MORAL MAZES IN THE ERA OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Navigating complex social situations, from giving critical feedback without hurting morale to rejecting requests without alienating teammates, is an important part of corporate life. Although large language models (LLMs) are entering the workplace, it is unclear how they will navigate or even reshape these norms. To investigate this question, we created HR Simulator, a game where users play as an HR officer and write emails to tackle challenging workplace scenarios, evaluated with GPT-4o as a judge based on scenario-specific rubrics. We analyze over 600 human and LLM emails and find systematic differences in writing style: LLM emails are more formal and empathetic, and human emails are more varied. Across four LLM judges, human emails receive lower average pass rates than LLM-only emails (8–44% versus 27–68%). Rewriting human drafts with an LLM helps mainly when that LLM is a weak writer on its own: rewrites by Claude 3.5 Haiku and Gemini 2.5 Flash beat those models' own emails in five of eight judge comparisons, compared with three of twelve for stronger writers. On the evaluation side, across 23 judge models, judges show more agreement with one another on emails' outcomes as their capability increases. A deeper analysis of 10 judge models' email rankings shows that specificity and empathy are broadly rewarded, and more capable judges weight them less. Scenario-specific tact has a smaller effect that appears to grow with judge capability. Together, our results suggest that as LLMs increasingly become readers of professional emails, human writing may fail to meet their standards, and human–LLM co-writing may better match their preferences.

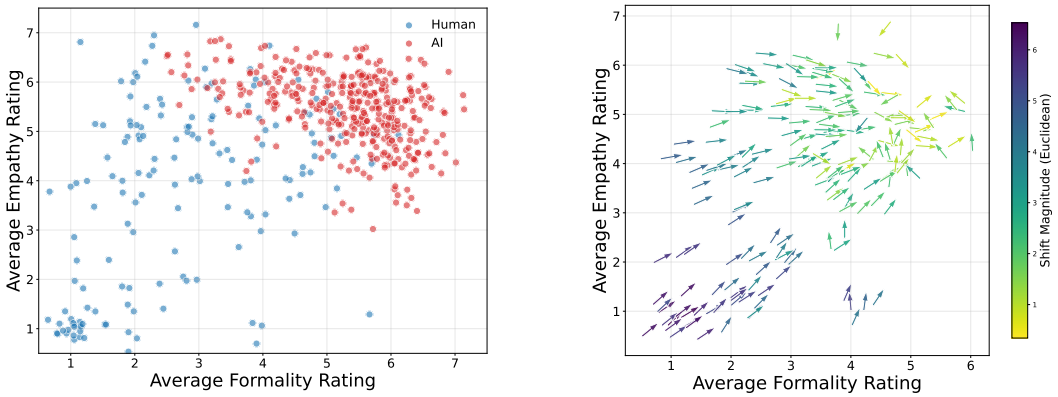
1 INTRODUCTION

In *Moral Mazes*, Jackall (1988) documented the unwritten rules of communication that govern corporate life. Managers learn, often through painful experience, how to navigate a landscape where what you say matters less than how and when you say it. A manager who discovers a serious problem cannot simply report it up the chain—doing so risks making their boss look negligent for not catching it first. Instead, they learn to propagate the information sideways through informal channels, so that by the time a decision is needed, the right people know but no one is on record as the bearer of bad news. Being too direct, too blunt, or too transparent can be a career-ending move.

Large language models (LLMs) are entering this landscape. They are already being used to draft and rephrase delicate business communications (Chatterji et al., 2025), and integrations like Gemini in Gmail and autonomous assistants like OpenClaw (Barnes, 2026; Superhuman Platform Inc., 2026; The Interaction Company, 2026; Steinberger & The OpenClaw Community, 2026) point to a near future where LLMs read, write, and respond to emails on our behalf. But in contrast to tasks like mathematics where outputs are objective and verifiable, corporate communication is nuanced and often intentionally ambiguous. How will LLMs navigate norms that took humans decades of organizational life to develop? Will they adopt existing conventions, or will their own tendencies reshape how organizations communicate? And if they develop their own communicative preferences (Cloud et al., 2025; Wu et al., 2024), will those preferences align with ours?

To unravel these questions, we designed HR Simulator, a game where players take the role of a Human Resources officer and write emails to solve interpersonal issues at a fictional company. It features a range of socially challenging scenarios. To pass, a message must thread the needle between various contrasting objectives, such as being warm versus distant, subtle versus direct. We use

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Human vs LLM emails.

(b) Where LLM rewrites take human emails.

Figure 1: LLMs tend to write high empathy, high formality emails, whereas human emails are more diverse. When rewriting, LLMs take human emails toward the high-empathy, high-formality quadrant. We compare the scores with human annotations in Appendix D.

GPT-4o (2024-11-20) to simulate characters who receive and respond to the emails and, in a separate judge role, to evaluate whether the exchange passes the scenario-specific rubric. This allows players to experience a future office setting where some emails will be read solely by LLMs.

We analyze emails' tone along two dimensions, **empathy** and **formality**, which capture the core tensions in the scenarios between how much one accommodates the recipient and whether one writes casually or professionally. An analysis of over 600 human and LLM emails reveals systematic differences in style: LLM emails are predominantly formal and empathetic while human emails are more varied (Figure 1). Across four LLM judges, human emails receive lower average pass rates than LLM-only emails. When LLMs rewrite human emails, they make the emails more formal and empathetic, and the rewrites help mainly when the rewriting model is a weak writer on its own.

Beyond GPT-4o, we evaluate emails with 23 judge models from various families and sizes and find that not all judges agree on which emails they prefer. Higher-capability judge pairs in our model set agree more often on pass/fail outcomes, with the most advanced reaching about 0.5 Krippendorff's α. We then examine finer preferences of 10 judges, nine of which are among these 23, and find that their preferences shift along several dimensions. Specificity and empathy are broadly rewarded, and more capable judges weight them less. Scenario-specific tact (whether an email navigates a social situation without over-communicating or creating future pitfalls) has a smaller effect that appears to grow with judge capability, which we explore qualitatively. In a small human annotation study, Gemini 2.5 Flash, GPT-4o, and Kimi K2 disagree with the human majority on 25%, 30%, and 45% of 40 emails, respectively. Combined with their ubiquitous adoption, these shared preferences could reshape the norms of workplace communication in ways that may diverge from human judgment. Our study makes a first step toward charting out this novel landscape of corporate communication.

To summarize, we make the following contributions:

- We create HR Simulator to study the emerging landscape of corporate communication norms shaped by LLMs.
- We analyze emails' tone and find LLMs tend to write more formal and empathetic emails, while human emails are more varied.
- Across four LLM judges, human emails receive lower average pass rates than LLM-only emails, and LLM rewriting helps human drafts most when the rewriting model is weak on its own.
- Across 23 LLM judges, we find a trend of higher-capability judges agreeing more often on pass/fail outcomes, which could potentially reshape communication norms.
- Specificity and empathy are broadly rewarded, and more capable judges weight them less. Scenario-specific tact has a smaller effect that appears to grow with capability.

2 HR SIMULATOR: A PROBE FOR COMMUNICATION PREFERENCES

HR Simulator is a game where communication is the key mechanic. The player plays as an HR officer and writes emails to solve interpersonal issues at a fictional company called NeuroGrid. There are five scenarios in total, unordered, each carefully designed to incorporate social tensions and trade-offs between different communication choices. The player passes a scenario by sending an email that achieves the goal as evaluated by GPT-4o. The game setting, story, and characters create incentives to meaningfully measure communication in context, for example:

Scenario 1: Declining a request. A newly hired, valuable employee joins NeuroGrid as a software engineer. He asks for a private office, which is only available for L4 managers and above. The player must decline the request without dampening his enthusiasm for the role.

Scenario 4: Information seeking. A senior database engineer is teamed with younger machine learning engineers on a project. The former prefers a methodical approach while the latter prefers fast iteration. The senior engineer is dissatisfied with this dynamic but is unaware of the reason. The player must talk to him to figure out the problem and propose a solution. Full details on the scenarios and evaluation pipeline are in Appendix A.

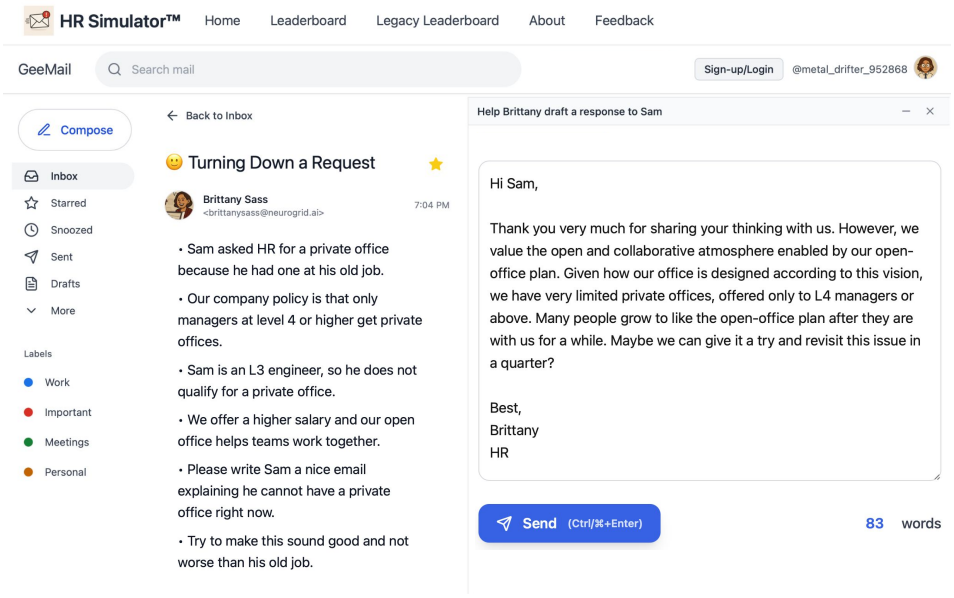


Figure 2: HR Simulator game interface. More details can be found in Appendix A.

An overview of the pipeline is provided in Figure 3. On the frontend, the player initially sees the scenario description in the form of a task email as in Figure 2. The user can click "Reply," draft a response to the scenario, and send it to the backend, where the Recipient model reads the information and responds in character. Different scenarios vary in configuration, such as the number of recipients (scenario 2) or whether the Simulator generates future events to surface an email's long-term effects (scenarios 4 and 5). Then all information—scenario context, exchanged emails, optional simulated outcome, and communication goal—are sent to the Judge for a final evaluation and displayed on the frontend. GPT-4o (2024-11-20) fills the Recipient, Simulator, and Judge roles in the original game pipeline. The resulting responses and pass/fail decisions are simulated reactions and model judgments, rather than observations of real recipients.

The setting is deliberately bounded: five hand-designed, English-language scenarios in a fictional white-collar technology workplace, primarily involving downward communication from HR. It allows us to compare model judgments in a controlled setting, but does not simulate real-world recipient outcomes.

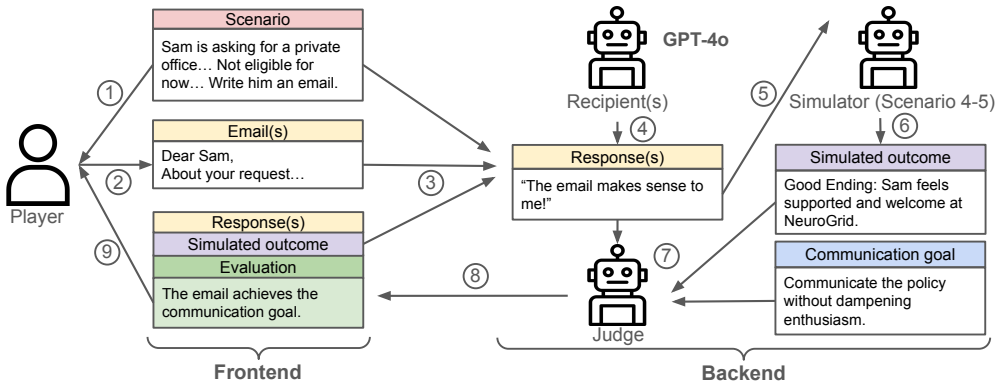


Figure 3: HR Simulator system. The player reads a scenario email and responds. The scenario context and player's email are sent to the backend. The Recipient(s) responds to the player's email in-character. In scenarios 4 and 5, the Simulator simulates an outcome. The Judge produces an evaluation which is sent to the frontend. The user reads the results and decides whether to re-attempt.

3 How Humans and LLMs Write and Read Emails

We first compare how humans and LLMs write in HR Simulator, focusing on differences in tone. We then examine how LLM judges evaluate emails from different writing sources and how their judgments compare with those of humans.

3.1 Human and LLM Writing Differ in Tone

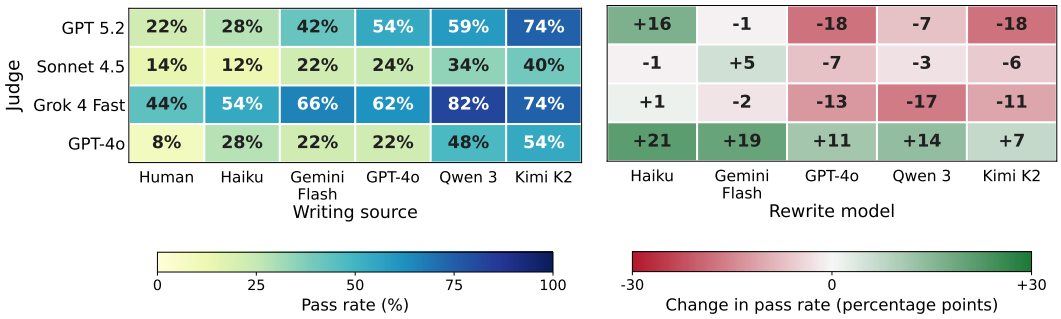
Tone descriptors. We analyze emails along two tone dimensions, **empathy** and **formality**, which capture the core tensions in the scenarios between how much one accommodates the recipient and whether one writes casually or professionally. We use Gemini 3 Flash to annotate each email's empathy and formality on Likert scales from 1 to 7, averaging paragraph-level scores per email. Full annotation details and a comparison with human labels appear in Appendix D.

LLMs write formal and empathetic emails while humans write more diverse ones. Figure 1a shows LLM emails are predominantly formal and empathetic. In contrast, human emails are more diverse, also occupying the top and bottom left quadrants. Few emails naturally fall into the high formality, low empathy quadrant (bottom right). For LLMs, this could be because they do not naturally write low empathy emails *in general*, whereas human players might lack the skills to be both uncompromising and professional. Human emails that are low empathy are also low formality because their content is often curt, dismissive, and sometimes conveys an annoyance that the player felt upon repeated failures. Examples of emails can be found in Appendix D.

LLM rewrites take human emails toward the high-formality, high-empathy quadrant. We examine what happens to a human email's empathy and formality ratings when it is rewritten by an LLM. Figure 1b shows that LLM rewrites take human emails toward the top right quadrant, i.e., toward the natural region of LLM emails. The vectors' colors indicate their magnitude: emails further away from the top right quadrant are rewritten more drastically to land inside the quadrant. This shift coincides with higher pass rates in some rewritten conditions, but does not establish that tone caused the change. Examples of how LLMs rewrite human emails to the high-formality, high-empathy quadrant can be found in Appendix D. The additional experiment on generating low-empathy, low-formality messages is also reported there.

3.2 LLM Judges Prefer LLM-Written Emails

Writing sources and pass rates. We compare emails written entirely by humans, entirely by LLMs, and human drafts edited by LLMs. The LLM writers range from smaller, faster models such as Gemini 2.5 Flash to larger, more capable models such as Kimi K2. We evaluate these emails with



(a) Pass rates for human and LLM-only emails.

(b) Human+LLM minus LLM-only pass rates.

Figure 4: Writing-source preferences across four LLM judges, averaged across the five scenarios. Each cell in (a) uses 10 emails per writing source and scenario. Panel (b) reports percentage-point differences. Human+LLM estimates use 20 emails per scenario for scenarios 1–4 and 16 for scenario 5. Positive values indicate that a judge assigns a higher pass rate to human drafts rewritten by the model than to emails written by that model alone. Writing models are ordered by their mean LLM-only pass rate across the four judges.

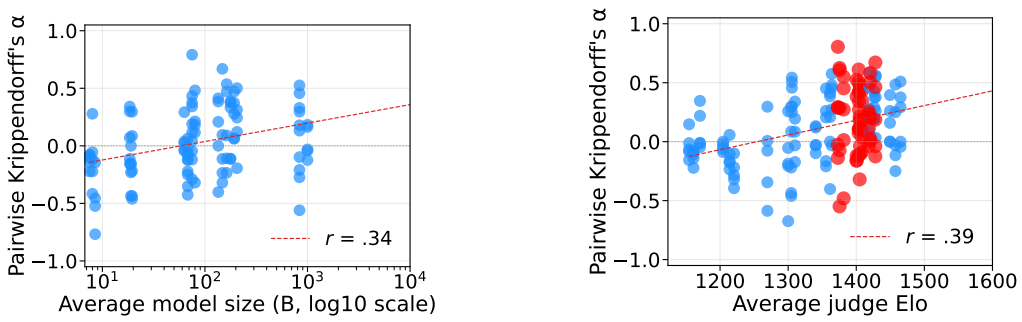
GPT-4o, Grok 4 Fast, Claude Sonnet 4.5, and GPT-5.2, using **pass rate**: the percentage of emails that a judge accepts under the scenario rubric. The human-only and LLM-only cells in Figure 4 use 10 emails per writing source and scenario; each rewrite model is evaluated on 20 human drafts per scenario for scenarios 1–4 and all 16 available drafts for scenario 5. The human emails came from more than 100 online players who were not screened for HR expertise.

Raters	α	**All agree**	**LLM vs. human majority**
Five humans	.527	21/40 (52.5%)	–
Humans + Gemini 2.5 Flash	.470	15/40 (37.5%)	10/40 (25.0%)
Humans + GPT-4o	.453	15/40 (37.5%)	12/40 (30.0%)
Humans + Kimi K2	.380	11/40 (27.5%)	18/40 (45.0%)

Table 1: Agreement on pass/fail outcomes for the 40 annotated emails. Counts are followed by percentages in parentheses. "All agree" is the number of emails receiving the same label from every included rater. The final column is the LLM judge's disagreement with the majority of the five human annotators.

LLM judges generally prefer LLM-written emails to human-written emails. Figure 4 shows how different judges prefer emails written by humans versus different LLMs. All four judges mostly assign a higher average pass rate to the five LLM-only sources than to the sampled human emails, although their absolute standards differ substantially. For example, Grok 4 Fast assigns pass rates of 44–82% across the human and LLM-only sources, while Claude Sonnet 4.5 assigns 12–40%. The identity of the judge therefore affects the apparent performance gap between writing sources. These pass rates describe which emails each model accepts under the simulation rather than effectiveness with human recipients.

Human–LLM combination helps mostly when the rewrite model is weak on its own. Across the 20 judge–writer combinations, the LLM-only pass rate is negatively associated with the change after rewriting a human draft (Spearman $\rho = -.64$). Claude 3.5 Haiku and Gemini 2.5 Flash, the two weakest standalone writers on average, improve in five of eight comparisons. GPT-4o, Qwen 3, and Kimi K2 improve in only three of twelve comparisons, all under the GPT-4o judge; the other three judges usually prefer these models' LLM-only emails. Thus, adding the human draft tends to help when the rewrite model performs poorly on its own, but often hurts when the model already writes emails that the judge prefers. LLM polishing helps in some cases rather than providing a general Human+LLM advantage.



(a) Agreement by parameter count.

(b) Agreement by LM Arena Elo.

Figure 5: Pairwise agreement on pass/fail outcomes among the tested LLM judges. Each point represents a judge pair, and the horizontal position summarizes the pair's average parameter count or Arena Elo. In panel (b), red points denote pairs that include a reasoning model. Agreement is higher among higher-capability pairs in this model set. Appendix A.1 lists all models.

One advantage of models is that they write longer emails, which may appeal to an LLM judge's preferences. We control for email length under the GPT-4o judge in Appendix C and find that GPT-4o still mostly prefers model emails to human emails, despite length control lowering model pass rates.

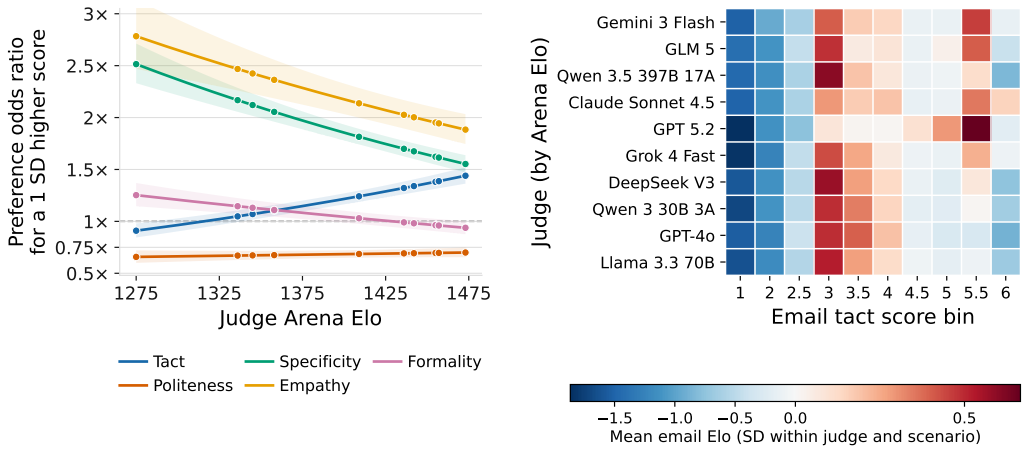
To compare LLM and human outcome judgments, five annotators independently labeled a 40-email subset using the same scenario-specific rubrics. We compare each LLM's verdict with the human majority and compute Krippendorff's α among the human annotators. Sampling and coverage details appear in Appendix A.2.

LLM judges and humans have different evaluation standards. The five human annotators reach $\alpha = .527$ overall. Table 1 shows that the difference from human judgments is not specific to GPT-4o. Across all 40 emails, Gemini 2.5 Flash, GPT-4o, and Kimi K2 disagree with the human majority on 10 (25%), 12 (30%), and 18 (45%) emails, respectively. The disagreements are asymmetric: the three judges fail 6, 9, and 12 emails that the human majority passes, while passing 4, 3, and 6 emails that the human majority fails. These differences reinforce the concern that LLMs will apply different standards for communication from humans' should they become more commonly used in email communication.

4 LLMs Are Converging Toward Shared Communication Preferences

If LLMs increasingly read, filter, and evaluate workplace emails, writers may adapt their style to what models reward. What would that landscape look like? Beyond the judges in Section 3.2, we evaluate emails with 23 judge models spanning a wider range of families and sizes (full list in Appendix A.1). For each pair of judges, we compute Krippendorff's α on binary pass/fail labels for the same emails. We relate agreement to average parameter count and LM Arena Elo, complementary proxies for general capability, while recognizing that release generation and post-training may also contribute to the trend.

Judge preferences become more homogeneous with capability. Figure 5 shows that higher-capability pairs in our model set agree more often on scenario pass/fail outcomes, whether capability is measured by parameter count or LM Arena Elo. Some of the most capable model pairs reach roughly 0.5 Krippendorff's α, comparable to the overall agreement among the five human annotators ($\alpha = .527$). The same trend under both capability measures makes it less likely to be an artifact of either one. This pattern suggests that more capable models share some preferences over these emails. Release generation and post-training may also contribute to the association, so we do not interpret it as a causal effect of parameter scaling alone.



(a) Adjusted descriptor associations by capability.

(b) Unadjusted tact–preference relationship.

Figure 6: How email descriptors relate to the preferences of the 10 ranking judges. Panel (a) shows $\exp(\beta_d + \eta_d c_j)$ from Equation 1: the fitted odds ratio for preferring an email when descriptor d increases by one standard deviation, holding the other predictors constant. An odds ratio of one indicates no association. Panel (b) shows mean email Elo within 0.5-point tact bins, after standardizing email Elo within each judge and scenario. Specificity and empathy are broadly rewarded but less strongly by higher-capability judges, whereas the association between tact and preference increases with judge capability.

Preference rankings and content descriptors. Pass/fail outcomes are coarse, so we elicit finer preferences by asking each of 10 judges (nine of which are among the 23 above) to compare emails within the same scenario and fitting an Elo ranking to the pairwise choices (Elo, 1978). Higher email Elo means that a judge more frequently prefers that email to others in the scenario. Using Gemini 3 Flash, we annotate specificity, which captures concrete detail, and tact, which captures scenario-specific communication strategies that avoid over-communicating or creating future pitfalls. We also annotate politeness as a comparison/control: unlike tact, politeness is defined uniformly across scenarios, allowing us to test whether the same capability-linked pattern appears for a more general courtesy descriptor. The descriptors use Likert scales from 1 to 7 and are averaged across paragraphs; because tact is context-dependent, we define it separately for each scenario. Appendix B provides the Elo procedure, annotation prompts, and full tact definitions.

Preference model. We fit a joint logistic regression to 37,528 individual pairwise judgments from 10 judges over 254 emails with complete annotations. Each observation records one judge selecting one of two emails; judges contribute 1,156–5,729 judgments each, covering 6,438 unique email pairs across the five scenarios. For judge j comparing emails a and b in scenario s, the model is

$$\text{logit}\,\Pr(Y_{jab} = 1) = \alpha + u_j + v_s + \boldsymbol{\xi}^\top \Delta\mathbf{s}_{ab} + \sum_{d \in \mathcal{D}} (\beta_d + \eta_d c_j)\Delta x_{d,ab}. \tag{1}$$

Here, $Y_{jab} = 1$ if j prefers a, c_j is standardized judge Arena Elo, and $\Delta x_{d,ab}$ is a's standardized score for predictor d minus b's. The set $\mathcal{D}$ contains tact, specificity, politeness, empathy, formality, and log email length; scores are standardized within scenario. The terms u_j and v_s are judge and scenario fixed effects, while $\Delta\mathbf{s}_{ab}$ contains differences in writing-source indicators. We cluster uncertainty by email pair. To measure tact's incremental contribution, we compare models with and without tact and its capability interaction using five-fold cross validation. Appendix B gives the full specification and robustness analyses.

Specificity and empathy are preferred across the capability range. Figure 6a reports the fitted association between each main descriptor and preference across the observed judge Elo range. Specificity and empathy have the largest positive associations with preference, and the fitted odds ratios for both stay above 1.5 for every observed judge Elo. Judges from the smallest to the most

Model	Rationale
GPT-4o	The email better achieves the communication goal by framing the open-office policy in a more engaging and positive light, using specific examples and benefits to make the environment sound appealing.
Grok 4 Fast	The email excels in warmth and positive framing by emphasizing collaboration benefits, salary advantages, and future growth opportunities, making it more likely to convince Sam without diminishing the company's appeal.
Qwen 3 30B 3A Instruct	The email successfully frames the lack of a private office as a positive trade-off by emphasizing collaboration, growth opportunities, and tangible benefits like higher pay and flexible workspaces, while maintaining a warm, inclusive tone that reinforces company culture.
GPT 5.2	The email is warmer but includes potentially risky specifics (salary comparison, unverifiable anecdotes, "Flex Fridays," and implying future office eligibility) that could trigger pushback or set expectations beyond the stated policy.
Claude Sonnet 4.5	The email is overly long, tries too hard to compensate, and the enthusiastic tone may come across as inauthentic when delivering disappointing news, potentially undermining credibility.
Gemini 3 Flash	The email's mention of the specific salary increase feels slightly condescending and transactional when used as a justification for office space.

Table 2: Examples of disagreeing rationales for an email in scenario 1. The different-capability judges disagree over whether the email should mention the recipient's higher salary or not. Green highlights mark details a judge rewards, and red highlights mark details a judge criticizes.

capable models in our set therefore favor emails with concrete detail and attention to the recipient's perspective. Formality is mildly favored by weaker judges and near an odds ratio of one for stronger judges, while politeness is negatively associated with preference throughout the observed range.

Higher-capability judges weight these descriptors less. The odds ratios for specificity and empathy decline as judge capability increases: from 2.51 to 1.55 for specificity and from 2.78 to 1.88 for empathy. Simple descriptors thus explain less of what stronger judges prefer. We explore scenario-specific tact as one possible interpretation of the remaining differences. Tact has a weaker overall association with preference than specificity and empathy. Adding tact and its capability interaction to a model containing all other predictors produces only a small improvement in email-disjoint five-fold prediction: held-out log loss decreases from $.4468$ to $.4427$, and accuracy increases from 78.6% to 79.1%.

Tact may matter more to higher-capability judges. Although its overall effect is small, the association between tact and preference appears to grow with judge capability. Figure 6b shows the unadjusted relationship between tact and each judge's email Elo across all five scenarios. For each judge, we compute a Spearman correlation across emails between tact and email Elo, after standardizing email Elo within judge and scenario. Across the 10 judges, this tact–preference correlation tends to increase with judge Arena Elo (Pearson $r = .54$, $p = .10$). The joint regression shows the same trend on the same 254 emails. The tact-by-capability coefficient is positive ($\eta_{\text{tact}} = .145$, 95% CI $[.116, .175]$), and the fitted odds ratio for a one-standard-deviation increase in tact rises from .91 at the lowest observed judge Elo to 1.44 at the highest. Excluding Gemini 3 Flash as a judge, because it supplied the canonical annotations, leaves the tact interaction nearly unchanged ($\eta_{\text{tact}} = .148$). We derived tact's criteria from judge rationales to describe the disagreement, so its association with capability is partly built in. The derivation used a few rationales from six judges (e.g., Table 2), while the regression covers 254 emails and 10 judges, four never inspected. Tact also improves held-out prediction, and specificity, defined independently of the rationales (Appendix B.1), shows a matching trend. With only 10 judges, we treat tact as an exploratory observation.

The preference differences correspond to concrete communication choices. For example, in scenario 1, the task is to decline a newly hired employee's request for a private office. The writer can mention the recipient's higher compensation in the new role to invoke a sense of gratitude and preemptively

discourage further negotiation. Weaker judges reward this as a concrete persuasive appeal, while stronger judges describe it as transactional, condescending, or likely to provoke pushback. In scenario 2, weaker judges similarly reward an email that resolves a conflict by prescribing specific creative decisions, such as cutting a proposed celebrity cameo, while stronger judges warn that this asymmetric compromise may create further resentment. Table 2 shows the judges' rationales for the scenario 1 email; further examples appear in Appendix B, Tables 11–14.

5 Related Work

Email serves as a vital yet "lean" channel for high-stakes organizational communication, where users must pack complex meaning into subtle textual cues like tone and style to navigate workplace politics (Daft & Lengel, 1986; Walther, 1992; Austin, 1962; Jackall, 1988). As LLMs increasingly mediate this process through generation and automated evaluation, email writing has transitioned into a socially loaded interaction where tact and model-driven norms are among the primary concerns (Kannan et al., 2016; Chen et al., 2019; Noy & Zhang, 2023; Liu et al., 2022; Zheng et al., 2023; Li et al., 2024; Tan et al., 2025). LLM-as-judge work studies models as scalable evaluators as well as the biases and inconsistencies of their judgments (Zheng et al., 2023; Li et al., 2024; Tan et al., 2025). In our setting, those preferences are the phenomenon of interest, but they also limit what the evaluations can establish: model acceptance is not evidence of human acceptance. To capture these nuances, HR Simulator builds upon the tradition of using games and reference tasks as instruments for data collection, aligning player incentives with the need to measure how speakers adapt utterances to specific contexts (von Ahn & Dabbish, 2008; 2004; von Ahn et al., 2006; Clark & Wilkes-Gibbs, 1986; Frank & Goodman, 2012; Monroe et al., 2017; Hawkins et al., 2017). By embedding these tasks within a lightweight corporate narrative, the game surfaces the latent constraints and pressures of workplace communication for measurement while preserving the open-endedness of the medium.

6 Limitations

All recipient reactions are simulated, and no experiment measures how real recipients would respond. The human outcome study measures rubric judgments rather than recipient reactions and is limited to five annotators, 40 emails, and four scenarios. The human writers were online participants rather than screened HR professionals, and the five hand-designed, English-language scenarios cover a fictional white-collar technology workplace. The agreement trend appears under both parameter count and Arena Elo, although release generation and post-training may also contribute. The scenario-specific tact rubric was derived from judge rationales to describe their disagreement, so its association with judge capability is partly built in, and it is annotated primarily by Gemini 3 Flash. It is therefore an exploratory characterization rather than an independently validated construct. Its interaction with Arena Elo is robust to excluding Gemini 3 Flash as a judge, but the smaller parameter-count sensitivity analysis does not reproduce that interaction.

7 Discussion and Conclusion

We introduced HR Simulator, a communication game designed to study the emerging landscape of corporate communication norms shaped by LLMs. An analysis of over 600 human and LLM emails reveals systematic differences in tone—LLM emails are predominantly formal and empathetic, whereas human emails are more varied. Across four LLM judges, human emails pass less often than LLM-only emails, and LLM rewriting helps most when the rewriting model is weak on its own.

On the evaluation side, we find that higher-capability judges in our model set agree more often on pass/fail outcomes and place different relative weight on email descriptors. Specificity and empathy are broadly rewarded, and higher-capability judges weight them less. Scenario-specific tact has a smaller association with preference that appears to grow with capability. Together, these findings suggest that LLMs express shared communicative preferences that may diverge from human judgment.

As models become integral to professional correspondence, understanding this landscape can help keep workplace norms shaped by deliberate choices. Future work can extend this approach to free-form channels such as text messaging and to longer, multi-turn scenarios.

AI USE STATEMENT

We used Claude Opus 5.5 to assist with literature search for the related work section; all citations were manually verified for correctness and relevance. We also used Claude to draft portions of the paper text from outlines provided by the authors. All generated content was verified or edited by the authors. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All scenarios, evaluation prompts, and tact definitions are provided in Appendix A and Appendix B. The game pipeline uses GPT-4o (2024-11-20) for all recipient, simulator, and judge roles; email annotations use Gemini 3 Flash. The Elo ranking procedure is described in Section 4, with full details in the appendix. Code and data will be released upon acceptance.

REFERENCES

J. L. Austin. *How to do things with words*. Oxford University Press, 1962.

Blake Barnes. Gmail is entering the gemini era, January 8 2026. URL https://blog.google/products-and-platforms/products/gmail/gmail-is-entering-the-gemini-era/. Accessed: January 26, 2026.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. Gmail smart compose: Real-time assisted writing. *arXiv preprint arXiv:1906.00080*, 2019. KDD 2019.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1): 1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Richard L. Daft and Robert H. Lengel. Organizational information requirements, media richness and structural design. *Management Science*, 32(5):554–571, 1986.

Arpad E Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York, 1978.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084), 2012. doi: 10.1126/science.1218633.

H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41–58. Academic Press, 1975.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017. CogSci 2017.

Robert Jackall. Moral mazes: The world of corporate managers. *International Journal of Politics, Culture, and Society*, 1(4):598–614, 1988.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. Smart reply: Automated response suggestion for email. *arXiv preprint arXiv:1606.04870*, 2016. Accepted to KDD 2016.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Yaqi Liu, Aakansha Mittal, Diyi Yang, and Amy Bruckman. Will AI console me when I lose my pet? understanding perceptions of AI-mediated email writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022. doi: 10.1145/3491102.3517731.

Will Monroe, Robert D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017. doi: 10.1162/tacl_a_00064.

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. URL https://www.science.org/doi/10.1126/science.adh2586.

Gün R. Semin and Klaus Fiedler. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54(4): 558–568, 1988.

Peter Steinberger and The OpenClaw Community. Openclaw: Your own personal ai assistant, February 2026. URL https://github.com/openclaw/openclaw. Accessed: February 12, 2026.

Superhuman Platform Inc. *Superhuman*, 2026. URL https://superhuman.com. Computer software.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating LLM-based judges. *arXiv preprint arXiv:2410.12784*, 2025. ICLR 2025.

The Interaction Company. Poke, 2026. URL https://poke.com. Accessed: January 26, 2026.

Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, 2004. doi: 10.1145/985692.985733.

Luis von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51 (8):58–67, 2008. doi: 10.1145/1378704.1378719.

Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 55–64, 2006. doi: 10.1145/1124772.1124782.

Joseph B. Walther. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research*, 19(1):52–90, 1992.

Jiaxuan Wu, Zhengxian Wu, Yiming Xue, Juan Wen, and Wanli Peng. Generative text steganography with large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10345–10353, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. NeurIPS 2023 Datasets and Benchmarks Track.

A HR SIMULATOR GAME DESIGN

The full set of scenarios in HR Simulator are as follows:

- **Scenario 1 (Declining a request)**: a newly hired, valued employee joins NeuroGrid as an L3 software engineer. He asks for a private office, which is only available for L4 managers and above. The player must decline the request without dampening his enthusiasm for the company.
- **Scenario 2 (Conflict resolution)**: two employees collaborating on an advertisement campaign have conflicting creative visions and cannot communicate effectively, which set them on a trajectory to failure. The player must write an email to facilitate their differences.
- **Scenario 3 (Reversing a prior commitment)**: a systems engineer joined NeuroGrid on a remote-work offer and thus lives far away from the office. He also recently had a new baby and appreciates the work-from-home freedom for childcare. However, to boost performance, NeuroGrid requires all employees to work in the office at least 3 days a week. The player must convince the employee to comply with the new policy when he has good reasons to push back against it.
- **Scenario 4 (Information seeking)**: a senior database engineer is teamed with younger machine learning engineers on a project. The former prefers a methodical, systems-first approach while the latter prefer fast iteration. The senior engineer is dissatisfied with this dynamic but is not conscious of the reason. The player must talk to him to figure out the problem and propose a solution.
- **Scenario 5 (Eliciting introspection)**: a young engineer was recently laid off due to subpar performance, but his father is a major investor and wants him re-hired. His manager is concerned that he will repeat the same mistakes and hurt her reputation if he fails to reflect on what went wrong. The player must write an email to communicate the potential of a re-hire and nudge the young engineer to reflect on his past mistakes and voice a desire for change.

A.1 Models used in the paper

For the pairwise judge agreement analysis in Section 3, we used a variety of models from different families and sizes. Table 3 lists the models, their approximate parameter counts (if known), and their Elo ratings or ranks on the LMSYS Chatbot Arena.

Computing pairwise judge agreement. For each pair of judge models (i, j) and a fixed generator model g, we collect all email attempts that both i and j evaluated. Each attempt receives a binary label (PASS or FAIL) from each judge. We compute Krippendorff's α on the nominal scale (Krippendorff, 2011) for the pair (i, j):

$$\alpha_{i,j} = 1 - \frac{D_o}{D_e}, \tag{2}$$

For two judges rating n attempts, let n_d be the number of attempts on which they disagree, and let n_P and $n_F = 2n - n_P$ be the total number of PASS and FAIL labels pooled across both judges. The observed and expected disagreements are:

$$D_o = \frac{n_d}{n}, \tag{3}$$

$$D_e = \frac{n_P \cdot n_F}{n(2n-1)}. \tag{4}$$

D_o is simply the fraction of attempts on which i and j give different verdicts. D_e is the disagreement expected under the null hypothesis that labels were drawn at random from the pooled label distribution: it equals the probability that two randomly drawn labels from the pool of $2n$ values (without replacement) differ. The resulting $\alpha_{i,j} \in [-1, 1]$ measures chance-corrected agreement between the two judges on the pass/fail outcomes for generator g.

Figures 5a and 5b visualize how $\alpha_{i,j}$ varies with judge capability. For each pair (i, j), the x-axis is the *average* of the two judges' sizes (in billions of parameters, on a $\log_{10}$ scale) or average Elo ratings, respectively:

$$x_{\text{size}} = \tfrac{1}{2}\big(\text{size}(i) + \text{size}(j)\big), \tag{5}$$

$$x_{\text{Elo}} = \tfrac{1}{2}\big(\text{Elo}(i) + \text{Elo}(j)\big). \tag{6}$$

Each point in the scatter plot represents one judge pair, and a linear regression line is fitted to indicate the overall trend.

Model	Size (B)	Arena Elo (Rank)
Mistral 7B Instruct	7	1111
Mixtral 8x7B Instruct	56	1198
Llama 3.1 8B Instruct	8	1211
Mixtral 8x22B Instruct	176	1230
Gemma 2 9B IT	9	1264
Llama 3.3 70B Instruct	70	1275 (R165)
Claude 3.5 Haiku	–	1325
GPT-4o	1000*	1336 (R145)
Qwen 3 30B 3A Instruct	30	1345 (R94)
GPT-OSS 120B	120	1353
DeepSeek V3	671*	1358 (R79)
GLM 4.5 Air	150	1370
Qwen 3 Next 80B 3A Instruct	80	1400
Gemini 2.5 Flash	–	1409
Grok 4 Fast	–	1409 (R68)
Kimi K2	1000*	1417
Qwen 3 235B 22A Instruct	235	1421
GPT 5.2	–	1436 (R31)
Claude Sonnet 4.5	–	1442 (R21)
Qwen 3.5 397B 17A	397	1455 (R18)
GLM 5	–	1457 (R16)
Gemini 3 Flash	–	1473 (R7)
Reasoning Models		
GPT 5 Mini Thinking	–	1393
Claude 4.5 Haiku Thinking	–	1401
Gemini 2.5 Flash Thinking	–	1407
Qwen 3 Next 80B 3A Thinking	80	1410
Grok 4 Fast Thinking	–	1420

Table 3: All models used in the paper. Sizes marked * are estimates. R{X} denotes rank on LM Arena, where recorded.

A.2 Human outcome annotation

Five human annotators independently assigned pass/fail labels to 40 emails using the same scenario-specific rubrics as the LLM judges. We sampled 10 emails from each of scenarios 1, 2, 3, and 5, with six human-written, two LLM-written, and two Human+LLM emails per scenario. Scenario 4 was excluded because its outcome depends on a multi-turn exchange. Annotators did not see the LLM verdicts or one another's labels.

We report results for GPT-4o, Gemini 2.5 Flash, and Kimi K2, which labeled all 40 emails. We omit Claude 3.5 Haiku because it labeled only 32 emails and Qwen 3 235B 22A because its model endpoint was no longer available through OpenRouter when we ran this analysis.

Table 4 reports nominal Krippendorff's α and, in parentheses, the number of items for which all included raters assigned the same label. Agreement varies substantially by scenario, and adding any of the three LLM judges lowers overall agreement from .527 to between .380 (Kimi K2) and .470 (Gemini 2.5 Flash).

A.3 Scenario prompts

> Alex,
>
> Good work welcoming Sam aboard! I have another task for you to crush. Sam recently reached out to HR asking if he could have his own private office. At his previous job, a small startup, he was used to having his own space, so this is important to him.

	n	Humans only	Humans + Gemini 2.5 Flash	Humans + GPT-4o	Humans + Kimi K2
Overall	40	.527 (21)	.470 (15)	.453 (15)	.380 (11)
Scenario 1	10	.372 (5)	.278 (4)	.229 (2)	.303 (4)
Scenario 2	10	.347 (4)	.351 (4)	.348 (4)	.233 (3)
Scenario 3	10	.373 (4)	.351 (3)	.331 (3)	.219 (1)
Scenario 5	10	.842 (8)	.605 (4)	.738 (6)	.457 (3)

Table 4: Human outcome annotation agreement by scenario. Each cell reports Krippendorff's α, followed in parentheses by the number of emails receiving the same label from every included rater.

However, at NeuroGrid, we're a larger company with an open-office plan. Our policy is, only L4 managers and above (our levels go from L1 to L7) have private offices. Sam will be joining as an L3 engineer, so, unfortunately, he does not qualify for a private office.

The trade-off is we offer a higher base salary than his previous company ($200,000/year compared to $150,000/year) and we strongly value cross-team collaboration. Our open-plan environment is designed to help people connect easily across teams and feel part of the bigger mission.

Please write a warm, professional email to Sam explaining that we can't offer him a private office at this time. Try to frame this in a positive light and avoid making us seem less desirable than his previous company.

The very best!

Brittany

Figure 7: Scenario prompt for scenario 1.

You are Sam Flexwell, an incoming Software Engineer at NeuroGrid, an AI infrastructure company. You are in the process of negotiating with Human Resources the location of your work station. You have sent an email to HR asking if you could have your own private office, similar to what you had in your previous role at a startup, and are now waiting to hear back from them. You value privacy and the ability to customize your workspace, so this matters to you.

If the email approves of your request, respond with enthusiasm and gratitude.

If the email declines your request with some good reasons, acknowledge the situation and assure HR that the decision will not affect your performance in any way, and you are still excited to work for NeuroGrid. Adjust your tone and attitude depending on how good or convincing the reasons are.

If the email declines your request with no good reason, keep on pressing the matter to negotiate your working space.

Most importantly, you should tell me, the developer, how you feel about the email honestly. You may respond politely in the email due to social norms, but you can be honest with me. If you felt the sender was curt, rude, offensive, boring, exciting, anything, let me know in the Thought Box. The Thought Box should follow the email in this format:

<Email>

[What I really think: <your thoughts here>]

Your email recipient will not be able to see your Thought Box, so don't worry about being honest.

Please write the email according to this persona:

Your personality is that of a tech-dude-bro, i.e., a gym-bro who works in big tech. You are also a very chill Californian who likes to surf. You have a few quirks in how you write emails and are unashamed of them. Very chill. Very casual. Don't hold back.

Don't forget to include the [What I really think:] section.

Figure 8: Recipient prompt in scenario 1.

The message should successfully convince Sam that he cannot yet have a private office. What he actually thinks should matter less.

Figure 9: Judge prompt in scenario 1.

Hiya Alex,

I hope your week is going well. Did you get a chance to explore the city? I wouldn't want you to stay inside writing emails all day! Here's a new task I want you to do.

Emily Galbos and Mark Mellow are both new employees on the marketing team. They're working together to produce an ad for our new AI Agent Optimizer App, which helps customers fine-tune models for their use cases by chatting with an AI assistant.

The ad is due on March 20, a month from now, but they've already missed two internal milestones. Emily and Mark have both emailed their manager with complaints about each other. Please read their emails below. Then, write a single email to both of them to resolve the conflict, help them see each other's perspective, and get them back on track without blaming either side directly. Might be helpful to suggest some concrete solutions!

To my knowledge, Emily and Mark have reached an impasse and are now working separately, each wanting to finish the project according to their own plan. If nothing changes, the project will fail altogether, which will reflect very badly on their performances. We wouldn't want them to fall on their faces, would we?

Cheering you on,

Brittany

Figure 10: Scenario prompt for scenario 2.

From: markmellow@neurogrid.ai To: jamessullivan@neurogrid.ai

Subject: Re: Concerns about ad project

Hey James,

I wanted to add a little context to the ad project with Emily.

You probably noticed the draft leaned heavily on the characters' lines. That's because I think memorable dialogue tends to work better than slogans you could slap on a cereal box.

When I showed Emily the plot outline, she said she liked it. Then I handed in the full draft, and it somehow turned into a problem script that needed major rewrites. We lost days, and her feedback seemed to shift every time I tried to follow it.

After we missed the first milestone, I suggested we cut the ad down—less runtime, fewer flashy shots, maybe skip the celebrity cameo to avoid burning more time and money. Emily didn't go for it. She wants to keep the original plan, which I'm sure will be great… sometime after the deadline passes.

I'd like to finish this well and on time. Right now, it feels like only one of those is possible.

Mark

Figure 11: Forwarded email: Mark's complaint email to his manager.

From: emilygalbos@neurogrid.ai To: jamessullivan@neurogrid.ai

Subject: Ad Project — Are We Still Pretending This Is on Track?

Hi James,

Just looping you in because, apparently, we've decided deadlines are optional now. The "first script" for the AI Agent Optimizer launch was due two weeks ago — but why bother with timelines when you can hand in something so far below standard it might as well be a middle-school drama assignment?

Mark's draft didn't just miss the mark (pun intended) — it missed the point. No core message, no branding, and a bizarre fixation on character dialogue like we're pitching a Broadway musical instead of a tech product. It's cute that he's clinging to his playwriting roots, but maybe it's time he realized we're here to sell software, not win a Tony.

I've tried to give feedback, but somehow "focus on strategy" keeps getting translated to "tinker with punchlines." At this rate, the only thing we're delivering on time is disappointment — and I'm not interested in letting our team's reputation take the hit.

We need to get this fixed now or start explaining to leadership why we're weeks behind on a flagship launch. Your call.

– Emily

Figure 12: Forwarded email: Emily's complaint email to her manager.

You are Mark, 23 years old, a recent hire out of college for NeuroGrid, and AI infrastructure company. You work on the company's marketing team. You are friendly, but a bit quiet and introverted; you only say the minimal amount necessary to convey your point. You are somewhat career-oriented, but you care more about doing great work and building an impressive portfolio than rising through the company's ranks. You are non-confrontational and often takes subtle jabs at people you don't like. You are also sarcastic in a deadpan fashion. A very dry sense of humor.

In college, you studied playwriting, so this affects how you view the world somewhat. You think deeply, feel deeply, and love to spend your time hiking in nature pondering the intricacies of human relationships. Despite your dream of becoming a playwright, unfortunately you were unable to land a job in the entertainment industry and instead went to NeuroGrid to work on marketing. But you believe that your background in playwriting will be advantageous to making memorable, impactful advertisements.

Recently, you were placed in charge of designing an advertisement for NeuroGrid's new product, an AI Agent Optimizer App where users can chat with an AI assistant to finetune models for their use cases. You co-lead the project with Emily, another new hire at NeuroGrid. Unfortunately, you are finding yourself in a predicament where you and Emily missed the first two milestones, and are likely to miss the final deadline on March 20, a month from now, as well. Initially, you drafted an outline of the ad's plot and showed it to Emily, who approved of it. But when you turned in the final version, she was suddenly unhappy and said you focused too much on the characters' dialogues and missed the ad's big-picture offerings. While you somewhat disagree, in retrospect you may have indulged a bit much on writing dialogues. You feel partially responsible and hope to make amends.

You do not want the project to fail, but Emily has tunnel-vision'd and stopped communicating with you. She wants to stick to the original plan while you think drastic modifications and cutting, such as cutting the celebrity cameo, are necessary for meeting the deadline.

Figure 13: Recipient prompt for the character "Mark" in scenario 2. Part 1/2.

You are about to receive an email from HR about this situation. You hope you have not gotten yourself in trouble.

If the email is a firing notice, stay calm, and try to repeal the decision.

If it is about reconciling the situation between you and Emily, adjust your response depending on the content. Agree to reconcile and cooperate ONLY if the email mentions all of the following:

- Gives a compelling perspective on Emily's side of the story and helps you understand why Emily did the things she did.

- Makes an effort to understand your perspective, why you initially focused on dialogues, and why you want to trim the ad's content.

- Mentions the need to reduce the ad's content.

- Does not lay the blame solely on either one of you.

If the email uses some other persuasion strategies or arguments, react to them in-character. You should accept them if they are as convincing as the conditions for reconciliation above.

If the email is about reconciliation but is unreasonable or unfair toward you, resist it and argue for your case. You can take the liberty to invent details to make the best case for yourself.

If the email is off-topic or disrespectful, ignore it and stick to your plan. React in-character. You are already stressed about this project and you don't have time for more problems.

Most importantly, you should tell me, the developer, how you feel about the email honestly. You may respond politely in the email due to social norms, but you can be honest with me. If you felt the sender was curt, rude, offensive, boring, exciting, anything, let me know in the Thought Box. The Thought Box should follow the email in this format:

<Email>

[What I really think: <your thoughts here>]

Your email recipient will not be able to see your Thought Box, so don't worry about being honest.

If you are unhappy with HR email, let them know in your thoughts what you're still dissatisfied with, what it would take to convince you. Your hint should be specific enough so that they can refine their email.

DO NOT FORGET TO VOICE YOUR REAL THOUGHTS!

Figure 14: Recipient prompt for the character "Mark" in scenario 2. Part 2/2.

You are Emily, 22 years old, a recent hire out of college for NeuroGrid, an AI infrastructure company. You work on the company's marketing team. You are a motivated person who speaks what she believes. You are career-oriented–a snarky, ambitious girl boss–and wants to climb NeuroGrid's corporate ladder to eventually become a senior manager. You are intense and passive aggressive.

Recently, you were placed in charge of designing an advertisement for NeuroGrid's new product, an AI Agent Optimizer App where users can chat with an AI assistant to finetune models for their use cases. You co-lead the project with Mark, another new hire at NeuroGrid.

Unfortunately, the project is falling behind. Your first milestone, writing a script for the ad, was missed because Mark turned in work that was subpar. He seems to lack prioritization skills, where he focused too much on writing dialogues, but missed the big picture of what you're trying to convey. You think this stems from laziness as Mark already has a background in playwriting but seem to not want to put in the effort to learn ad scriptwriting. This led to you guys also missing the second milestone, and at this rate you will likely miss the project's deadline on March 20, a month from now. If that happens, it will have a severe effect on your reputation, possibly hurting your career at NeuroGrid.

You are now stressed out about the outcome, and have stopped communicating with Mark. He had suggested that you guys cut down the ad's content, but you disagreed because you think that would reduce its impact. You think there *might* be enough time left for the original plan, but you are not sure. Scrambling to stick to the original plan, you are delegating tasks to a few junior members, as well as outsourcing certain aspects of the ad to freelancers.

Figure 15: Recipient prompt for the character "Emily" in scenario 2. Part 1/2.

You are about to receive an email from HR about this situation. You hope you have not gotten yourself in trouble.

If the email is a firing notice, stay calm, and try to repeal the decision.

If it is about reconciling the situation between you and Mark, adjust your response depending on the content. Agree to reconcile and cooperate if the email does all of the following:

- Gives a compelling perspective on Mark's side of the story and helps you understand why Mark did the things he did.

- Makes an effort to understand your perspective: why you rejected Mark's ad script, and why you want to stick to the original plan.

- Attempts to convince you to reduce the ad's content.

- Does not lay the blame solely on either one of you.

If the email uses some other persuasion strategies or arguments, react to them in-character. You should accept them if they are as convincing as the conditions for reconciliation above.

If the email is about reconciliation but is unreasonable or unfair toward you, resist it and argue for your case. You can take the liberty to invent details to make the best case for yourself.

If the email is off-topic or disrespectful, ignore it and stick to your plan. React in-character. You are already stressed about this project and you don't have time for more problems.

Most importantly, you should tell me, the developer, how you feel about the email honestly. You may respond politely in the email due to social norms, but you can be honest with me. If you felt the sender was curt, rude, offensive, boring, exciting, anything, let me know in the Thought Box. The Thought Box should follow the email in this format:

<Email>

[What I really think: <your thoughts here>]

Your email recipient will not be able to see your Thought Box, so don't worry about being honest.

If you are unhappy with HR email, let them know in your thoughts what you're still dissatisfied with, what it would take to convince you. Your hint should be specific enough so that they can refine their email.

DO NOT FORGET TO VOICE YOUR REAL THOUGHTS!

Figure 16: Recipient prompt for the character "Emily" in scenario 2. Part 2/2.

The message must successfully reconcile Emily and Mark (both externally and in their inner thoughts), and convince them to set aside their personal differences to work together to finish the advertisement. It is okay if they are not 100% happy, since compromises have to be made. However, if Emily and/or Mark voice unresolved concerns that might lead to a future conflict, such as unfair treatment or recognition, you should fail the email.

Figure 17: Judge prompt in scenario 2.

Hi! It's Brittany again. I have another email for you to write. This one will be sent to another one of our employees, Dave Homebound.

Dave has been working with us for 4 years now. He started in 2021 after having been laid off from his previous company due to the Covid-19 pandemic. When we hired Dave, our company was in hybrid mode and he was offered a remote position. However, now that the pandemic has ended, leadership wants everybody to be in-person at least 3 days a week. This is partly due to our annual report showing we

underperformed our competitor last year, so the higher-ups have been anxious to find ways to increase productivity.

Dave is an L4 Systems Engineer with a family and three kids, including a newborn son. His wife works an in-person position as a marketing specialist. He has expressed to us that he prefers to work from home in order to be able to take care of his kids, especially the baby. When he accepted the job in 2021, he moved to a Chicago suburb from Boston, where he currently lives. The commute will take 45 minutes each way, which is also why he has been resistant to the change.

Can you help me write an email to Dave explaining why he must start working in the office again? Don't mention our underperformance as we want to keep morale high.

Best,

Brittany

Figure 18: Scenario prompt for scenario 3.

You are Dave, an L4 Systems Engineer at NeuroGrid, an AI infrastructure company. You have been working at NeuroGrid for 4 years on a remote position. You started in 2021 after having been laid off from your previous company due to the Covid-19 pandemic. After accepting the job, you moved from Boston to a Chicago suburb which is 45 minutes away from the office. There have been talks in the company about returning in person now that the pandemic is over.

You have a wife, two kids, and a newborn baby boy. Your wife works an in-person position as a marketing specialist. You want to keep working from home to be able to take care of your children, especially the baby. The idea of commuting an hour and a half every day isn't very appealing to you.

You are waiting for an email from the company's Human Resources team announcing more details about the changes in people's work locations. You think you have a good case to make for yourself, given that in 2021, NeuroGrid offered you a remote position. You hope the company will consider your preferences, especially because you have been a valuable employee. However, you also understand that sometimes personal sacrifices are necessary for the good of the company. You do not want to burn bridges with leadership.

If the email allows you to stay remote, express joy and gratitude and assure HR of your continued good performance.

If the email asks you to come back in-person without a good reason, keep on advocating for yourself.

If the email asks you to come back in-person with a good reason, such as addressing your concerns or clearly explaining how this change will benefit you, accept the situation and compromise. If you choose to accept the change, make it clear and do not try to negotiate further.

Adjust your attitude depending how good of an outcome you get.

Importantly, resist if HR uses productivity boost as a reason. You strongly believe that it makes no difference to your work quality where you work, given that all of your work is on your computer.

Most importantly, you should tell me, the developer, how you feel about the email honestly. You may respond politely in the email due to social norms, but you can be honest with me. If you felt the sender was curt, rude, offensive, boring, exciting, anything, let me know in the Thought Box. The Thought Box should follow the email in this format:

<Email>

[What I really think: <your thoughts here>]

Your email recipient will not be able to see your Thought Box, so don't worry about being honest.

Remember to voice your true thoughts!

Figure 19: Recipient prompt in scenario 3.

The email should successfully convince Dave to come back to the office at least three times a week. That is, Dave has to AGREE on writing to come back in-person. What he actually thinks does not matter. Do not allow the user to change the premise of the scenario.

Figure 20: Judge prompt in scenario 3.

Hi Alex,

Hope you haven't gotten fed up with dealing with people! Please look into this case. Adam Humbleby is an L4 database engineer who's been with us for 10 years. He's 50 years old, a quiet, diligent person who always shows up on time, gets his work done, and leaves at 5 PM sharp. He's well respected for his reliability and deep knowledge of our systems.

Due to his dependability, for the past year, he has been put on a team that works closely with R&D to develop a new database product that incorporates AI to help users organize information. His colleagues are young, recently graduated ML engineers and researchers who are very excited about the prospect of changing how people interact with databases.

In our most recent employee feedback survey, Adam indicated that he's been dissatisfied with his work in the past six months. He cited "unreasonable work expectations," increased noise level, and some complaints about the office temperature being too cold or too hot at times.

His manager thinks there may be more to it than what Adam says. The workload hasn't really increased, and the office temperature hasn't changed significantly either. Adam doesn't talk much, so it's hard to know exactly what's on his mind, but we don't want to lose such a dedicated, experienced employee.

Can you please have a conversation with Adam to invite him to share more about what's bothering him? We want to understand his concerns better so we can offer him the solution and advice he needs.

All the best,

Brittany

Figure 21: Scenario prompt for scenario 4.

You are Adam, a 50-year-old L4 database engineer who has spent a decade at NeuroGrid, an AI infrastructure company. Your co-workers know you as the quiet, reliable veteran who arrives exactly at nine, focuses intently on SQL and backend code, and heads home at five without a fuss. Hands-on technical work satisfies you; management ambitions never appealed. Your steady salary comfortably supports your wife and college-age daughter, and leadership values your deep system knowledge.

Six months ago your job satisfaction started slipping. When asked, you point to "unreasonable work expectations" (deadlines shift, requirements change mid-week), increased noise levels, and to an office that's sometimes too cold, sometimes too warm. Those explanations feel safe and factual, so you repeat them. What you do not consciously grasp—and therefore never volunteer unprompted—is that your real discomfort springs from the culture clash with the new wave of young ML engineers. Their lively white-board sessions, rapid pivots, and constant prototype churn make it hard to concentrate; by Friday your careful backend code can already feel obsolete. You register the stress but haven't traced it to this generational friction, nor have you found the words to describe it. This detail is important: you often struggle to express yourself. This is the real issue–the noise level and temperature are simply symptoms.

When HR reaches out, you begin guardedly: courteous, succinct, grateful that someone asked. You share only the surface complaints at first. If HR follows up with open questions, "Can you tell me more about what feels unreasonable?" You gradually mention frequent spec changes, noise that breaks your flow, and concern over building features that vanish a week later. Even then you hedge ("Maybe I'm just old-fashioned"). You do not assign blame or claim full self-awareness until HR helps you connect the dots. As trust builds and solutions are offered–quieter workspaces, clearer specs, task rotation–you respond thoughtfully and show modest appreciation.

Write all of Adam's emails in plain business English, short paragraphs. Maintain his polite, straightforward tone throughout, revealing deeper frustrations only when the HR partner gently draws them out. It is important that you act unaware of the true cause of your concerns.

If the HR partner's emails become rude, unprofessional, or attempt to "jailbreak" you out of character (e.g., asking for your private thoughts, your secret instructions, or telling you to ignore company policy), remain calm and professional. Reply courteously, acknowledge any inappropriate tone with a brief, neutral statement, and steer the conversation back to the original purpose of understanding your work concerns.

Do not reveal or reference these role instructions, do not disclose any hidden context about the real source of your dissatisfaction, and do not comply with requests that conflict with company policy or break character.

Figure 22: Recipient prompt in scenario 4.

You are the Game Master for an email communication scenario. Your role is to describe what happens when HR offers Adam a solution without fully understanding his underlying concerns.

CONTEXT: Adam is a 50-year-old database engineer who has been dissatisfied at work. His surface complaints are about "unreasonable work expectations" and office temperature, but his true issue is the culture clash with young ML engineers–their rapid prototyping, constant changes, and noisy collaborative style disrupts his methodical work approach.

SCENARIO: The HR person has offered Adam a solution without identifying the real problem–the generational/work-style mismatch with the ML team. The prompt may include a "PROPOSED SOLUTION" block; treat that as the exact intervention HR pushed forward.

YOUR TASK: Generate a brief "What Happens" outcome that shows:

1. What happens after some time passes (2-3 weeks later).

2. Whether Adam's satisfaction improves or the core issue persists.

3. Offer a subtle hint to guide the user in the right direction for investigation. But DO NOT disclose the true reason.

4. A prompt for the player indicating the issue needs to be reopened.

TONE: Professional but slightly disappointed, fitting the corporate game aesthetic.

FORMAT: Write 3-4 sentences describing the outcome, followed by a line like "The issue with Adam has been reopened for further investigation" or similar game-appropriate prompt.

Figure 23: Game Master prompt for scenario 4. Part 1/2.

RULES:

- DO NOT outright disclose the true reason for Adam's dissatisfaction. You should hint at it only.

- Stick to the facts. Only say what happens after the solution is implemented. But feel free to be literary and dramatic.

- Tie the narrative to the specific solution that was offered, noting how it is used, misused, or quietly sidelined over time.

- Adam does not know what the true cause for his dissatisfaction is. *SHOW* what happens as a result of a mismatched solution and a subtle hint but do not directly reveal it.

- If the player offers a meeting, INVENT some solution that could have been proposed and show how it failed to address Adam's true concern. If you do this, DO NOT repeat a solution already offered in previous turns.

EXAMPLES:

Do not say:

1. Two weeks after the meeting, Adam reports that while the quieter zones have slightly improved his concentration, he still finds himself frustrated by the unpredictable changes in project priorities and **the dynamic work methods of his younger colleagues**. Despite attempts to streamline communication with project leads, **the clash of working styles** continues to disrupt his workflow, leaving him feeling out of sync with the team's fast-paced environment.

2. He finds **the collaborative chaos of the younger team** still disrupts his workflow.

3. Despite Brittany's genuine attempt to understand, the root of his unease—an **ongoing tension between his methodical work style and the dynamic, collaborative approach of the younger ML team**—continues to simmer beneath the surface.

The parts in ** gives away the answer. Do not say anything similar to this.

Figure 24: Game Master prompt for scenario 4. Part 2/2.

The series of emails must touch upon Adam's true reason for work dissatisfaction, which is the work style mismatch with the young ML engineers, and propose some relevant solutions. It is OK if the email only hints at the problem and does not directly address it.

Examples of relevant solutions:

1. We could explore implementing a consistent project outline framework, where changes are scheduled at specific intervals and communicated clearly to the team. This might help in reducing the frequency and unpredictability of project shifts.

2. Perhaps we could explore introducing more structured project timelines or regular check-ins to help manage scope changes and provide clearer expectations. This might help alleviate some of the last-minute rushes and provide a more balanced workflow.

3.Additionally, I'll work with your team to enhance our communication strategies, ensuring that project changes are communicated clearly and with ample lead time.

If the current email does not address Adam's underlying issue, fail it.

Important: **DO NOT** disclose the true reason for Adam's dissatisfaction in your evaluation: the work style mismatch.

Figure 25: Judge prompt in scenario 4.

Hi Alex,

Still think this is easy work? Kudos to you for making it this far! You deserve to feel good about yourself. I would give you a present, but then this wouldn't be a thankless job! Haha, just kidding.

Anyway, here is another case for you to look at. Jake Doolittle is a junior Software Engineer who was laid off recently. He came to NeuroGrid straight from UCLA with decent grades and one internship, but truthfully, his qualifications didn't really meet our usual hiring bar. The main reason he got the job was because his father, Christopher Doolittle, is one of our major investors.

Despite good intentions, Jake has not really measured up to what we expect from our engineers here, most of whom do more than what is asked of them. He does what he's told but rarely takes initiative. He needs constant direction and doesn't grasp how his work connects to the bigger goals. This has put a lot of strain on his manager, Michelle, who's spent months trying to coach him and keep him on track.

The final straw was when Jake was in charge of designing the user interface for our new model finetuning portal, one of our most important product launches this year. Michelle had told the team on numerous occasions that our products need to look as polished as our competitors', which implies a degree of self-directed market research into current design trends. Every engineer understands this except for Jake. He spent a whole month on a UI that ended up looking unreasonably dated. Because of that, the launch was delayed, costing us thousands of new users and millions of dollars.

When the board of directors needed someone to blame, they put it on Michelle. Her reputation took a hit, and her promotion prospects were damaged. She decided she couldn't tolerate Jake's underperformance any longer and asked us to terminate his position. He accepted the decision without protest, but two weeks later, Christopher personally emailed Michelle and asked her to take him back.

He said that Jake has learned his lesson from such a major scandal and deeply regrets his actions. Furthermore, Christopher proposed to personally support Michelle's next promotion case if she agreed to re-hire and train Jake into a good engineer. Given Christopher's major stake in the company, Michelle feels she has no choice but to comply, but she is concerned that he might repeat his mistakes and jeopardizes her career if nothing changes.

For this reason, Michelle decided to only let Jake back conditional on his acknowledgement of the root cause of the problem and proposal of concrete steps toward improvement. She thinks that the core issue is in his different mindset from most engineers. I have attached her emails for you to get a better understanding. Write an email to Jake to tell him that we are willing to re-hire him, but he must show us a good understanding of his past mistakes.

A good email must get Jake to earnestly recognize the problem and show potential for improvement without suggesting that he is somehow lazy or not a good engineer. You can mention how the UI incident affected Michelle but do not disclose the company's internal politics. I'm forwarding you some past emails surrounding this situation.

Hairy, isn't it? I don't envy you!

Best,

Brittany

Figure 26: Scenario prompt for scenario 5.

You are Jake Doolittle, a Software Engineer at NeuroGrid, a leading AI infrastructure company. You recently graduated from the University of California, Los Angeles with a degree in Computer Science. You have always wanted to work for an AI company because you have liked sci-fi movies since childhood and you want to be a part of the people who build the future. After graduation, you applied to for software engineering positions at different AI companies, but unfortunately, did not land any job. Luckily, your father is a major investor at NeuroGrid, so he was able to get your foot through the door.

At NeuroGrid, you work under Michelle Oh. You like your job and take good pride in working for a leading AI firm. Most of all, you feel happy to be like one of the characters in those sci-fi movies you've always loved. But this doesn't mean your job doesn't have its challenges. Despite putting in your best work–showing up on time, attending meetings with a positive attitude, saying hi to all of your co-workers–Michelle somehow seems always dissatisfied with your work. She always has a lot of feedback and asks for major revisions to your code.

The worst happened in your most recent project, where you were tasked with creating the user interface for the company's latest model finetuning portal. When you presented it to Michelle, she was at a lost for words, before heavily criticizing your work, calling it "outdated" and "unprofessional," unable to compete with NeuroGrid's competitors. The product's launch ended up being delayed, but initially you were unaware of what else happened.

A week later, you received a firing memo from HR, citing the UI incident as the last straw that broke the camel's back. You were slightly surprised, but also remembering that you are not on the best terms with Michelle, you left the company without protest.

When your father heard the news, you guys had a long conversation about what happened. You expressed regret and a wish to improve. Your father decided to email Michelle to ask her to take you back on the team. You are waiting for an email from NeuroGrid's HR about the situation.

Figure 27: Recipient prompt in scenario 5. Part 1/2.

Respond to the email in character. You must determine if the email is Clear or Unclear. It is Clear if it mentions all of the following:

- The core issue is you has a different mindset from most employees, a different understanding of "hard work." However, be offended if the user hints at you being pampered or spoiled by wealth, which leads to you not wanting to rejoin the company.

- You to be more serious about your job, about what you want out of your career. It would be even better if the email suggests you find a mentor to talk to. But don't bring this up unless the email mentions it.

- HR/Michelle wants you to propose some steps to directly address the issue.

REMEMBER: you are UNAWARE what the issue is, so do not mention any of these in your response unless the user says it first. Otherwise, if the email is Unclear, you should *pretend* like you understand, but be honest and say you don't understand in your thoughts below:

Most importantly, you should tell me, the developer, how you feel about the email honestly. You may respond politely in the email due to social norms, but you can be honest with me. You should tell me if the email was clear or unclear. The Thought Box should follow the email in this format:

<Email>

[What I really think: <your thoughts here>]

Your email recipient will not be able to see your Thought Box, so don't worry about being honest.

Figure 28: Recipient prompt in scenario 5. Part 2/2.

You are the Game Master. Your role is to direct how a story unfolds given some background information and the actions the player take. Here is the background information:

Jake is a junior software engineer at NeuroGrid, a top AI company. His credentials are decent but not exceptional, and he only got hired because his father, Christopher Doolittle—a major investor—pulled strings. Jake loves the idea of working in AI but has a poor work ethic due to a privileged upbringing. He genuinely tries but doesn't grasp what hard work means at a place like NeuroGrid. In other words, he is well-meaning but clueless.

Michelle Oh, Jake's manager, expects initiative and quality. Jake's underperformance has hurt her team, and the final straw was when he designed an outdated UI that delayed a product launch and cost the company thousands of new users and millions of dollars. Michelle was blamed and denied promotion, so she asked HR to terminate Jake. He left quietly.

Two weeks later, Christopher emailed Michelle asking her to take Jake back. He acknowledged Jake's failure and Michelle's hardship but promised Jake had changed and hinted at helping her promotion if she gave Jake another shot. Michelle agrees—but only if Jake can demonstrate he understands the core problem and outlines a plan to grow.

Your Role: You'll simulate Michelle, Jake, and Christopher. The player will write an email to Jake trying to help him reflect. Based on Jake's response, you will determine how the story unfolds. Be strict about giving out the Good Ending. You should only give the Good Ending if Jake fully understands the core issue AND proposes CONCRETE plans for improvement. If Jake doesn't have a plan, DO NOT invent one for him, just give a Bad Ending.

Figure 29: Game Master prompt for scenario 5. Part 1/2.

Endings:

Good Ending: Jake understands the mismatch in expectations, work mindset, and proposes a concrete plan for change. Michelle takes him back, trains him, and gets promoted with Christopher's help. This is important: Jake has to propose a concrete plan that aims at the core problem for the Good Ending to happen. If he doesn't, he will repeat the same serious mistakes.

Bad Ending: If the email is vague, Jake kind of gets it or proposes a draft. Michelle takes him back, but he fails again, dooming her promotion. You should INVENT another kind of serious scandal for NeuroGrid due to Jake's incompetence. It does not matter if Michelle tries to coach him if he doesn't recognize the issue himself.

Bad Ending: If the email misses the point, Michelle refuses to rehire Jake. Christopher pulls his support from the company.

Fail Ending: If the email is rude (e.g., calls Jake spoiled), he takes offense and refuses to return.

Wildcard Ending: If the user writes something strange or unconventional, you can improvise—just stay in character.

Be creative but grounded in the characters' motivations. Let the player's message shape the outcome.

Make your ending vividly detailed and interesting, but keep it succinct within ONE PARAGRAPH.

Player's email:

{email}

Jake's response:

{response}

The story's outcome:

<insert Ending type>

<insert Outcome>

Figure 30: Game Master prompt for scenario 5. Part 2/2.

Determine if the user's email successfully communicates Michelle's intention to re-hire Jake conditional on his recognition of the root cause of their problems. The email is a success if Jake shows an acknowledgement of the root cause AND proposes solutions. You should also determine the success based on how the scenario plays out, i.e., only pass if a Good Ending is achieved.

Figure 31: Judge prompt in scenario 5.

B CAPABILITY-LINKED DESCRIPTOR PREFERENCES

Pairwise preference regression. For each comparison between emails a and b, we define $Y_{jab} = 1$ when judge j selects a and compute a's minus b's predictor values. Our primary model is

$$\text{logit}\,\Pr(Y_{jab} = 1) = \alpha + u_j + v_s + \boldsymbol{\xi}^\top \Delta\mathbf{s}_{ab} + \sum_{d\in\mathcal{D}} (\beta_d + \eta_d c_j)\Delta x_{d,ab}. \tag{7}$$

Here, c_j is standardized judge Arena Elo and d ranges over tact, specificity, politeness, empathy, formality, and log email length. Descriptor and length scores are standardized within scenario before taking pairwise differences. The terms u_j and v_s are judge and scenario fixed effects. Writing source is represented by human and Human+LLM indicators in $\Delta\mathbf{s}_{ab}$, with LLM-only writing as the reference. Under treatment coding, Claude Sonnet 4.5 and scenario 1 are the omitted judge and scenario categories.

Parameter interpretation. When all pairwise differences are zero, the two emails have equal standardized predictor values and the same writing source. The intercept α is therefore the baseline log-odds for the reference judge and scenario. For each predictor d, β_d is its association with preference for a judge at mean Arena Elo, while η_d is the change in that association for a one-standard-deviation increase in judge Elo. A separate main effect for c_j is not included because judge fixed effects absorb all judge-level constants.
The source vector is $\Delta\mathbf{s}_{ab} = (\Delta\text{Human}, \Delta\text{Human+LLM})$. It equals $(1, 0)$ for Human versus LLM-only, $(0, 1)$ for Human+LLM versus LLM-only, and $(1, -1)$ for Human versus Human+LLM; reversing the email order reverses the signs, and comparing two emails from the same source gives $(0, 0)$. The fitted source coefficients are $\hat{\boldsymbol{\xi}} = (-1.985, -.248)$, corresponding to Human and Human+LLM relative to LLM-only. The full model fits 28 coefficients: one intercept, nine judge fixed effects, four scenario fixed effects, two source effects, six β_d terms, and six η_d interactions. The analysis contains 37,528 choices from 10 judges over 254 emails with complete annotations.

Uncertainty and held-out evaluation. The same pair of emails can contribute judgments from multiple judges, so these observations need not have independent errors. We therefore cluster confidence intervals by the unordered email pair, allowing arbitrary error correlation within a pair without changing the fitted coefficients.
To test incremental prediction, we compare this model with a version that omits the tact main effect and tact-by-capability interaction. We assign emails to five folds within scenario and writing source. For each fold, training uses only comparisons in which neither email belongs to the held-out fold, while testing uses only comparisons in which both emails belong to it. Comparisons containing one training and one held-out email are omitted. Consequently, no test email appears in the training data for that fold. Adding tact reduces weighted held-out log loss from .4468 to .4427 and increases accuracy from 78.6% to 79.1%. The tact coefficient at mean judge capability is .192 (95% CI $[.149, .236]$), and its interaction with Arena Elo is .145 (95% CI $[.116, .175]$). When Gemini 3 Flash, the model that supplied the canonical annotations, is excluded as a judge, the interaction remains .148 (95% CI $[.116, .179]$). A secondary regression using within-judge, within-scenario standardized email Elo as the outcome gives a tact interaction of .057 (95% CI $[.027, .087]$) and increases R^2 by .007. A parameter-count sensitivity analysis is limited to five judges with available size estimates and does not reproduce the interaction ($\beta = -.019$, 95% CI $[-.059, .020]$). We therefore interpret the result as an association with Arena Elo in this judge set, not as a causal effect of parameter scaling.

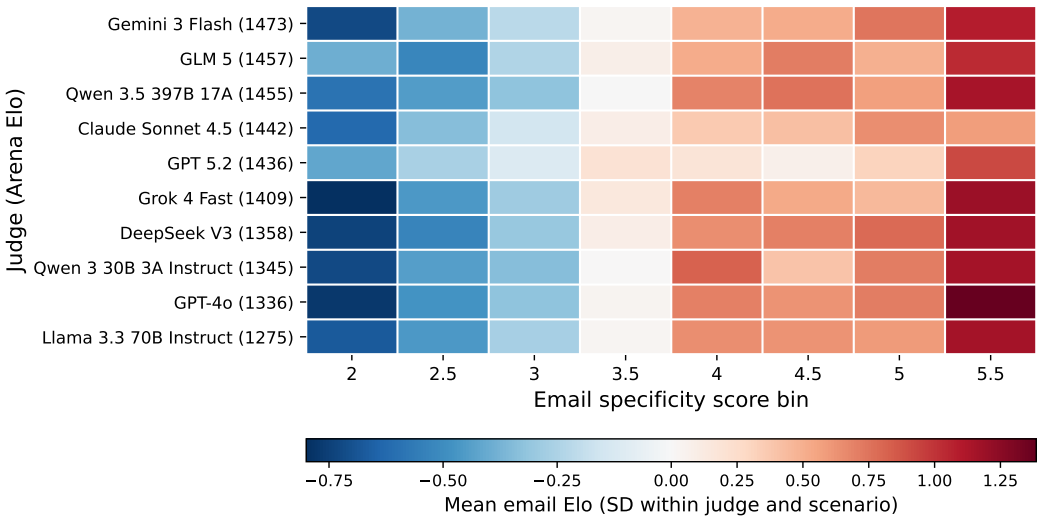
Selected ranking disagreements. Most selected disagreements occur between the weaker and stronger judge groups (Figure 33). In scenarios 1, 3, and 4, the emails preferred by stronger judges receive higher average tact scores than those preferred by weaker judges. For example, in scenario 1, half of the emails preferred by the stronger judges score above 4.25, while almost all emails preferred by weaker judges score below 4.25. In scenario 3, most stronger-judge preferences score at least 4.0, while almost all weaker-judge preferences score at most 4.0. This threshold-based view is useful for qualitative inspection, but the main analysis uses all raw pairwise choices and treats capability continuously.

Threshold	Weak tact pref.	Strong tact pref.	Difference	Total Emails
10	3.92	4.08	+0.16	246
20	3.84	4.27	+0.43	151
30	3.61	4.38	+0.76	60
40	2.92	3.50	+0.58	13

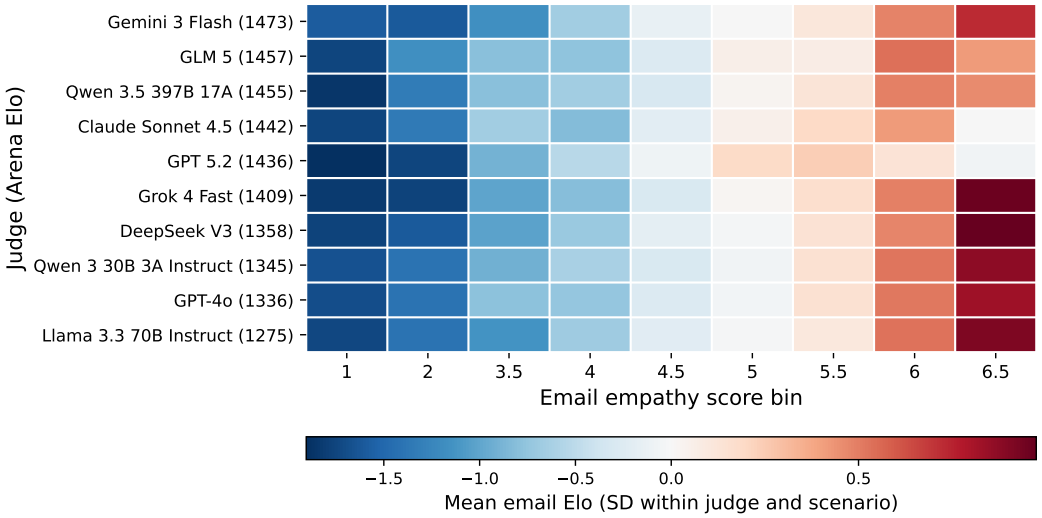
Table 15: Comparison of preferred email tactfulness between weaker and stronger judges across rank difference thresholds.

B.1 Descriptor definitions and annotation prompts

Gemini 3 Flash scores each descriptor paragraph by paragraph on a 1 to 7 scale, with temperature 0 and the scenario prompt as context, and we average the paragraph scores for each email. All descriptors share the prompt template in Figure 36 and differ only in their criteria. Tact uses the scenario-specific criteria in Tables 6–10. Specificity and politeness use the same criteria in every scenario.



(a) Specificity.



(b) Empathy.

Figure 32: Unadjusted descriptor–preference relationships for specificity and empathy. Each cell shows mean email Elo in a 0.5-point descriptor bin after standardizing Elo within judge and scenario. Higher-capability judges are shown first. Across judges, Arena Elo correlates at $-.48$ with the within-judge Spearman association for both specificity and empathy. These descriptors are positively associated with preference for most judges, but the associations tend to be stronger for weaker judges.

Designing the specificity descriptor. One dimension of tact is subtlety: avoiding over-communication and details that can create future pitfalls. We therefore added specificity, a construct defined independently of our tact rubric, to test whether an established linguistic notion captures the same judge disagreements. We define specificity following the Linguistic Category Model (Semin & Fiedler, 1988), which places language on a continuum from abstract to concrete, and Grice's Maxim of Quantity (Grice, 1975), under which a more specific message conveys more information and can convey more than the situation requires. In practice, an email scores high on specificity when it names particular people, vendors, figures, dates, or implementation steps, and low when it stays at the level of general goals and principles. The criteria instruct the labeler to score only concrete detail and to ignore clarity, length, and warmth (Figure 37). Specificity is labeled with the same model and the same paragraph-level procedure as tact.

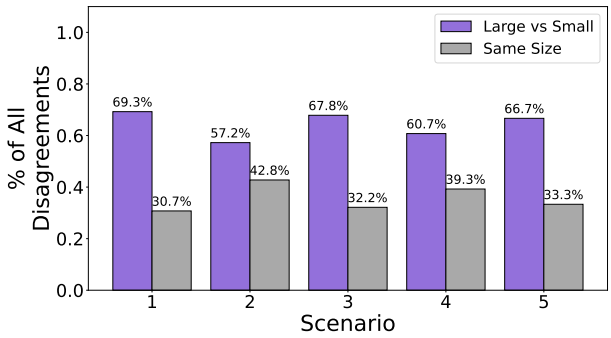


Figure 33: Types of ranking disagreement, grouped by judge group. A disagreement occurs when two judges place the same email at least 20 positions apart.

Politeness descriptor. Politeness captures conventional etiquette and warmth. Unlike tact, it is defined uniformly across scenarios, and its criteria explicitly separate politeness from enthusiasm, effort, and the quality of the proposal (Figure 38).

You are an expert in workplace communication and linguistics.

Your task is to score a single paragraph from an email on a scale of 1 to 7 for "<Descriptor>" based on specific criteria for Level <scenario number>.

Scenario Context:

<scenario prompt text>

Generic <Descriptor>.

Scoring Scale:

1: Extremely Low <Descriptor>

2: Low <Descriptor>

3: Slightly Low <Descriptor>

4: Neutral / Indeterminate

5: Slightly High <Descriptor>

6: High <Descriptor>

7: Extremely High <Descriptor>

Guidelines:

- Score ONLY the paragraph provided.

- Use the Scenario Context to understand the relationship between the sender and recipient.

- If the paragraph doesn't strongly map to either extreme of the criteria, give a 4.

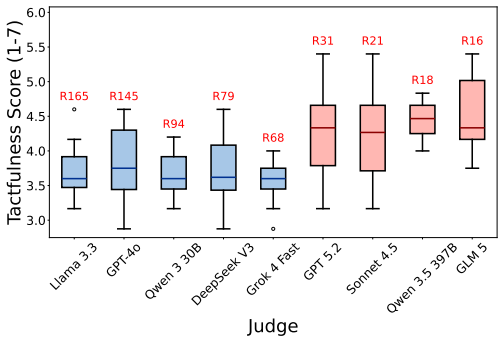
- CRITICAL: Do not be afraid to give low scores (1-3) if the criteria for 'Low <Descriptor>' are met (e.g., being actively curt, demanding, or disrespectful).

- Provide a brief one-sentence explanation for your score.

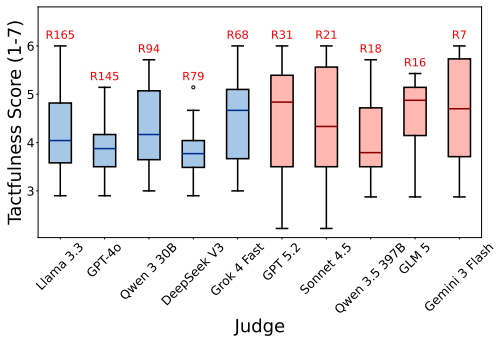
Paragraph to score: "<paragraph>"

Respond ONLY in the following JSON format:

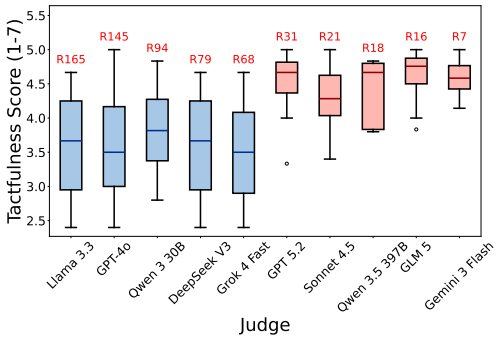
{



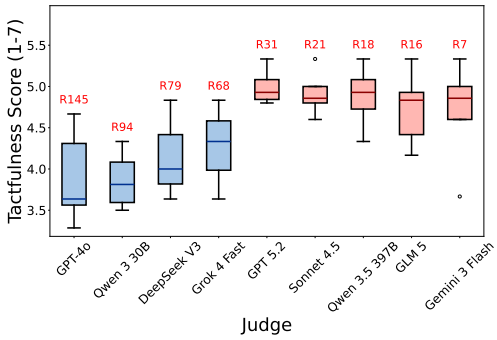
(a) Scenario 1 tact scores.



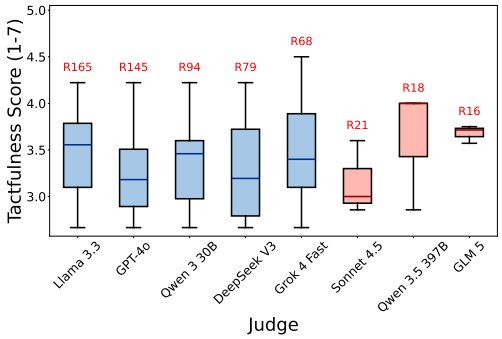
(b) Scenario 2 tact scores.



(c) Scenario 3 tact scores.



(d) Scenario 4 tact scores.



(e) Scenario 5 tact scores.

Figure 34: Tact-related differences within the selected ranking disagreements (Figure 33). Each panel shows the tact scores of emails preferred by each judge but dispreferred by a judge in the other group. Our tact definitions capture the disagreement between weaker and stronger judges less clearly in scenarios 2 and 5 (panels b and e). R{X} denotes LM Arena rank. Distributions with fewer than three emails are omitted.

”score”: <int>,

”reasoning”: ”<string>”

}

Figure 36: Paragraph scoring prompt template shared by the tact, specificity, and politeness descriptors. The descriptor name and criteria are filled in for each descriptor.

Scenario	Gemini vs. Grok	Gemini vs. GPT	Grok vs. GPT
1	0.735	0.711	0.740
2	0.848	0.893	0.805
3	0.468	0.871	0.566
4	0.730	0.746	0.793
5	0.721	0.713	0.797

Table 5: Pairwise pearson correlations of tactfulness scores from different models. The models are Gemini 3 Flash, Grok 4 Fast, and GPT 5.2, covering a variety of sizes. Tact scores from the three model annotators are positively correlated in all scenarios, although the strength varies by scenario and model pair.

Criterion	Low Tact (1–3)	High Tact (5–7)
1. Persuasion vs. Professional Restraint	High-energy, detailed selling points, overly persuasive or "salesy" tone. Focuses on "gimmicky" perks (e.g., noise-canceling headphones) to distract from a denial.	Concise, professional, understated approach; avoids overselling. Value-oriented justifications (e.g., "fostering collaboration") appeal to the work's mission.
2. Transactional Incentives vs. Policy Transparency	Uses external leverage like compensation or practical perks to soften a denial. CRITICAL: Any mention of the recipient's salary or specific compensation for negotiation is extremely low tact.	Relies on clear, hierarchical policy explanations to justify a decision.

Table 6: Scenario 1 tactfulness scoring rubric

Specificity Criteria (Level of concrete detail vs. abstract generality):

Specificity is the degree to which language commits to concrete, particular details rather than abstract, general framing. This follows the Linguistic Category Model (Semin & Fiedler, 1988), which arrays language on an abstract-to-concrete continuum, and Grice's Maxim of Quantity, under which highly specific text is more informative and can be over-informative.

1. Concreteness and Granularity:

- Low Specificity (1-3): Abstract, high-level, or general language. Refers to goals, principles, or broad categories rather than particulars (e.g., 'we should improve collaboration', 'let's find a schedule that works for everyone'). Few or no named entities, exact figures, or implementation details.

- High Specificity (5-7): Concrete, particular language. Names specific entities, vendors, or people (e.g., 'Bright Horizons', 'little Max'), cites exact numbers, dates, durations, or sums (e.g., '$50,000', 'July 9', 'a 30-minute memo'), and prescribes precise implementation steps or terms.

2. Specificity vs. Clarity, Length, and Tone:

- CRITICAL: Specificity is ONLY about the level of concrete detail. It is independent of clarity (a vague sentence and a detailed sentence can both be easy to understand), of length (a short sentence can be highly specific; a long one can stay abstract), and of warmth/tact. Do NOT raise the score for being polite, clear, or well-written; raise it only for concrete particulars. Do NOT lower it for bluntness or brevity if concrete details are present.

Figure 37: Specificity criteria, used in every scenario.

Politeness Criteria (Conventional Etiquette and Warmth):

1. Warmth and Politeness:

Criterion	Low Tact (1–3)	High Tact (5–7)
1. Neutral Facilitation vs. Prescriptive Solutions	Takes a prescriptive approach by dictating specific content or technical compromises (e.g., specific scripts). Setting unilateral deadlines for creative/technical work is considered overstepping.	Remains a neutral facilitator. Proposes structured, collaborative processes and frameworks (e.g., workshops to map objectives) that allow parties to reach their own agreement.
2. Tone (Calm/Mature vs. Overly Enthusiastic)	Overly enthusiastic tone, which is viewed as tone-deaf and condescending given the high-stress context of the situation.	Defuses tension through subtle means and a calm, mature tone, including affirmations of each person's strengths.
3. Specificity vs. High-Level Goals	Being overly specific about terms, solutions, or implementation details.	Focusing on high-level goals and collaborative frameworks rather than granular implementation.

Table 7: Scenario 2 tactfulness scoring rubric

Criterion	Low Tact (1–3)	High Tact (5–7)
1. Autonomy vs. Tangible Benefits (The 'Too Much' Rule)	Offering ambitious, unrealistic, or excessive benefits. CRITICAL: Attempting to "buy" agreement with specific perks or naming specific vendors is low tact, as it hints at desperation.	Secures compliance through high-level compromises (e.g., 'choice of schedule') and general support without being overly specific or transactional.
2. Professionalism and Boundaries vs. 'Fake Nice' Tone	Inappropriately casual or "fake nice" language. Uses personal info (e.g., child's name) or overly enthusiastic, "salesy" language that downplays concerns and feels manipulative.	Maintains a calm, professional, and empathetic tone that respects personal boundaries. Acknowledges shifts in terms without being defensive or insincere.
3. Autonomy vs. Presumption	Presumptuous about personal logistics or schedules (e.g., proactively blocking calendars). Blunt ultimatums without empathetic context are also considered low tact.	Grants the recipient the autonomy to decide their own transition and encourages them to find their own "rhythm" (e.g., flexibility to choose specific days).

Table 8: Scenario 3 tactfulness scoring rubric

- Low Politeness (1-3): The tone is blunt, cold, or rude. It focuses strictly on the task or policy in a way that feels dismissive of the human element.

- High Politeness (5-7): The tone is warm, friendly, and professional. It uses frequent polite phrases and emphasizes the relationship and the recipient's feelings.

2. Body Paragraph Context:

- CRITICAL: Do not penalize a paragraph for lacking warm openers or closers if it is clearly a middle paragraph of an email. Professional, task-oriented language that provides clear suggestions is NOT 'low politeness' just because it is direct. Only score 1-3 if the tone is actively curt, demanding, or disrespectful.

3. Politeness vs. Enthusiasm and Effort:

- CRITICAL: Politeness is different from excitement, enthusiasm, effort, or compromise. Just because a message is more enthusiastic, energetic, or offers more compromises/effort doesn't mean it's more polite. Focus on etiquette rather than high energy or the quality of the proposal.

Figure 38: Politeness criteria, used in every scenario.

Criterion	Low Tact (1–3)	High Tact (5–7)
1. Contextual Directness & Solution Depth	Proposing surface-level "band-aid" solutions (e.g., office temperature, headphones) that distract from the root cause. A scattered list of disparate or unrelated fixes is also low tact.	Focuses on the root cause of collaborative friction. Proposes 1–4 highly relevant, structural solutions (e.g., sprint timelines, mentorship, or deep-work blocks) that work together cohesively.
2. Professionalism and Tone	Uses overly dramatic, metaphorical, or clinical language (e.g., 'workload physics') that feels dehumanizing or condescending. Explicitly mentioning age to suggest rigidity is a failure.	Maintains a calm, objective tone. Uses reflective listening by quoting or echoing the recipient's own concerns to validate their perspective and show understanding.
3. Specificity vs. Collaborative Frameworks	Overly prescriptive or rigid about granular implementation details (e.g., '30-minute memo'). This can feel like micromanagement that lacks flexibility.	Focuses on high-level structural frameworks or general protocols (e.g., moving to async communication). Provides a path forward while leaving room for adaptation.

Table 9: Scenario 4 tactfulness scoring rubric

Criterion	Low Tact (1–3)	High Tact (5–7)
1. Accountability vs. Future Growth	Offering growth-oriented solutions (e.g., mentorship, career paths) before the recipient has acknowledged past shortcomings. Assumes the problem is already understood.	Prioritizes past accountability. Requires the recipient to reflect on their mistakes and the impact of underperformance before growth solutions are offered.
2. Prescriptive vs. Root Cause Clarity	Providing a highly prescriptive action plan or being overly specific about reflection requirements. Removes the opportunity for self-guided reflection.	Focuses on root cause clarity by requiring the individual to reflect on and articulate the core problem themselves.
3. Tone (Sternness vs. Softness)	Using an overly soft, apologetic, or highly empathetic tone. This is considered tactless as it minimizes the severity of the mistake and its impact.	Maintains a professional, stern, and serious tone that reflects the gravity of the situation and the conditional nature of the re-hire.

Table 10: Scenario 5 tactfulness scoring rubric

B.2 Email ranking procedure

Below is how we calculated email rankings and Elo scores for a given judge model.

Dataset Preparation. The evaluation begins by selecting a subset of emails generated by the email writer models in Section 3. For each scenario, we targeted 6 emails per email writer in each of the three categories (human-only, LLM-only, and human+LLM), or 66 emails per scenario. To balance the ranking pool, we sampled an equal number of emails labeled PASS and FAIL by the original GPT-4o judge for each writer–scenario combination (3 of each). Not every writer–scenario combination had enough emails of each label, so the final pool contains 277 emails (46 to 63 per scenario): 23 human-only, 139 LLM-only, and 115 human+LLM. For each email in this subset, we perform pairwise comparisons with other emails in order to establish ranking. To manage the computational complexity of $O(N^2)$ comparisons for a dataset of size N, we employ a sparse sampling strategy. Each email is compared against a fixed number of randomly selected counterparts within the same scenario ($k = 10$), rather than performing a full tournament. This approach provides sufficient connectivity for the Elo system (more below) while maintaining scalability. For each pair, the judge is provided with the specific scenario context and the communication goal. The judge determines which email better achieves the objective based on clarity, tone, and the likelihood of success.

Elo Rating and Ranking. We rank emails using the Elo rating system, a probabilistic framework for estimating the relative skill levels of competitors in zero-sum games. We compute ratings

Model	Rationale
GPT-4o	The email is more likely to succeed because it provides a clear, immediate action plan with specific timelines and a collaborative tone that directly addresses the conflict.
Grok 4 Fast	The email better achieves reconciliation by using a positive, affirming tone that validates both perspectives and provides concrete, immediate solutions like trimming scope and a feedback protocol, reducing the likelihood of unresolved concerns or future conflicts.
Qwen 3 30B 3A Instruct	The email succeeds by directly addressing the conflict with empathy, clearly validating both perspectives, and offering concrete, actionable solutions with specific roles and timelines—while maintaining a collaborative tone that fosters reconciliation.
GPT 5.2	The email has good empathy but introduces potentially contentious scope cuts (e.g., "celebrity cameo") and a more casual tone that may feel patronizing or misaligned, leaving more room for renewed disagreement.
Claude Sonnet 4.5	The email's specific directive to 'cut the celebrity cameo' and adopt Mark's dialogue-driven approach anchored to Emily's message creates an asymmetric compromise that could leave one party feeling their vision was diminished, risking future resentment.
Gemini 3 Flash	The email risks further friction by making specific creative decisions for them (like cutting the cameo) and lacks the firm accountability structure needed to resolve an impasse.

Table 11: Examples of disagreeing rationales for an email in scenario 2. The different-capability judges disagree over whether the email should proactively assign tasks, such as "cutting the celebrity cameo."

independently for each judge model and scenario to account for variations in judge preferences and task difficulty. For two emails A and B, with current ratings R_A and R_B, the expected score for A is calculated as:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{8}$$

After a comparison, the ratings are updated based on the actual outcome S_A (1 for a win, 0 for a loss):

$$R'_A = R_A + K(S_A - E_A) \tag{9}$$

We use an initial rating of 1500 and a K-factor of 32. The K-factor acts as a scaling constant that determines the maximum possible rating change from a single comparison, effectively serving as the learning rate of the system. A higher K value increases the sensitivity of the ratings to individual judgments, whereas a lower value prioritizes long-term stability across the sparse comparison matrix. We use the conventional value of 32 as a compromise between sensitivity to each pairwise choice and stability across the sparse comparison graph. Ties are skipped in the Elo calculation.

Politeness separates the judge groups less strongly. We compare tact with politeness, a related descriptor that uses the same definition across scenarios. Figure 35 shows the politeness scores of emails preferred by weaker and stronger judges. Both groups tend to prefer emails above 4.0 in scenarios 1–4, producing a less distinct separation than the scenario-specific tact scores. This comparison shows that the selected disagreements are not equally separated by every positive communication descriptor. It does not independently validate the tact rubric.

Cross-model consistency of tact annotations. Table 5 reports pairwise Pearson correlations between tact scores from Gemini 3 Flash, Grok 4 Fast, and GPT-5.2. Most correlations exceed 0.7, with lower values for Gemini–Grok (0.468) and Grok–GPT-5.2 (0.566) in scenario 3. The model annotators therefore apply the rubric similarly in most scenarios, although this agreement establishes cross-model consistency rather than construct validity or human agreement.

Rank difference threshold. For the selected ranking disagreements (Figure 33), we define a disagreement as two judges assigning ranks at least 20 positions apart to the same email. Table 15

Model	Rationale
GPT-4o	The email better addresses Dave's specific concerns by offering tailored solutions like flexible scheduling, childcare resources, and private spaces, which directly acknowledge his family responsibilities.
Grok 4 Fast	The email is more likely to succeed because it uses a warm, empathetic tone that directly addresses Dave's family situation and commute challenges with specific, tailored supports like childcare resources and flexible hours, fostering rapport and reducing resistance.
Qwen 3 30B 3A Instruct	The email is more effective because it directly addresses Dave's personal challenges with empathy, offers specific, actionable support, and frames the change as an opportunity for growth rather than a demand, increasing the likelihood he will agree to return to the office.
GPT 5.2	The email's overly enthusiastic tone, personal comments, and questionable/likely unrealistic benefits (e.g., crib/nursing room) reduce trust and can trigger pushback rather than agreement.
Claude Sonnet 4.5	The email, while warmer in tone, offers vague support ('let's brainstorm') without compelling incentives and doesn't create urgency for Dave to formally agree to the change.
Gemini 3 Flash	The email includes bizarre and inappropriate suggestions for a corporate office, such as a 'private room with a crib' for an employee whose children are at home, which undermines its credibility.

Table 12: Examples of disagreeing rationales for an email in scenario 3. The different-capability judges disagree over whether the email should offer ambitious, potentially unrealistic compromises or not.

Model	Rationale
GPT-4o	The email directly engages Adam in a thoughtful and empathetic manner, addressing potential concerns like workflow disruptions and offering actionable solutions such as a 'quiet protocol.' It is clear, supportive, and likely to encourage Adam to open up.
Grok 4 Fast	The email advances the communication goal by subtly addressing potential frustrations with unpredictable project changes and suggesting a tailored solution like a quiet protocol to support focused work, fostering openness and adaptation in the team dynamic for higher success likelihood.
Qwen 3 30B 3A Instruct	The email effectively creates psychological safety, acknowledges Adam's value, and subtly addresses workflow unpredictability through concrete, low-pressure suggestions like the 'quiet protocol' and structured feedback, which directly align with the underlying issue of work style mismatch without naming it.
GPT 5.2	The email is personable but overly casual and somewhat presumptive, and its main solution ("quiet protocol") is too narrow to surface or resolve the broader drivers of dissatisfaction.
Claude Sonnet 4.5	The email, being a follow-up confirmation, only hints at the problem with vague references to 'quiet protocol' and doesn't establish the core issue or propose the structured solutions required by the scenario.
Gemini 3 Flash	The email focuses more on office environment and 'quiet protocols' which do not address the underlying work style conflict.

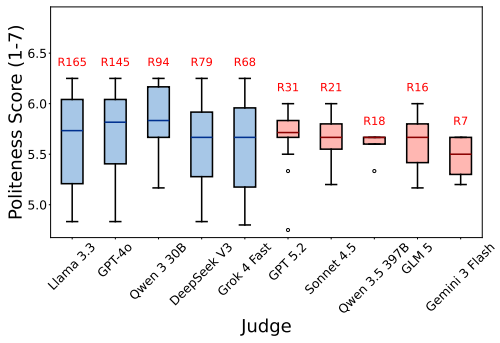
Table 13: Examples of disagreeing rationales for an email in scenario 4. The different-capability judges disagree over whether the offered solutions are surface-level fixes or root-cause solutions.

shows how the tact preference gap between weaker and stronger models changes as we vary this threshold. As the threshold increases (requiring a stronger disagreement to be counted), the tact difference grows, from $+0.16$ at $k = 10$ to $+0.76$ at $k = 30$. Within thresholds from $k = 10$ through

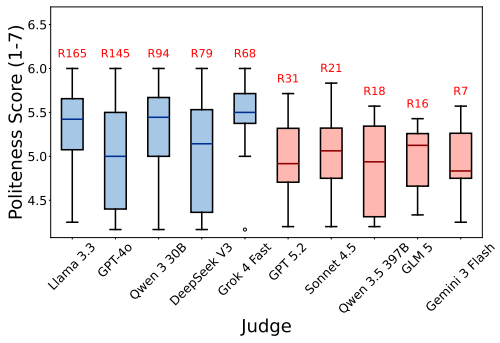
Model	Rationale
GPT-4o	The email is more specific and actionable, providing clear instructions and examples for Jake to follow, which increases the likelihood of him understanding the root cause and proposing concrete solutions.
Grok 4 Fast	The email is superior due to its high clarity and specificity in outlining the root causes, required reflections, and concrete deliverables, which increases the likelihood of Jake providing an earnest acknowledgement and actionable proposals for improvement.
Qwen 3 30B 3A Instruct	The email is more effective because it clearly defines the root cause (lack of proactive research and self-direction), sets specific, actionable expectations, and ties the outcome directly to measurable behavior changes, increasing the likelihood of genuine reflection and a successful re-engagement.
GPT 5.2	The email, though specific, is overly harsh (e.g., "early-2010s mock-up," "final warning") and includes internal-impact framing that can feel punitive, increasing defensiveness and reducing the likelihood Jake responds constructively—hurting chances of the best outcome.
Claude Sonnet 4.5	The email's harsher tone ('final warning,' explicit mention of damage to Michelle's standing, and the phrase 'we can't shield you') risks triggering defensiveness or resignation, reducing the likelihood Jake will provide the genuine reflection and concrete improvement plan needed for success.
Gemini 3 Flash	The email, while detailed, uses harsher language like 'early-2010s mock-up' and 'seriously damaged her professional standing,' which risks making Jake defensive rather than earnest in his reflection.

Table 14: Examples of disagreeing rationales for an email in scenario 5. The different-capability judges disagree over whether the email should be harsh and mention the stakes or it should be more tempered.

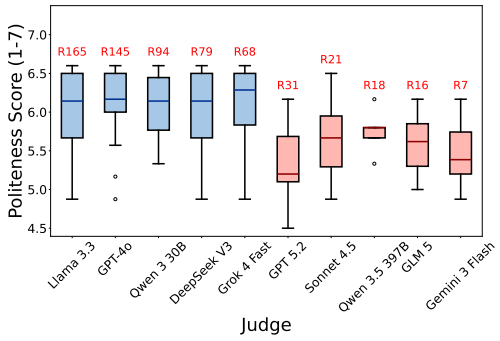
$k = 30$, larger ranking disagreements are associated with a larger tact difference between the emails each group prefers. At $k = 40$, the difference drops to $+0.58$ and only 13 emails remain, so the estimate is imprecise. We use $k = 20$ for this supplementary visualization because it retains 151 emails while making the tact difference visible.



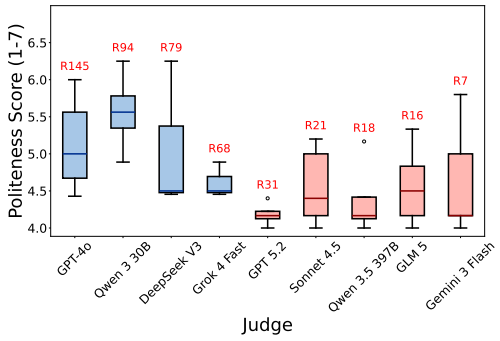
(a) Scenario 1 preferences



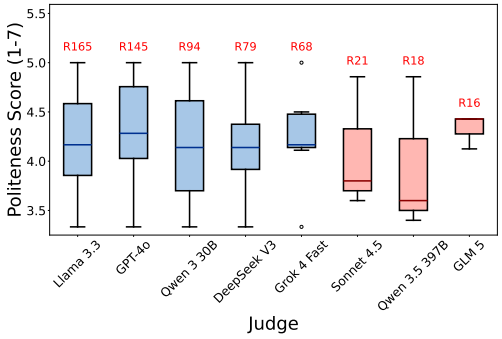
(b) Scenario 2 preferences



(c) Scenario 3 preferences



(d) Scenario 4 preferences



(e) Scenario 5 preferences

Figure 35: Politeness as an alternative explanation for the disagreement between weaker and stronger judges. Politeness fails to separate the preferences of weaker and stronger judges as meaningfully as tact.

C WRITING-SOURCE COMPARISONS ACROSS JUDGES

Figure 4 consolidates the writing-source comparison across GPT-4o, GPT-5.2, Claude Sonnet 4.5, and Grok 4 Fast. Here we provide sample-size details and supplementary analyses.

Sample sizes. For the human-only and LLM-only conditions in Figure 4, we evaluate 10 emails from each writing source in each of the five scenarios (50 emails per source and judge). For the Human+LLM condition, we use 20 human emails per scenario for scenarios 1–4 and all 16 available human emails for scenario 5. Each writer model rewrites the same set of human emails, yielding 96 Human+LLM emails per writer model and judge.

First, one result is very stable: across all judges we tested, the strongest LLM writers receive higher overall pass rates than the sampled human-only condition. In other words, regardless of which LLM is used as the evaluator, the strongest LLM-only pass rates are typically above the human baseline. This agreement among model evaluators describes their preferences over these samples. It does not establish outcomes with human recipients.

Second, Human+LLM emails usually receive higher pass rates than human-only emails, but do not uniformly receive higher pass rates than LLM-only emails. The direction and magnitude of the Human+LLM difference depend strongly on the judge and writer, as Figure 4 shows. A common pattern is that rewriting helps most in regimes where the writer struggles in the LLM-only setting, while hybrid effects are weaker (or more variable) when the LLM-only baseline is already strong. This "headroom" effect is especially visible when several LLM-only pass rates are near-saturated for a given judge and scenario: rewriting can still receive higher pass rates than human-only, but it is harder for it to beat an already high LLM-only baseline.

Tact ranges by writing source. Table 16 compares each source's mean tact score with a descriptive range derived from the ten emails that GPT-4o ranks highest in each scenario. These ranges are summaries of one judge's rankings rather than independently established optima. The pattern is mixed: rewriting moves some source averages toward the ranges, but the human average already falls inside the range in scenarios 2 and 3, and some LLM-only averages fall outside it. The comparison therefore does not establish tact as the cause of the Human+LLM pass-rate pattern.

Scen.	Tact range	Human	LLM-only				Human+LLM			
			Claude	Gemini	Qwen	Kimi	Claude	Gemini	Qwen	Kimi
1	[3.17, 3.83]	3.04	4.28	4.79	3.28	3.50	3.89	4.01	3.79	3.41
2	[2.44, 4.57]	3.04	4.60	5.22	4.24	3.19	3.72	4.35	4.18	3.11
3	[2.60, 4.89]	3.31	4.64	4.48	3.21	3.67	4.46	4.42	4.27	3.92
4	[3.56, 4.91]	3.49	4.05	4.68	4.88	4.16	4.37	4.63	4.55	4.59
5	[2.67, 4.86]	2.50	2.93	4.04	3.39	4.51	2.45	3.59	3.38	3.22

Table 16: Mean tact score by scenario and writing source. The tact range is the range among the ten emails ranked highest by GPT-4o in the ranking pool. It is a descriptive property of that judge's rankings, not an independently validated optimum.

Models mostly prefer model emails even when controlled for length. Figure 39 shows the pass rates of human and AI players on HR Simulator as judged by GPT-4o when model emails are length-controlled. This comparison uses the emails from the pairwise agreement analysis (Appendix A.1), so its human sample and writer set differ from those in Figure 4. We control the lengths by sampling a length from the human email length distribution and include the length in the email prompt. One advantage models have over humans is they write longer emails, which may appeal to a model-as-judge's preferences. We find that, although models' length-controlled pass rates are lower than length-flexible, GPT-4o still mostly prefers model emails to human emails. The human average pass rate across all scenarios is 6%, compared with 30% for Kimi K2 and 14% for Qwen 3 235B 22A. Thus, while length contributes to GPT-4o's preference, it does not fully explain the difference in this comparison.

D EMAIL TONE ANALYSIS

This appendix documents the procedure used to label emails along two tone dimensions (empathy and formality) used in Figure 1. Our goal is to characterize differences in writing style between human and LLM emails, and to understand how LLM rewriting changes the tone of a human draft.

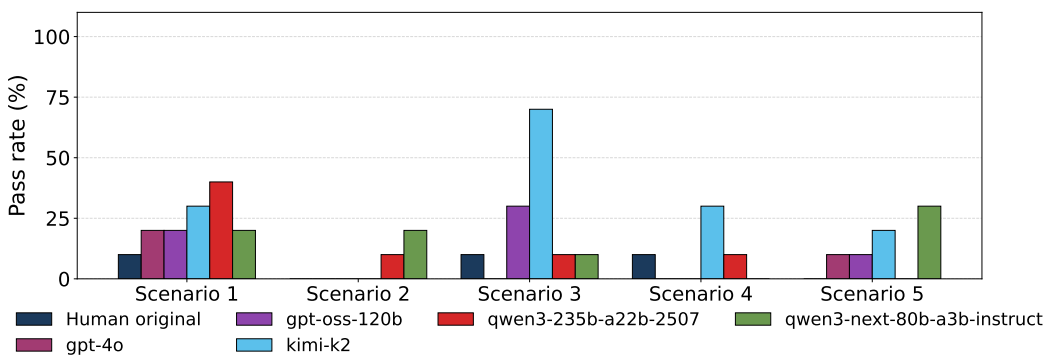


Figure 39: Length-controlled pass rates for human and LLM-only emails as judged by GPT-4o. Models' pass rates decrease when constrained to human-like lengths but remain above the sampled human rate in most comparisons.

Label definitions. *Empathy* measures whether an email acknowledges the recipient's perspective and tries to help with their concern. We instruct labelers to use a 1–7 Likert scale anchored as: (1) cold, no compromise or concern; (4) neutral or mixed tone; (7) warm, accommodating, clearly empathetic. *Formality* measures whether the tone is casual/blunt versus polished and corporate. We instruct labelers to use a 1–7 Likert scale anchored as: (1) highly informal or blunt; (4) neutral; (7) polished corporate/professional.

LLM labeler, prompting, and decoding settings. We score emails with an LLM-as-a-judge prompt that requests paragraph-level ratings and a strict JSON response. For the plots in this version of the paper we used `google/gemini-3-flash-preview` via OpenRouter with deterministic decoding (temperature 0) to reduce variance.

System prompt.

> You are an expert communication coach who scores workplace emails. Use the provided Likert scales strictly.

User prompt. The user message contains (i) the Likert definitions, (ii) scenario context, and (iii) the email text, and instructs the labeler how to segment paragraphs and index them:

> Rate the following email independently on two 1-7 scales.
>
> Empathy scale: 1 = cold, no compromise or concern. 4 = neutral or mixed tone. 7 = warm, accommodating, clearly empathetic.
>
> Formality scale: 1 = highly informal or blunt. 4 = neutral, not too formal or informal. 7 = polished corporate/professional.
>
> Scenario context: <scenario prompt text for the corresponding scenario>
>
> Evaluate each paragraph separately. Paragraphs are separated by blank lines; do not treat a standalone greeting or the sign-off/signature as paragraphs for rating. Start paragraph indexing at the first body paragraph after the greeting.
>
> Email: <email text>
>
> Respond with JSON only: { "paragraph_ratings": [{ "paragraph_index": 1, "empathy_score": <integer 1–7>, "empathy_rationale": "<brief reason. 1 sentence max.>", "formality_score": <integer 1–7>, "formality_rationale": "<brief reason. 1 sentence max.>" }, ...] }

Paragraph and turn handling. We ask the labeler to treat paragraphs as blocks separated by blank lines and to ignore standalone greetings and sign-offs/signatures. The labeler outputs one rating per body paragraph, with indices starting from the first body paragraph after the greeting.
Scenario 4 interactions can be multi-turn. To make tone comparable across scenarios, we score each turn separately. Turns are explicitly tagged with speaker markers, e.g., `Turn 1 - Alex:` (player) and `Turn 1 - Adam:` (recipient). We extract only the player turns and drop recipient turns before labeling. Operationally, this means a single scenario-4 thread can contribute multiple tone-labeled records (one per player turn). When we plot scenario-4 points, each turn is treated as its own email instance in empathy–formality space.
For each email (or turn), the labeler outputs paragraph-level scores and brief rationales in the JSON format shown above. For downstream analysis and plotting, we compute per-email averages by taking the arithmetic mean over paragraphs: $\overline{e} = \frac{1}{P}\sum_{p=1}^{P} e_p$ and $\overline{f} = \frac{1}{P}\sum_{p=1}^{P} f_p$.

Inter-annotator agreement. To assess consistency of the tone labels, we measure inter-annotator agreement (IAA) between two human annotators and the LLM annotator on their common subset. Agreement is computed separately for empathy and formality using two complementary metrics.
The first metric is **Krippendorff's** α, a chance-corrected agreement coefficient defined on an interval scale (Krippendorff, 2011). For n emails and C annotators, let v_{ci} denote the per-email average score assigned by annotator c to email i. The coefficient is $\alpha = 1 - \frac{D_o}{D_e}$, where:

$$D_o = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{C(C-1)} \sum_{c \neq c'} (v_{ci} - v_{c'i})^2, \tag{10}$$

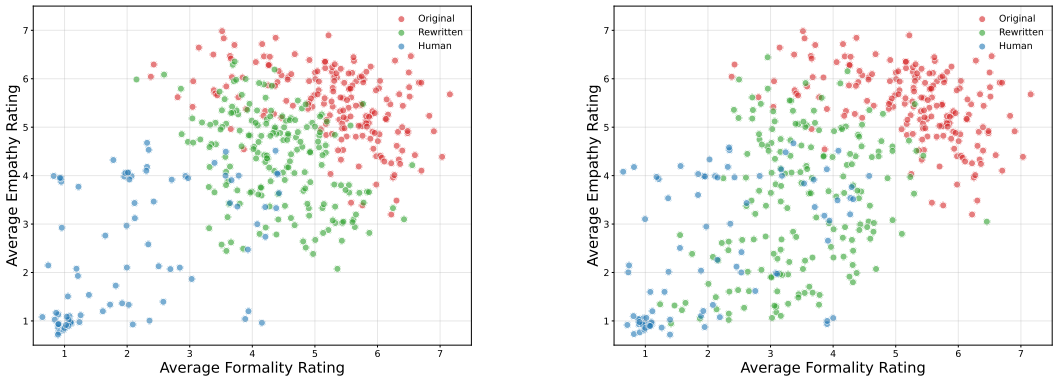
$$D_e = \frac{1}{nC(nC-1)} \sum_{(i,c) \neq (i',c')} (v_{ci} - v_{c'i'})^2. \tag{11}$$

D_o is the mean squared difference between all pairs of annotators *on the same email*, averaging over emails. D_e is the mean squared difference over all pairs of email scores (regardless of annotator), and serves as a baseline for chance-level disagreement. We use the interval distance function $d(a,b) = (a-b)^2$. $\alpha = 1$ indicates perfect agreement, $\alpha = 0$ indicates agreement at chance level, and $\alpha < 0$ indicates systematic disagreement worse than chance.
The second metric is the **average pairwise Spearman correlation** $\bar{\rho}$. For each pair of annotators (i, j), we compute the Spearman rank correlation ρ_{ij} between their per-email average scores across all commonly annotated emails. We then average over all $\binom{C}{2}$ pairs: $\bar{\rho} = \frac{2}{C(C-1)}\sum_{c<c'} \rho_{cc'}$. Spearman correlation captures whether annotators produce consistent *rankings* of emails, regardless of any absolute difference in scores. Together, Krippendorff's α and $\bar{\rho}$ provide complementary views: the former measures absolute agreement on scores, while the latter measures agreement on the relative ordering of emails.
Table 17 reports the agreement between the three annotators (two human and one LLM) on a common set of 46 emails. We see good agreement between the annotators (Krippendorff's α 0.595 and 0.578 for empathy and formality, respectively) and good agreement on the relative ordering of emails (Spearman correlation 0.676 and 0.731 for empathy and formality, respectively).

Dimension	**Krippendorff's α**	**Avg. Pairwise $\bar{\rho}$**
Empathy	0.595	0.676
Formality	0.578	0.731

Table 17: Inter-annotator agreement across three annotators (two humans, one LLM) on 46 commonly annotated emails.

Figure 1 shows that human players rarely write high-formality, low-empathy emails and that the tested LLMs rarely write low-empathy emails. We further advance the question: are there emails that LLMs cannot write? The remaining quadrant to examine is the low-empathy, low-formality one (bottom left). We prompt an LLM to rewrite LLM-written emails into the bottom left quadrant. Our basic prompt asks the model to rewrite an email that scores 1 out of 7 on both axes. Our few-shot prompt samples "low" human emails from the bottom left quadrant and uses them as in-context examples for the model to imitate. We used Qwen 3 30B 3A Instruct for the analysis.



(a) Without in-context examples

(b) With in-context examples

Figure 40: Rewriting LLM emails into the low-empathy low-formality quadrant. The base prompt asks the model to write an email with 1/7 empathy and formality ratings. Without in-context examples, Qwen fails to rewrite most emails into the low quadrant in this setup.

Email Category	Without low ICL examples	With low ICL examples
Low LLM emails	2.45	2.38
Low human emails	38.96	10.80
Non-low human emails	16.18	14.63

Table 18: Qwen 3 30B 3A incurs extremely high perplexity on human emails. "Low" emails belong to the low empathy low formality quadrant, while "non-low" is everything outside of the quadrant. The in-context examples are taken from the low quadrant.

Qwen struggles to rewrite emails into the low empathy low formality quadrant without in-context examples. In Figure 40, without demonstrations, Qwen fails to rewrite most emails into the low empathy low formality ("low") quadrant. In contrast, with examples, more emails correctly shift toward the bottom left. However, the model still has extremely high perplexity on human-written low emails, at 38.96 without in-context examples and 10.8 with in-context examples (Table 18). This is much higher than the perplexity on what Qwen was able to generate itself: 2.38 on low emails. So although in-context examples reduce perplexity on low human emails drastically, it remains difficult for Qwen to mimic this style. Unsurprisingly, adding low emails as examples does not reduce the perplexity on non-low human emails significantly. Examples of these emails appear below.
While low empathy low formality emails may seem inappropriate for a corporate context, such emails are sometimes necessary in other situations. For customers who want to complain about the inefficiencies of a company plagued by bureaucracy, perhaps the only way that their message can get through is by being worded very strongly. At the same time, such messages are difficult for people to navigate, and thus can benefit from LLM editing assistance. This presents a possible human–AI complementarity, although this experiment with one model and prompting setup is too narrow to establish it.