option

AI usage

How often do you use AI to compose, edit, and draft emails? *

Select one: *

Select an option

Perception with others

I feel concerned about how my coworkers might perceive me if they knew I used AI to help write my emails. *

☐ 1
Strongly Disagree

☐ 2
Disagree

☐ 3
Slightly Disagree

☐ 4
Neutral

☐ 5
Slightly Agree

☐ 6
Agree

☐ 7
Strongly Agree

Comfort with AI

I feel comfortable with an AI drafting emails for me. *

☐ 1
Strongly Disagree

☐ 2
Disagree

☐ 3
Slightly Disagree

☐ 4
Neutral

☐ 5
Slightly Agree

☐ 6
Agree

☐ 7
Strongly Agree

Figure 9: Exit survey (screenshot part 1) collecting open-ended reflections on decision factors, email volume, AI usage frequency, and AI attitudes. The open-ended question asked participants to describe what factors they considered when deciding whether an email is worth drafting.

Confidence in writing

I feel more confident using AI to draft emails than writing them myself. *

1

2

3

4

5

6

7

Strongly Disagree

Disagree

Slightly Disagree

Neutral

Slightly Agree

Agree

Strongly Agree

I prefer to write emails with AI because it saves me time and effort. *

1

2

3

4

5

6

7

Strongly Disagree

Disagree

Slightly Disagree

Neutral

Slightly Agree

Agree

Strongly Agree

Ranking dimensions

Rank the four dimensions of how important you think is in terms of an email worth drafting. *

1. Urgency

2. Effort to compose

3. Sender importance

4. Routine vs. Novel

Enter a unique rank (1 = most important, 4 = least important) for each dimension. *

Background

What best describes your role? *

Select one:*

Select an option

General feedback

Any other thoughts or feedback about email drafting or this study?

Figure 10: Exit survey (screenshot part 2) collecting forced rankings of the four dimensions (Urgency, Effort to Compose, Sender Importance, Email Type) from most to least important in participants' decision-making, along with current role and optional feedback.

A.5 Individual Differences in Preference Patterns

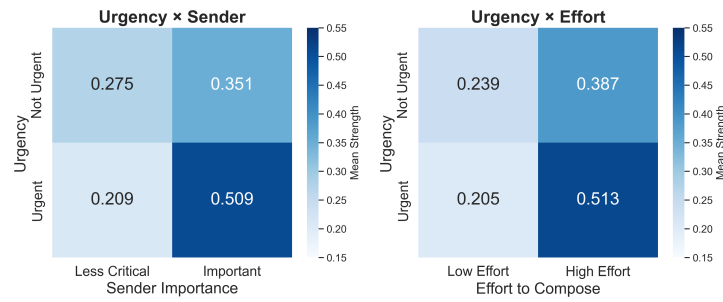


Figure 11: Remaining pairwise dimension interactions showing mean Bradley-Terry preference strengths. Urgency $\times$ Sender and Urgency $\times$ Effort show relatively flat patterns, providing no evidence that the observed interactions overturn urgency's near-zero scenario-level correlation ($\rho \approx 0$).

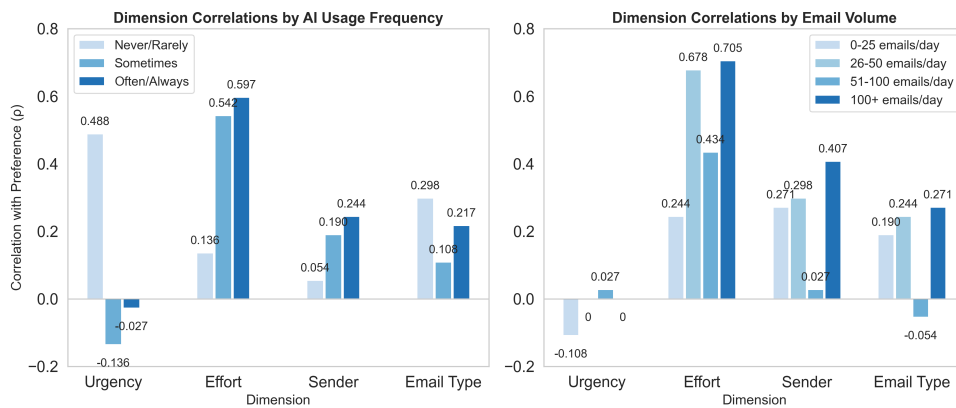


Figure 12: Individual variation in preference patterns across user segments. Left panel shows dimension correlations by general AI-use frequency; right panel shows correlations by email volume. Effort dominates consistently across all groups (blue bars remain high), while Email Type (routine vs. novel preference) exhibits meaningful variation. Participants reporting more frequent general AI use show weaker routine preference ($\rho = 0.217$) compared to infrequent users ($\rho = 0.298$), and high-volume users (100+ emails/day) show stronger routine preference ($\rho = 0.271$) than moderate-volume users.

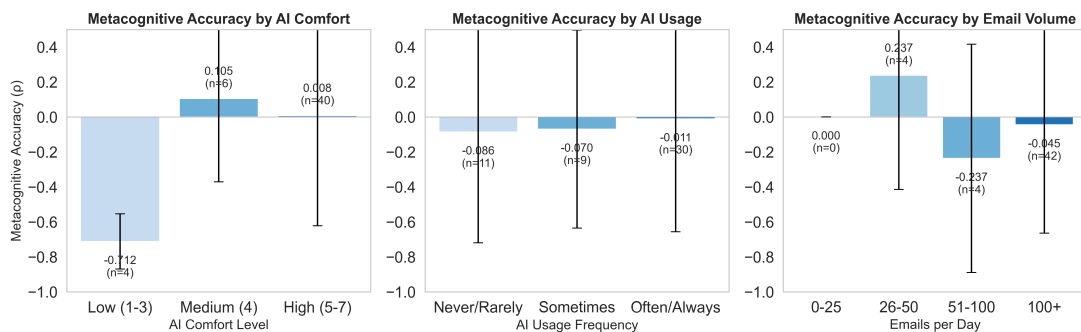


Figure 13: Individual differences in stated-revealed alignment across user characteristics. Left panel shows an exploratory association with AI comfort, with mean participant-level alignment of +0.008 for 40 high-comfort participants and a mean of -0.712 for the four low-comfort participants. Middle and right panels show general AI-use frequency and email volume are not significantly associated with stated-revealed alignment. Error bars represent standard deviation across users in each group.