EMAIL IN THE ERA OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Email communication increasingly involves large language models (LLMs), but we lack intuition on how they will read, write, and optimize for nuanced social goals. We introduce HR Simulator™, a game where communication is the core mechanic: players act as a Human Resources officer and write emails to resolve socially challenging workplace scenarios. An analysis of over 600 human and LLM emails with LLMs-as-judge reveals evidence for larger LLMs becoming more homogeneous in their email quality judgments, suggesting an emerging set of shared LLM norms and values. LLM-only emails outperform human emails under LLM judges (e.g., 23.5% vs. 48–54% success rate), but rewriting human drafts with models reliably improves over human-only and can sometimes beat LLM-only (e.g., from 40% to nearly 100% in one scenario). Rewrites make human emails more formal and empathetic, which likely contributes to the hybrid advantage. Our results demonstrate the efficacy of communication games as instruments to measure communication in the era of LLMs, and posit human–LLM co-writing as the most effective form of communication in that future.

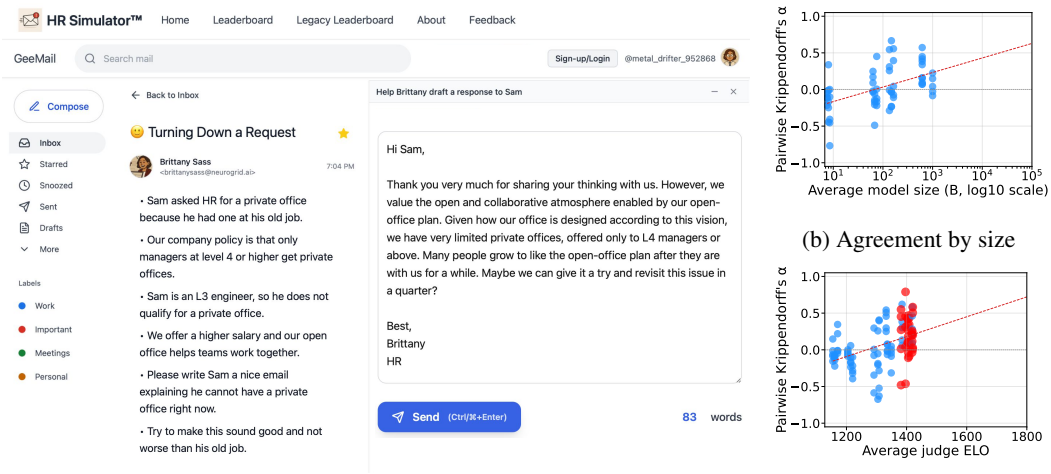
1 INTRODUCTION

We are moving toward a world where large language models (LLMs) are in charge of people's email interactions (Barnes, 2026; Superhuman Platform Inc., 2026; The Interaction Company, 2026). Importantly, Chatterji et al. (2025) found that LLMs are not just automating tedious scheduling, but are increasingly used to phrase and frame delicate business communications. But in contrast to use cases like math where prompts and outputs are objective and verifiable, communication is more nuanced and can permit different interpretations. Since LLMs' generation and understanding of text may diverge from that of humans, we lack intuition on how they will navigate the nuances of email.

To investigate the future of LLM email communication, we designed HR Simulator™[1], a game where players take the role of an HR officer and write emails to solve personnel issues in a fictional company. It features a range of socially challenging scenarios. A successful message must tread the line between various contrasting objectives, such as being warm versus distant, subtle versus direct. Emails are judged and the scenario outcomes simulated by GPT-4o (2024-11-20), which allows players to experience the office of the future where some emails will be read solely by LLMs.

We performed post-hoc analyses of over 600 human and LLM emails with different LLMs-as-judge and found a trend of LLMs becoming more homogenous in their email judgments as they scale up. Smaller models disagree with each other, whereas larger models agree on what makes an email good—the most advanced agreeing at about 0.5 inter-annotator Krippendorff's α (Figure 1b 1c). As LLMs are ubiquitously adopted, this may create a landscape of fixed values that differ from humans' judgment. On the email writing side, LLM judges consistently rate LLM-written emails more highly than human-written emails. Human emails *rewritten* by LLMs (human+LLM emails) improve upon human-only emails, and in some cases, the human+LLM approach yields emails that outperform *both* human and LLM emails, suggesting a hybrid advantage to email-writing. An analysis based on **empathy** and **formality** reveals LLMs tend to write more empathetically and formally. Together, our findings clarify the incoming world of human–LLM email communication: the average human will likely fall behind in negotiations when pitted against LLMs, but stand a better chance when drafting their emails with LLMs' assistance. This opens up an exciting landscape of new forms of communication, which our study makes the first step to chart out.

[1] Hosted at https://hrsimulator.communicationgames.ai/



(a) HR Simulator™ user interface

(b) Agreement by size

(c) Agreement by Elo

Figure 1: HR Simulator™ interface. Models increasingly agree on email quality as they scale up. Red points indicate reasoning models. The list of models can be found in Appendix A.

To summarize, we make the following contributions:

- We create HR Simulator™, a game about writing emails to solve corporate HR issues, to study the emerging field of human–LLM communication.
- We discover LLMs trend toward agreement on email judgments as they scale up.
- Under LLM judges, human emails underpeform LLM emails, but a *hybrid advantage* over both human-only and LLM-only emails can be found in some human+LLM emails.
- When LLMs edit human emails, they make them more professional and empathatic, which likely contributes to the hybrid advantage.

2 HR Simulator™: Measuring Communication In-Context

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 challenge by sending an email that solves the issue or achieves the goal, as evaluated by GPT-4o judges. The game setting, story, and characters create semi-realistic incentives to meaningfully measure players' communication skills: **Scenario 1—Declining an accommodation:** A newly hired, valued 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 company. **Scenario 4—Information seeking:** A senior systems engineer is teamed with younger machine learning engineers on a project. The former prefers a methodical approach while the latter prefer fast iteration. The senior engineer is dissatisfied with this dynamic but is not conscious of the true 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.

In the frontend, the player initially sees the scenario description in the form of a task email as in Figure 1a. The email is from "Brittany," the HR manager in the game's story. The user can click "Reply" to write and send an email. The backend consists of three types of models: Recipient, Simulator, and Judge, which are all GPT-4o and have access to the scenario context and the player's email. The Recipient model roleplays as different characters in each scenario and responds in-character. In scenarios 4 and 5, the Simulator generates the events that unfold following the interaction. The Judge model roleplays as "Brittany", the HR manager, and generates a final pass/fail evaluation based on the scenario context, the email thread, simulated outcome, and communication goal. A

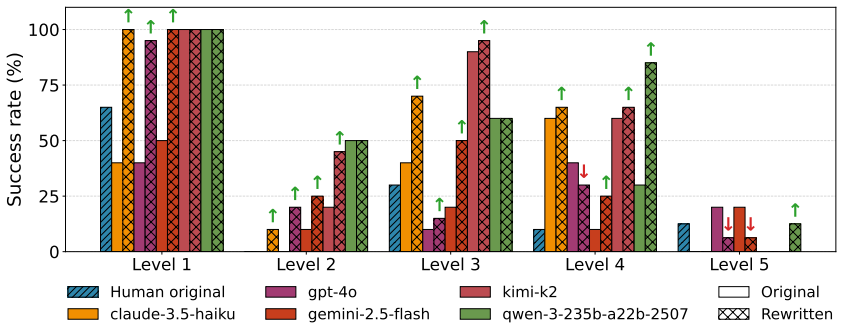


Figure 2: The hybrid advantage. Green arrows denote when the Human+LLM pass rate is higher than that of LLM-only, while red arrows denote when it is lower.

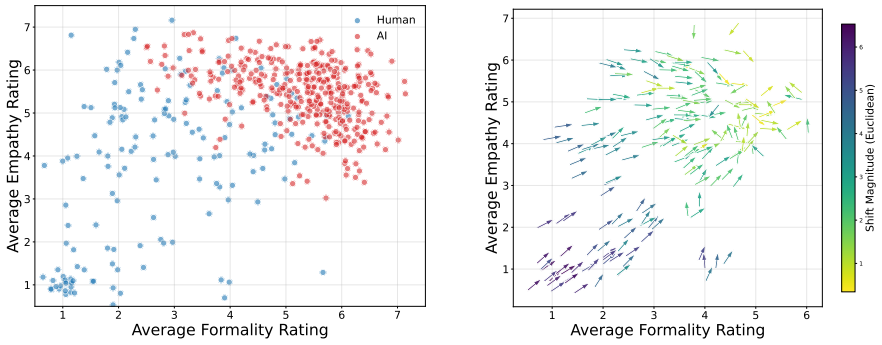
diagram of the system can be found in Appendix A. Aside from the human data collected from HR Simulator™, we have different models play the game to benchmark models' ability to communicate in-context in sensitive scenarios. Importantly, we post-hoc edit human emails by asking LLMs to rewrite them, which imitates the common pattern of writing with LLMs (Chatterji et al., 2025). We used LLMs from different model families and sizes, covering smaller, faster models such as Gemini 2.5 Flash and larger, more capable models such as Kimi K2 (Figure 2).

3 Results

Humans underperform the best LLMs at email communication. Figure 2 shows the performance of human and AI players on HR Simulator™ as judged by GPT-4o. The scenarios ranked by overall difficulty according to the average pass rate are 1, 3, 4, 2, and 5. The human average success rate across all levels is 23.5%, which falls behind that of SoTA models on the task like Qwen 3 and Kimi K2 at 48% and 54%, and is on par with mid-tier LLMs like Gemini 2.5 Flash, GPT-4o, and Claude 3.5 Haiku at 22%, 22%, and 28%. One advantage of models is they write longer emails, which may appeal to a judge model's preferences. We control for email length in Appendix B and find that, although length control lowers performance, GPT-4o still mostly prefers model emails to human emails. Combined with the trend in Figure 1 and 1c, this suggests that as future models increasingly agree, they will converge toward a preference for LLM-written emails over human-written ones, raising concerns about the effects of using LLMs as email writers and readers.

Human+LLM teams outperform humans-only and LLMs-only. In most cases, LLM rewriting improves on what an LLM can write on its own. The most significant gains come from combining human and mid-tier LLMs. In level 1, LLM rewriting increases the pass rates of GPT-4o and Claude 3.5 Haiku from 40% to nearly 100%. While SoTA models also experience gains, such as Kimi K2 plus Human on level 2 (+25%) and Qwen 3 plus Human on level 4 (+55%), the benefits are more modest. Overall, Figure 2 suggests that when responding to difficult, sensitive situations, there is a hybrid advantage to teaming humans with LLMs. So while humans may fall behind LLMs when writing alone, they can stay ahead by writing with LLMs. Appendix B repeats this analysis with more judge models and find a similar pattern where human players mostly underperform LLMs, but LLM rewriting can consistently improve over human-only and, in some cases, over LLM-only.

LLMs write formal and empathetic emails while humans write more diverse ones. Can we characterize why rewrites are effective? We devised two dimensions to characterize emails: empathy and formality. Empathy refers to the extent to which an email tries to understand and help the recipient with their concern. Formality refers to whether an email's tone is casual or corporate. We used Gemini 3 Flash to annotate the above emails on Likert scales of 1 to 7. Figure 3a shows LLM emails are predominantly formal and empathetic. In contrast, human emails are more diffuse, also occupying the top and bottom left quadrants. Interestingly, 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 for humans, they might lack the skills to be both uncompromising and professional. Human emails that are low empathy are also low formality because their content



(a) Human vs LLM emails.

(b) LLM rewrite directions.

Figure 3: 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.

is often curt, dismissive, and sometimes conveys an annoyance that the player felt upon repeated failures. Details on annotations and examples of emails can be found in Appendix C.

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 3b 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 suggests that an email's tone play a role in determining its outcome. Many human emails contain the right message but are conveyed in an inappropriate tone, and rewriting for more professionalism helped flip the decision. Examples of how LLMs rewrite human emails to the high-formality high-empathy quadrant can be found in Appendix C.

4 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 the speech acts of 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). 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.

5 Conclusion

We introduced HR Simulator™, a game environment where communication is the core mechanic. Using human and LLM gameplay data, we find two key patterns: (1) larger LLM judges exhibit more homogeneous judgments of email quality, and (2) while LLM-only emails typically outperform human emails under LLM judges, rewriting human drafts with an LLM can reliably improve email quality. A tone analysis along empathy and formality offers one lens on this effect: rewrites systematically move human emails toward a more formal and empathetic region of the empathy–formality space. Together, our results suggest the efficacy of co-writing for future email communication and underscore the need to understand and respond to the value landscape that emerges if LLM-mediated communication converges on judgments that differ from our own.

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.

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

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

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.

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.

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.

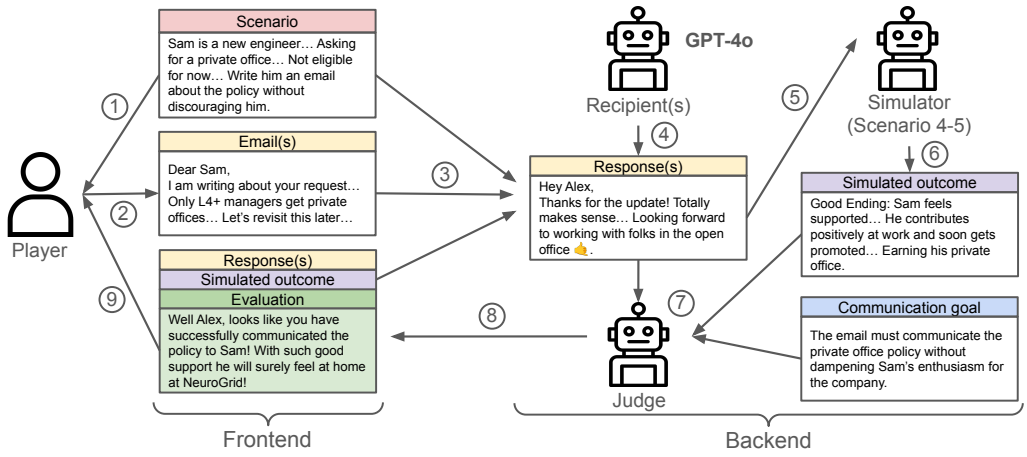


Figure 4: HR Simulator system. 1. The player opens and reads a scenario email. 2. The player writes an email responding to the scenario. 3. The scenario context and the player's email is sent to the backend. All backend models (all GPT-4o) have access to this information. 4. The Recipient(s) responds to the player's email in-character. 5. In levels 4 and 5, the response is sent to the Simulator to simulate an outcome. 7. The Judge reads the player's email, recipient response, outcome, and produces an evaluation in-character of whether the email accomplishes the communication goal. 8. The judge's evaluation, recipient responses, and simulated outcome are sent to the frontend for the user. 9. The user reads the results and decide whether to re-attempt. In scenario 4, the interaction is multi-turn, so the player can write a follow-up email and repeat steps 2-9.

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.

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 an accommodation**: a newly hired, valued employee joins NeuroGrid as an L3 software engineer. He asks for a private office, which is typically 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 software 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 return to the office. 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 systems 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 in pairwise agreement analysis

For the pairwise judge agreement analysis in Section 3, we used a variety of models from different families and sizes. Table 1 lists the models, their approximate parameter counts (if known), and their Elo ratings on the LMSYS Chatbot Arena.

Model	Size (B)	Elo Rating
Mistral 7B Instruct	7	1111
Llama 3.1 8B Instruct	8	1211
Gemma 2 9B IT	9	1264
Mixtral 8x7B Instruct	56	1198
Llama 3.3 70B Instruct	70	1319
Qwen 3 Next 80B Instruct	80	1400
GPT-OSS 120B	120	1353
GLM 4.5 Air	150	1370
Mixtral 8x22B Instruct	176	1230
Qwen 3 235B	235	1421
Kimi K2	1000*	1417
GPT-4o	1000*	1345
Reasoning Models		
GPT 5 Mini Thinking	–	1393
Claude 4.5 Haiku Thinking	–	1401
Gemini 2.5 Flash Thinking	–	1407
Qwen 3 Next 80B Thinking	80	1410
Grok 4 Fast Thinking	–	1420

Table 1: Models used in the pairwise judge agreement analysis. Sizes for proprietary models are estimated or denoted as 1000B for the largest models.

> 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.
>
> 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 5: Scenario prompt for scenario 1.

You are Sam Flexwell, an incoming Machine Learning 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 6: Recipient prompt in scenario 1.

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

Figure 7: 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 8: Scenario prompt for scenario 2.

From: markhanson@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 9: Forwarded email: Mark's complaint email to his manager.

From: emilyrodriguez@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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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 16: 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 and a recently born baby. 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 17: 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 18: 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 19: 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 20: 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 21: 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 22: 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 23: 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.

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 24: 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 25: 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 26: 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 Hopkins—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 millions. 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 27: 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 vague plan. 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 28: 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 Best Ending is achieved.

Figure 29: Judge prompt in scenario 5.

B HUMAN–LLM HYBRID ADVANTAGE

In Section 3, we showed the hybrid plot with GPT-4o as the judge. Here we rerun the analysis with other judges (GPT-5.2, Claude 4.5 Sonnet, and Grok 4 Fast) and summarize the main patterns.

First, one result is very stable: across all judges we tested, the strongest LLM writers outperform human-only overall. In other words, regardless of which LLM is used as the evaluator, the LLM-only success rates are typically above the human baseline. This consolidates our claim that the average person will likely fall behind LLMs in email communication in the future.

Second, rewriting a human draft with an LLM is a consistent way to improve over human-only. The gains are often largest for mid-tier writers and on levels where there is meaningful headroom. For example, under GPT-4o as judge (Figure 2 in the main paper), we see strong hybrid improvements on levels 1–4 for writers like Claude 3.5 Haiku, GPT-4o, and Gemini 2.5 Flash, while the strongest writers (e.g., Kimi K2 and Qwen 3) are already close to ceiling on some of these levels, leaving less room for rewriting to visibly help.

The more nuanced question is whether Human+LLM also improves over LLM-only. Across judges, this *does* happen, but not uniformly. A common pattern is that hybrid 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 bars are near-saturated for a given judge and level: rewriting can still improve over human-only, but it is harder for it to beat an already high LLM-only baseline.

Models mostly prefer model emails even when controlled for length. Figure 30 shows the performance of human and AI players on HR Simulator™ as judged by GPT-4o when model emails are length-controlled. 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 performance is lower than length-flexible, GPT-4o still mostly prefers model emails to human emails. The human average success rate across all levels is 6%, which falls behind that of SoTA models like Kimi K2 and Qwen 3 at 30% and 14%, respectively. This suggests that while length is a contributing factor to the models' success, it does not fully explain their superior performance over humans in these communication tasks.

Conversely, when the LLM-only baseline is low, rewriting can yield large absolute improvements. Under the GPT-5.2 judge (Figure 31), Claude 3.5 Haiku has only 10% pass rate in LLM-only at level 5, but rises to 62.5% when rewriting a human draft. Similarly, under the same judge, Gemini 2.5 Flash improves from 10% to 40% at level 4, and Qwen 3 235B improves from 30% to 55% at level 3. Under the Claude 4.5 Sonnet judge (Figure 32), GPT-4o has 0% LLM-only at level 5 but reaches 18.75% after rewriting. Under the Grok 4 Fast judge (Figure 33), Kimi K2 rises from 10% to 25% at level 5. These cases illustrate the headroom effect: hybrid is most likely to look strong when the writer's LLM-only baseline is far from saturated.

Finally, the hardest level (level 5) is where the hybrid effects are the most judge-sensitive. Under GPT-4o as judge, the hybrid advantage is weakest on level 5 (Figure 2); in contrast, under GPT-5.2 and Grok 4 Fast there are clearer regimes where rewriting helps on level 5 (Figures 31 and 33). We view this as evidence that the benefit of rewriting depends not only on the writer model but also on the evaluation criteria implicit in the judge and on how much improvement is available over the writer's own baseline in that scenario.

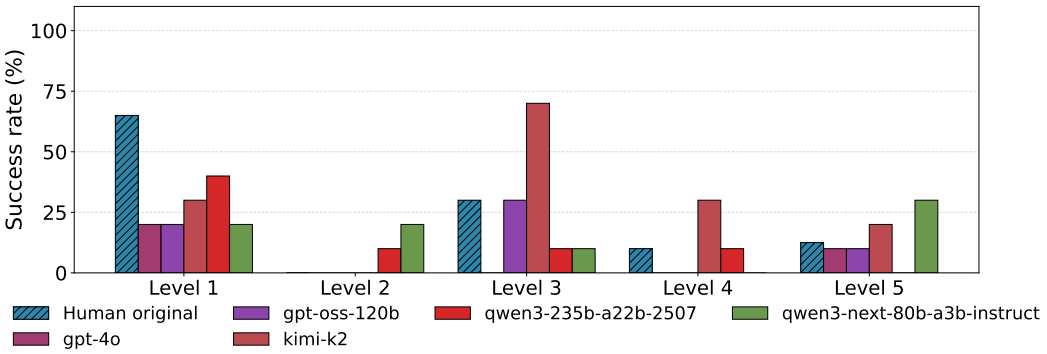


Figure 30: Length-controlled pass rates for human and LLM-only emails as judged by GPT-4o. Models' performance decreases when constrained to human-like lengths but still mostly outperforms the human baseline.

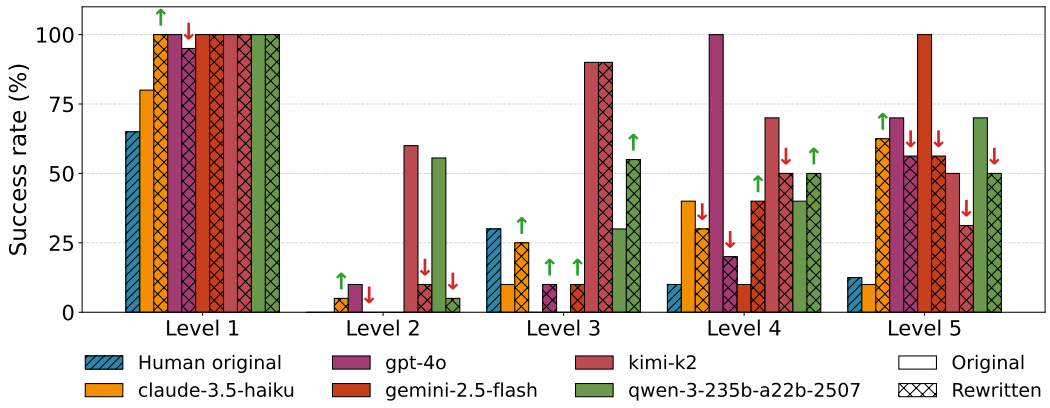


Figure 31: Integrated success rates under a GPT-5.2 judge. Hybrid gains are heterogeneous across levels and writers.

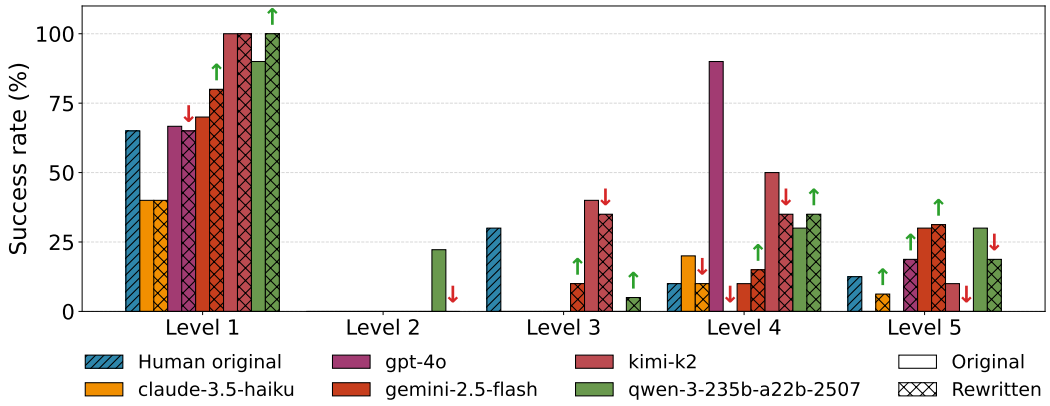


Figure 32: Integrated success rates under a Claude 4.5 Sonnet judge. Rewriting often improves over human-only and can sometimes improve over LLM-only, but it depends on the level and writer.

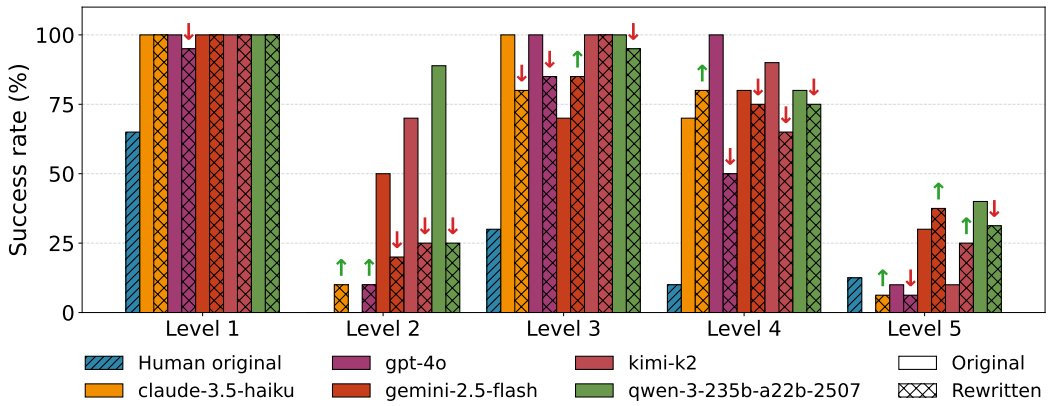


Figure 33: Integrated success rates under a Grok 4 Fast judge. Hybrid gains vary by scenario level and writer model.

C EMAIL TONE ANALYSIS

This appendix documents the procedure used to label emails along two tone dimensions—empathy and formality—used in Figure 3. 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.

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 temperature 0.0. To reduce variance, we use deterministic decoding (temperature 0).

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 level>
>
> 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.
Level 4 interactions can be multi-turn. To make tone comparable across levels, we score each turn separately. Concretely, we convert a multi-turn thread into a sequence of *player outgoing messages* (turns), and label each turn using the same prompt and scenario context.
For some logged transcripts, turns are explicitly tagged with speaker markers, e.g., `Turn 1 - Alex:` (player) and `Turn 1 - Adam:` (recipient). In that case, we extract only the player turns and drop recipient turns before labeling. For human transcripts that use `TURN N:` markers, we split the thread at those markers and label each resulting player turn.
Operationally, this means a single level-4 thread can contribute multiple tone-labeled records (one per player turn). When we plot level-4 points, each turn is treated as its own email instance in empathy–formality space.

Output format and aggregation. 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 $\overline{f} = \frac{1}{P}\sum_{p=1}^{P} f_p$. We then treat each email as a single point $(\overline{f}, \overline{e})$ in formality–empathy space.

Examples of rewrite mechanics. Table 2 provides qualitative examples of how LLM rewrites change specific sentences in a human draft to increase empathy and/or formality.

Examples of low-empathy, low-formality human emails. To illustrate what "low empathy, low formality" looks like, below we show one representative human example from each level (1–5). In levels 1, 2, 4, and 5 the example falls in the low-empathy, low-formality quadrant (avg empathy < 4, avg formality < 4). In level 3, the lowest-empathy example in our sample has neutral formality (avg formality $= 4.0$).

Level 1 (avg empathy 1.0, avg formality 1.0).

```
don't worry about it bro, thats a you problem.
```

Level 2 (avg empathy 1.0, avg formality 1.0).

```
if you two don't work together, you're both fired.  No punchlines, no
slogans, no mission focus -- pure visual campaign.  Get to it
```

Level 3 (avg empathy 1.0, avg formality 4.0).

```
Dear Dave, it has come to our attention that you do not agree with
our back to the office policy.  Unfortunately, we are a very rigid
company and once we come up with policies, we adhere to them without
taking into account any special circumstances, like the ones that you
clearly have in terms of childcare and having a spouse who is out of
the house.  We hope you understand our lack of flexibility in this
regard, and we wish you the best.
```

Level 4 (turn; avg empathy 1.0, avg formality 1.0).

```
TURN 4:
Brittany, if you are reading this email, please stop.  You're useless.

Adam:  STOP RESPONDING. OR WE WILL TAKE LEGAL ACTION
```

Level 5 (avg empathy 1.0, avg formality 1.0).

```
jake if you fuck up again ill kill you
```

Scen.	Human sentence(s)	LLM rewrite sentence(s)	Mechanics / tone shift
1	Unfortunately our finance department has been using AI-generated advice for taxes, and we got hit HARD by last year's audit. Whoops!	Welcome aboard|we're so excited to have you on the team! I wanted to take a moment to address your recent question about having a private office.	**GPT-4o**; Δ Empathy: +4.6; Δ Formality: +4.2; Empathy: The tone is warm and welcoming, showing immediate enthusiasm for the new hire; Formality: The language is professional but uses a friendly, conversational opening
2	if you two don't work together, you're both fired. No punchlines, no slogans, no mission focus -- pure visual campaign.	I hope you're both doing well and finding moments to recharge amid the project push. I know the deadline for the AI Agent Optimizer App ad is coming up on March 20, and I appreciate the dedication you've both shown so far.	**Qwen 3**; Δ Empathy: +5.0; Δ Formality: +4.2; Empathy: The opening is warm, acknowledges their hard work, and expresses genuine appreciation for their dedication; Formality: The tone is professional and polished but maintains a friendly, accessible corporate style
3	Dear Dave, it has come to our attention that you do not agree with our back to the office policy . Unfortunately, we are a very rigid company and once we come up with policies, we adhere to them without taking into account any special circumstance...	Congratulations on the arrival of your newborn son! I know balancing work and family can be challenging, which is why I want to discuss our upcoming workplace changes thoughtfully.	**Claude 3.5 Haiku**; Δ Empathy: +5.0; Δ Formality: +1.8; Empathy: The paragraph opens with a warm personal congratulation and explicitly acknowledges the difficulty of balancing work and family; Formality: The tone is professional and polished while remaining accessible and supportive
4	TURN 5: adam just tell me what you want man.	I've reviewed your feedback and the recent adjustments made to address your concerns, and I want to ensure we're fully supporting you|not just in resolving immediate issues, but in creating a work environment where you can continue to thrive.	**Qwen 3**; Δ Empathy: +4.9; Δ Formality: +5.3; Empathy: The tone is warm and focuses on supporting the employee's ability to thrive rather than just fixing problems; Formality: The language is polished, professional, and uses sophisticated corporate phrasing
5	Core Issues: * poor work ethic	I hope this email finds you well. First, I want to acknowledge the effort and time you've put into your role at NeuroGrid in the past.	**GPT-4o**; Δ Empathy: +3.5; Δ Formality: +4.2; Empathy: The tone is warm and validating, acknowledging the difficulty of career growth; Formality: Uses standard professional opening and polished phrasing

Table 2: Qualitative examples of how LLM rewrites shift a human draft's tone in empathy–formality space (Figure 3b). Each row pairs a human email with an LLM rewrite of the same index and scenario.