Cold Outreach Response Rate - Email Formats

- Dylan Sivori

Do longer or shorter subject lines and email bodies lead to higher response rates for cold outreach messages? This is a particularly interesting question because we are living in an increasingly digital world filled with more and more emails, messages, notifications, and spam. How can we break through all this noise when reaching out to someone we don't know personally? Generally agreed upon best practices along with intuition say that a cold outreach email should be short, anywhere from 25-120 words, with a shorter subject line; but does this really increase response rate? The intervention of this experiment is the four variations of cold outreach emails: short subject, short body; short subject, long body; long subject, short body; long subject; long body. All the emails will end with the question "could you share your perspective on the industry and your experience? This was chosen as the intervention because it is a scalable treatment, allowing us to email many industry professionals to gain a sufficient sample to measure the average treatment effect, it is cost effective, and it is four distinct email variations that mirror how people would normally structure their cold outreach messages. The results that our experiment generated are that a short subject line with a short email body has the strongest average treatment effect, increasing response rate by 10% with a p-value of 0.035. We reject the sharp null hypothesis and find that it is unlikely that we measured an average treatment effect as strong as we did merely by chance. The implications of our findings are that anyone engaging in cold digital outreach should employ a shorter subject line with a shorter email body to increase the chances that they will actually get a response.

Bayesian Influence Diagrams & LLM Performance in Perfect-Information Games

- Jonathan Hernandez

Does including Bayesian Influence Diagrams improve LLM performance in solving perfect-information games?

Various studies have demonstrated that LLMs fail to perform consistently in discrete information games, which require them to divine an optimal strategy, given an existing history of past actions and theoretical future actions - because they are stochastically generating a mean-reverting strategy rather than a strategy specific to the existing context. Ideally, by using game theory, the Nash Equilibrium can be modeled and found for simple games but can only be approximated for games with a high level of complexity - which is true even for artificial intelligence.

We replicated prompts from the paper, “The Illusion of Thinking” and augmented it to include a generalized chess-specific Bayesian Influence diagram designed to approximate toward a Nash Equilibrium. We inserted the prompts into the model’s context while adversarially playing the LLM in a turn-by-turn sequence, such that the user’s moves were taken from a discrete chess engine and input into the prompt’s context to provide the model with a reasonable turn-by-turn game history that didn’t require us to recruit human players, and which didn’t require us to attempt to measure human performance.

We find that LLM subjects did not win at chess more often than existing literature suggests, most likely due to fading token salience within the context window. However, we found that LLMs receiving our treatment prompts were 15% more successful than LLMs in control in the total number of users’ pieces eliminated during the game.

Our results suggest that including directed acyclic graphs into LLM prompts while executing perfect-information tasks may include turn-by-turn efficiency, potentially by providing an additional scaffold to reason across that buttresses the lack of specific causal prompts when LLMs are merely instructed to take on a role, and therefore are often assumed to take on the behaviors of that role. On the contrary, our research suggests that motion isn’t meaning for LLM prompts, and further context engineering research is needed to provide structured decision-making frameworks to weight LLM behavior - going beyond merely establishing role-specific identities and assuming behavior follows.

Empathic Response Variation Across Race & Sex in Historically Slave-Owning States

- Jonathan Hernandez

Does the gender or race of a potential customer affect response rates when requesting catering orders from U.S. states that historically supported slavery?

It is well known that the U.S. fought a civil war over the right to own slaves roughly 160 years ago, and it is widely assumed in popular culture that there remains significant divergence in the experiences of men and women of color in historically slave-owning states.

In an audit study employing blocking, we contacted restaurants and caterers sampled randomly across the blocks of states that composed the “Union” and the “Confederacy”, evenly distributing the identity of the messenger by race, and then by gender, with a standardized vague message stating the number of people to be fed with a moderate budget.

We found that, in historically slave-owning states compared to other states, catering requests from profiles identifying as men and women of color received response rates at a rate that was 7.5 percentage points lower. This observed difference was statistically significant (e.g. 95% confidence interval: [-9.8, -5.2] percentage points; $p < 0.02$). Additionally, their proposed catering budgets were 15 percentage points less likely to be accepted, a finding also statistically significant (e.g., 95% Confidence Interval: [-17.5, -12.5] percentage points; $p < 0.01$).

These results further elucidate the potential contrast in attitudes that exist in the U.S. today. However, we caution that further research is needed to better understand whether response rates may have been impacted by local economic conditions, education, or technology.

Course-RAG Study Assistant and Trust Calibration

- Bjorn Melin

Does adding a course-specific retrieval (vector database) to an LLM study assistant improve learning and reduce harmful overreliance?

Retrieval-augmented generation grounds answers in class materials and is promising for knowledge-intensive tasks, while users frequently over-rely on unguided AI advice.

We randomize students to (a) base LLM or (b) RAG-augmented chatbot linked to slides/readings, plus a brief “verify before accept” prompt; chosen to couple grounding with a minimal behavioral nudge.

The experiment “finds” RAG raises immediate quiz accuracy by 9 points and one-week retention by 6 points, halves acceptance of incorrect AI claims, and improves trust calibration scores without reducing usage.

Implications: Lightweight RAG plus verification nudges can boost learning and curb overreliance, guiding safe adoption of agentic study tools in courses.