Causal Alignment: Augmenting Language Models with A/B Tests

Panagiotis Angelopoulos[1], Kevin Lee[2], Sanjog Misra[3]

[1]Persado.
[2]University of Michigan, Ross School of Business.
[3]University of Chicago, Booth School of Business.

Abstract

We develop a framework for improving the effectiveness of marketing communications by aligning a language model to the causal signal in A/B tests. The core idea is that randomized experiments reveal not just which content performs best, but directions of improvement that can serve as supervision for a generative model. Using within-experiment comparisons, we fine-tune a language model to create an *improver* architecture that transforms lower-performing content into higher-performing content. The resulting model can be composed with any baseline decision-maker, e.g. a human expert, to suggest improvements to new content. We validate our framework through a large-scale field experiment in email marketing spanning 36 new campaigns and 283 million impressions. Subject lines improved by our aligned model achieve click-through rates 33% higher than those from unassisted human experts, while a general-purpose language model with 30 times as many parameters as our fine-tuned model does not improve outcomes. This suggests that alignment to experimental data adds value beyond model scale alone and remains relevant even as general-purpose models become more capable. Our findings provide the first field evidence that aligning language models to A/B tests improves outcomes in real marketing campaigns and show that integrating AI with experimentation can train systems that make better future decisions.

Keywords: Generative AI, LLM, Human-AI Collaboration, Field Experiment, NLP, Decision Support, Email Marketing

JEL Classification: M31 , C93 , L21

This paper was previously circulated as "Value Aligned Large Language Models". We thank Eric Budish, Jennifer Hu, Walter Zhang, and numerous seminar participants for helpful comments. All remaining errors are our own.

1 Introduction

Firms routinely run randomized A/B tests to improve marketing communications such as email subject lines, push notifications, ad copy and product descriptions among others. The standard use of these experiments is typically local. Firms seek to identify the best performer among a small set of variants and deploy it. This practice is well-founded: randomization ensures internal validity, and the winning variant is causally identified as superior among the options tested. However, this leaves most of the value of experimentation on the table.

We posit that A/B tests do more than select winners. Within each campaign, randomization identifies an ordinal relationship between content alternatives. That is, one message causally outperforms another in that specific context. If email subject line y_A outperforms y_B under randomization, then the experiment reveals not only which is better, but a *direction* of movement. This within-context comparison can be used to learn how to move from a worse decision to a better one in the space of relevant content. Aggregated across many experiments, these within-campaign comparisons constitute a large, structured training set for learning generalizable improvement rules. This directional signal is almost entirely discarded by current practice. The question we address in this paper is how to extract and operationalize it.

Existing approaches to training language models on historical decision data either ignore this signal or use it in ways that are brittle. Imitation learning trains a model to reproduce the winning arm of past experiments, discarding the losing variants entirely. This throws away evidence about what does not work and how winners differ from losers. Reward optimization methods use both winners and losers to estimate a scalar reward, but policies trained to maximize an estimated reward can exploit regions where the estimate is inaccurate, leading to failures in new contexts. This is an especially costly risk when generating marketing content that is deployed at scale in customer-facing settings.

We propose a different path. Rather than imitating winners or optimizing a learned reward, we fine-tune a language model to learn a transformation: given a lower-performing piece of content from an experiment, produce the higher-performing piece. We call the resulting model an *improver*. Concretely, for each historical A/B test with context x, winning variant y_A, and losing variant y_B, we fine-tune a language model G_θ to maximize $\log G_\theta(y_A \mid x, y_B)$. This objective teaches the model a conditional mapping that takes a context and a reasonable decision as input and outputs a better one. It operationalizes what randomized experiments uniquely provide: clean, within-context directions of improvement. Because only relative performance within a campaign is causally identified, the improver is trained exclusively on within-experiment ordinal comparisons. Cardinal comparisons across campaigns are confounded by audience composition, seasonality, and other factors unrelated to content quality. We refer to this broader approach of aligning the generative process with causal information contained in A/B experiments as *causal alignment*.

The improver is designed for deployment as a component rather than an autonomous system. At inference time it is composed with a baseline policy π_0, which returns some reasonable decision

y_0 given context x. In our field experiment, the baseline policy is a human expert. The expert proposes a subject line; the model proposes an improved version. This human-proposing/AI-improving architecture anchors the model's output to a reasonable baseline, making it less likely to produce something drastically worse than the starting point. We further train a preference model on within-experiment comparisons to re-rank candidates at inference time, with the original human-proposed decision included in the candidate set so the system can fall back to the baseline when no generated alternative scores higher.

We validate this framework in the setting of large-scale field experiments in email marketing. Using subject lines from 20,000 past campaigns, we fine-tune a language model on contrastive pairs constructed from randomized A/B tests and evaluate performance in 36 new campaigns totaling 283 million impressions. In each campaign, recipients are randomized across three conditions: a subject line written by an unassisted human expert, the same subject line improved by a prompted general-purpose language model, and the subject line improved by our fine-tuned improver. The improver achieves click-through rates 33% higher on average than the unassisted expert, improving outcomes in 35 of 36 campaigns and shifting the entire distribution of campaign-level effects upward. The general-purpose model, despite having roughly 30 times as many parameters, does not improve outcomes on average. This contrast suggests that the gains come from alignment to experimental data rather than model scale alone, while the absence of systematic harm supports the practical robustness of the human-proposing/AI-improving architecture.

Related Work

Our paper speaks to the broad literature on using data to guide marketing decisions. First, we extend work that estimates causal effects to guide decision-making. There is a large set of papers in this area including, for example, Yoganarasimhan et al. (2020), which runs an experiment to optimize the length of software free trials, and Dubé and Misra (2023), which runs an experiment to find the profit-maximizing price. These are both low-dimensional treatments while we consider a similar problem for a high-dimensional, unstructured treatments. Our paper also relates to work on feature-based approaches for marketing interventions. Ellickson et al. (2023) estimate the effects of a high- dimensional treatment in the setting of email subject lines, projecting estimated treatment effects onto manually defined text features such as indicator variables for the presence of key words. Their approach yields an interpretable decomposition of treatment effects as an intermediate step to optimizing policies. Sisodia et al. (2025) extract features from product images to represent survey respondents' preferences over these image "treatments". In comparison, we implicitly represent the treatment within a language model and directly generate improved decisions. Our framework allows for explicit structure to be imposed but does not require it.

Other work uses data at the A/B test level to aid inference of treatment effects, in the same spirit as a meta-study. Ye et al. (2024) uses an LLM as a prior to increase the statistical efficiency of estimating the effects of a fixed set of text treatments. Gordon et al. (2023) pools a set of A/B tests and observational studies to enable better estimation of causal effects of observational studies.

We complement these works by addressing how the treatments themselves could be improved by leveraging data across multiple A/B tests.

Finally, we add to an empirical line of work on the alignment of language models towards high-level objectives. Ouyang et al. (2022) formalizes the alignment problem as a decision-theoretic *policy optimization* problem. For a given objective function, the goal is to find a decision rule, represented by a language model, that yields high-quality decisions with respect to an objective function. In Ouyang et al. (2022), the objective was the predicted human preference rating for a chatbot's response, but this formalization naturally extends to more general objectives and decisions. Reisenbichler et al. (2022) outline the first direct adaptation of language models towards a business objective, fine-tuning GPT-2 (Radford et al. (2019)) to generate webpage content that will rank highly in search engine queries. Our work differs from theirs in two key ways. First, in the search engine optimization setting, examples of high-quality decisions exist. That is, the contents of top-ranking pages are observed, and so it suffices to train a language model to imitate them. In contrast, the reason for running A/B tests for marketing communications is precisely that it is not clear what the most effective message will be. In other words, we examine the case where there is no observable expert policy. Second, as a consequence of the earlier point, we develop a contrastive approach to learning rather than one based on imitation (imitation learning requires observable expert policies). We train our model on both better and worse examples of decisions (contrastive learning) instead of just imitating the better decision. We note here that these approaches are complementary. A model trained to imitate could provide the initial input for a model trained to improve.

Our contributions are fourfold. Our first contribution is conceptual. We introduce the *improver* as a new paradigm for aligning language models to causal experimental data, positioned as an alternative to both imitation learning and reward optimization. The key observation is that A/B tests do not merely rank alternatives; they reveal *directions of improvement* in the space of decisions. We operationalize the directional signal by training a language model to map lower-performing content to higher-performing content from the same experiment. The resulting operator, when composed with any reasonable baseline policy, yields a human-proposing/AI-improving architecture that is both practically robust and conceptually distinct from standard alignment approaches.

Our second contribution is methodological. We show how the ordinal information embedded in randomized experiments, data that firms already collect at scale, can be structured as contrastive training pairs that teach a language model generalizable improvement rules, providing a methodological bridge between experimental causal inference and language model fine-tuning. This differs from preference-learning approaches such as DPO (Rafailov et al., 2023), which use pairwise comparisons to optimize an implicit objective directly. We instead treat within-experiment comparisons as supervision for a *transformation*: given a worse decision, produce a better one. We further show how a preference model trained on within-experiment ordinal comparisons can be used to re-rank candidates at inference time, providing additional refinement without requiring cardinal performance estimates.

Our third contribution is empirical. We provide the first field-experimental evidence that aligning language models to A/B test results improves outcomes in new, unseen campaigns. The field experiment rules out the simulation artifacts and unobserved confounds that often limit offline evaluations of content-optimization systems, and the scale of the experiment also allows us to characterize the distribution of effects across campaigns rather than only average performance.

Our fourth contribution is practical and speaks directly to the deployment of AI in decision-making contexts. A general-purpose language model with roughly 30 times as many parameters does not improve outcomes, while our aligned improver achieves a 33% average lift. This shows that performance gains come from learning the signal in causal experimental data rather than model scale alone and suggests that our approach remains valuable even as general-purpose models become more capable. At the same time, our deployment architecture that composes AI improvement with a human baseline bounds downside risk while allowing gains from learned improvement. This offers a practical template for firms seeking to deploy AI in decision contexts where safety is a binding constraint.

The remainder of this paper is organized as follows: we first present a formal framework to precisely define estimands and motivate our approach. We then describe the empirical setting and procedures for estimating our model. Subsequently, we present the results from our field experiment and discuss the sources of performance gains. Finally, we discuss implications and conclude. Implementation details not critical to understanding the main idea are provided in the appendices.

2 The Framework

To clarify and motivate our approach, we introduce some notation. Consider decision problems in marketing where a context x and decision y lead to an outcome or "reward" $r(x, y)$. For example, x may be a product's features, y its price, and $r(x, y)$ the profit. Or x can be ad topic, y ad content, and $r(x, y)$ the clickthrough rate. Letting ϕ denote parameters of the reward function, a typical decision rule, or *policy*, π is to maximize the predicted reward:

$$\pi(x) = \arg\max_y \ r_\phi(x, y).$$

ϕ can be estimated from historical data of the form $D = \{(x_i, y_i, r_i)\}_{i=1}^N$. For data consisting of A/B tests, each observation is of the form (x, y_A, y_B, r_A, r_B), where we adopt the convention that $r_A > r_B$.[1]

When r is differentiable in y, e.g. if y is a real-valued quantity like a price, $r_\phi(x, y)$ can be maximized by gradient ascent. When y is unstructured, however, like the visual appearance of a product or the text in an advertisement, the optimization problem is intractable even if a high quality predictive model of r is available. Gradients no longer exist, and the space of possible decisions is too large to exhaustively evaluate. In what follows, we present our approach to the problem and comparisons to the main alternatives.

[1] This can be generalized to more than two treatment groups.

2.1 The Improver

We propose to train a language model to perform a transformation: given the context and a lower-performing variant from an experiment, generate the higher-performing variant. Consider a language model G_θ that specifies a distribution over text y given input. The fine-tuning objective[2] is:

$$\max_{\theta} \; E_{(x,y_A,y_B)\in D} \left[\log G_\theta(y_A \mid x, y_B)\right]. \tag{1}$$

This objective teaches the model a conditional mapping that takes a context and a reasonable decision as input and outputs a better one. It treats within-experiment pairwise comparisons as direct supervision for conditional generation.

The fine-tuned model is composed with a baseline policy π_0 to define an improved policy:

$$\pi(x) := G_\theta(\cdot \mid x, \pi_0(x)).$$

The baseline π_0 could be a prompted general-purpose language model, a fine-tuned model, or a human expert—any policy known to have reasonable worst-case performance. In our field experiment, the baseline is a human expert who proposes a subject line, and the improver proposes an alternative.

This compositional design addresses a central concern in deploying language models for customer-facing decisions: robust performance in new contexts (i.e. out-of-distribution inputs). Because the improver is composed with a baseline of known quality, its generation is anchored to a reasonable starting point. The model is unlikely to produce something drastically different from its conditioning input—an empirical regularity of fine-tuned language models that our field results confirm. In our field experiment, even when the model lacks a useful training signal, outcomes are not systematically harmed, which provides direct evidence that the anchoring mechanism works in practice. This is analogous to trust-region methods in reinforcement learning (Schulman et al., 2015), which constrain policy updates to remain close to a reasonable baseline policy to ensure monotonic improvement: worst-case performance is anchored to the baseline, while expected performance is strictly higher whenever the model has learned useful improvement regularities from the data. Our method differs in that we allow the baseline policy to be specified flexibly at the time of deployment rather than during the model estimation stage.

2.2 Preference Model for Re-Ranking

To further refine the candidates generated by the improver, we train a preference model to score and rank them. Given an initial decision y_0 from the baseline policy and alternatives $\{y_1, \ldots, y_N\}$ drawn from $G_\theta(\cdot \mid x, y_0)$, we want a function $r(\cdot)$ that ranks the alternatives by quality. It is important that this function measures the effect of the text of y_i and not confounding factors like

[2]Perhaps a more precise notation would be to write θ as an adjustment to pre-trained parameters ω. We simplify this to directly using θ as the parameters of the fine-tuning objective.

message topic or seasonality. Political ads about certain topics (e.g., social security) tend to receive much more engagement than others. If r predicted the click-through rate of an ad based on its text alone, it would rank content about social security over content about the original topic in y_0, which is not the intended use. Similarly, seasonality can inflate engagement for campaigns that happen to run during the holiday season, and the text of those campaigns would be incorrectly attributed as leading to higher performance.

To isolate the effect of making *changes* to a decision, we make r a function of both the original decision y_0 and the candidate y_i, and train it on ordinal comparisons from historical A/B tests. Specifically, we use experiments with at least three treatment groups, denoted y_L, y_M, y_H in order of performance. We train the preference model to rank potential improvements to the middle variant so that $r(y_M, y_H) > r(y_M, y_L)$. Formally, we model these comparisons using the Bradley-Terry framework (Bradley and Terry, 1952):

$$\Pr(y_H \succ y_L \mid y_M) = \frac{\exp(r(y_M, y_H))}{\exp(r(y_M, y_H)) + \exp(r(y_M, y_L))}.$$

The latent reward $r(y_0, y_i)$ is obtained by a linear transformation of an embedding of the concatenated pair:

$$r(y_0, y_i) = \alpha + \beta' e(y_0, y_i),$$

where the embedding vector e is taken from the final hidden layer of a language model. Parameters are estimated via maximum likelihood—this is equivalent to logistic regression on reward differences.[3]

An alternative to ordinal comparisons is to train r to predict the within-campaign z-score at the treatment-arm level, which performs similarly in settings where sample sizes are large and campaigns have many arms. Ordinal comparisons are more robust to noise and better accommodate experiments with a small number of arms.

The final improved policy takes as input a proposed decision y_0, draws N alternatives from $G_\theta(\cdot \mid x, y_0)$, and selects the highest-ranked candidate:

$$\pi(x) = \underset{y' \in \{y_0, y_1, \ldots, y_N\}}{\arg\max} \; r(y_0, y') \quad \text{s.t.} \quad y_i \sim G_\theta(\cdot \mid x, y_0), \quad y_0 = \pi_0(x).$$

Because the original decision y_0 is included in the candidate set, the re-ranking step admits a fallback to the baseline: whenever no generated alternative scores higher than the original, the system selects y_0. This ensures the improved policy cannot be made worse than the baseline by the re-ranking step, provided the preference model is reasonably calibrated.[4]

Reweighting draws from a proposal distribution to better approximate a target distribution is an established technique (e.g., importance sampling, actor-critic methods). Our methodological

[3] This reward function is related to the implicit objective in DPO (Rafailov et al., 2023), which also uses a Bradley-Terry likelihood over pairwise comparisons. The distinction is in how the learned function is used: DPO recovers a policy by treating the reward as an implicit function of the policy parameters, whereas we estimate the reward model explicitly and use it solely to re-rank candidates from a separately trained generator.

[4] If the preference model is badly miscalibrated—ranking a poor generated candidate above y_0—this guarantee does not hold. Our field results, where outcomes improve in 35 of 36 campaigns, provide empirical reassurance that this is not a concern in practice.

contribution is to leverage A/B test data to train both the proposal distribution and the reweighting rule to learn *improvements* to a baseline policy.

2.3 Discussion: Alternative Approaches

The improver differs from two standard approaches to training language models on A/B test data. We briefly describe each and highlight the differences.

Imitation Learning.

Given A/B test data, a natural approach is to train G_θ to reproduce the winning variant:

$$\max_\theta \; E_{(x,y_A)\in D} \left[\log G_\theta(y_A \mid x)\right]. \tag{2}$$

This is simple, but it uses only part of the experimental information. Losing variants reveal what does not work and how winners differ from losers; imitation learning discards this information. Moreover, the performance of the resulting policy is upper-bounded by the quality of the implicit policy that generated the training data—imitation cannot surpass the demonstrator.

Reward Optimization.

Alternatively, the parameters of G_θ could be updated to place higher probability on decisions with high predicted reward:

$$\max_\theta \; E_{x\in D,\, y\sim G_\theta(\cdot|x)} \left[r_\phi(x,y)\right]. \tag{3}$$

This can be solved using policy gradient methods (Schulman et al., 2017). Preference-learning methods such as Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) are instances of this broader approach. Reward optimization uses the full signal in A/B test data. That is, both winners and losers contribute to the reward estimate. However, this approach may introduce a fragility that imitation learning avoids. For example, in RLHF, the policy is trained to maximize an estimated reward, and estimation error in r_ϕ can lead the policy to exploit regions where the reward model is inaccurate, producing distributional failures in new contexts (Amodei et al., 2016). Critically, the learned policy generates from scratch rather than conditioning on a baseline, so there is no architectural anchor when the reward model is wrong. This is an especially serious concern when deploying G_θ as the policy for customer-facing marketing communications. Since our writing of the first version of this paper, there have been other new ideas in this area. We conjecture that adapting DPO based approaches could also be used in conjunction with our improver idea. We leave that to future research.

What the Improver Gains.

The improver uses the full experimental signal (both winners and losers) while avoiding the main vulnerability of reward optimization. By conditioning on a baseline decision rather than generating

from scratch, the improver is architecturally tethered to a reasonable starting point. And by treating within-experiment comparisons as direct supervision for conditional generation rather than routing them through a scalar reward estimate, the training objective avoids the reward-exploitation failure mode. The preference model is used only at inference time for re-ranking, not as the training signal for the policy, which limits the scope of reward misspecification to candidate selection rather than the generative process itself.

The safety aspect of our framework needs to be underscored. As we discuss in the subsequent section, we find that even when the model lacks a useful training signal, the compositional design prevents systematic harm. We now describe the details of the empirical setting and estimation that obtain the results just mentioned.

3 Empirical Setting

To demonstrate the effectiveness of our approach, we conduct a field experiment on real marketing campaigns in a setting with practical relevance and measurable outcomes. Email marketing is one of the most common and effective ways for firms to communicate with their customers, e.g. to announce a new product or a promotion. In these emails, a compelling subject line can grab the reader's attention, increase open rates, and ultimately drive more conversions. Effect sizes are economically significant; in our sample, the best subject line in a campaign attained response metrics that were on average 73% and up to 445% higher than the worst one. However, crafting effective subject lines can be a challenging task, especially for marketers who have to constantly come up with new ideas and variations.

We focus on the task of designing effective marketing email subject lines. Given a topic, the goal is to generate subject line text with a high click rate while remaining relevant and appropriate. Clicks are defined as events where the recipient clicks a link inside the email. We use clicks rather than email opens as the target outcome because they are more reliably measured and a better surrogate for downstream outcomes.

Our training data comes from 10 years of marketing campaigns from a private marketing platform. For the exploratory phase of each campaign, various subject lines were tested, and click rates were recorded. The campaigns come from 337 well-known brands spanning various industries including retail, e-commerce, fashion, financial services, and insurance. The dataset consists of 286k individual subject lines from 20k campaigns, with each campaign sent to a median of 798k recipients. Each campaign typically has 16 variants of subject lines auto-generated from a grammar template, where recipients are randomly assigned and observed differences in subject line performance are interpreted as causal. Subject lines are 60-100 characters in length.

4 Methods

We now describe how the framework is implemented in our empirical setting, including model estimation, deployment with human oversight, and the design of the field experiment used to evaluate it.

Table 1 Fine-tuning data examples

Input (Worse subject line)	Output (Better subject line)
Ready to redecorate? Save up to 70% on home must-haves	You've lucked out: up to 70% off home must-haves
Shop sunny-day styles for less. Shorts, capris & more from 17.99	You've been selected to shop sunny-day styles for less — shorts, capris & more from 17.99
Get ready for Easter family photos with up to 60% off fashion & shoes	You're getting up to 60% off clothing & shoes for Easter

4.1 Model Estimation

Language model fine-tuning.

We fine-tune a language model on contrastive pairs constructed from past A/B tests. For each campaign, if subject line y_A outperformed subject line y_B, we create a training example where the model receives as input "Edit the given email subject line to increase the click-through rate: $\{y_B\}$" and is trained to generate y_A. The objective is to update the parameters of the language model to increase the log probability of generating the higher-performing subject line y_A conditional on a lower-performing subject line from the same campaign. Examples of contrastive pairs are shown in Table 1.

For the pretrained language model, we use T5-base (Raffel et al., 2020), a relatively small model with 220 million parameters. Fine-tuning is done with a batch size of 16, learning rate of 0.0003, AdamW optimizer, and run for 3 epochs. Following standard practices, we regularize via early stopping on the validation loss. Overall training costs were modest. The T5-base model was fine-tuned on a V100 GPU ($2.50 per hour on Google Cloud as of early 2023) running for 20 hours.

At the time of the field experiment, T5 provided a practical foundation for fine-tuning in this application. In subsequent offline analyses, we examine more recent base models.

Preference model estimation.

We use a preference model to rank candidate improvements relative to the initial subject line. Given an initial subject line y_0 and a candidate alternative y_i, the model assigns a score, or *reward*, $r(y_0, y_i)$ that captures whether y_i is likely to improve on y_0. To train this model, we use historical experiments with at least three treatment arms, denoted y_L, y_M, y_H in ascending order of realized performance. From each such experiment, we construct training triplets in which the model is asked to prefer the higher-performing alternative y_H to the lower-performing alternative y_L conditional on the same reference point y_M. This yields training comparisons of the form $r(y_M, y_H) > r(y_M, y_L)$, which isolate the quality of a proposed change relative to a common baseline.

We parameterize the reward $r(y_0, y_i)$ by applying a linear transformation to an embedding of the pair (y_0, y_i). The embedding is from the RoBERTa model (Liu et al., 2019). Estimation proceeds by maximizing the Bradley-Terry likelihood over the training triplets, jointly updating the parameters of the linear transformation and of the embedding model. We train with a learning rate of 5×10^{-6}

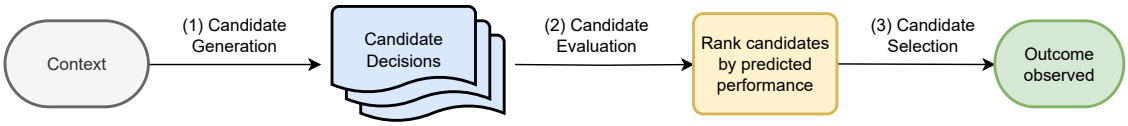

Fig. 1 Deployment of improver. Given some context, a human expert proposes an initial email subject line. Then (1) a generative model proposes improved alternatives, (2) a reward model re-ranks the candidates, and (3) a human makes the final selection.

and use early stopping based on holdout accuracy. On a held-out set of experiments, the preference model correctly ranks the higher-performing alternative in 69.7% of comparisons.

4.2 Deployment

The improver is designed for deployment as a component within a decision-making process rather than as an autonomous system. In our field experiment, a human expert proposes an initial subject line, the model generates improved alternatives, a preference model re-ranks those alternatives, and the human makes the final selection. Figure 1 illustrates this process.

This design serves two purposes. First, conditioning the generative model on a human-authored baseline anchors the model's output to a reasonable starting point. Second, retaining a final human review step serves as a guardrail against undesirable outputs, such as inaccurate or misleading subject lines (Bender et al., 2021). The resulting workflow allows the system to benefit from experimental data while preserving human oversight in a customer-facing setting.

The framework is flexible and can accommodate additional constraints or controls when needed. For example, one can append attributes such as desired emotional valence to the model input during fine-tuning, allowing the system to generate improvements with specified stylistic properties. More generally, features that are useful for predicting performance can be used to steer generation. Separately, generated outputs can be screened for factual accuracy, relevance, and appropriateness before being shown to the human for final review. These extensions are not central to our contribution, and we provide details in Appendices A-C.

4.3 Experiment Design

We evaluate our framework in new email marketing campaigns conducted through a private marketing platform that provides content optimization services to clients across industries. For each campaign, the control subject line is the human-authored subject line that the brand would otherwise send. We compare this control to two treatment arms that measure the value of AI assistance with and without alignment to historical A/B test data.

In the first treatment arm, the human-in-the-loop workflow of Figure 1 is implemented using a general-purpose language model that is not aligned to historical A/B test data. Specifically, the control subject line is entered as part of a prompt to ChatGPT, which returns five alternative subject lines, from which the Brand Content Strategist (BCS) selects one.[5] This arm provides a benchmark for the value of the improver workflow when the generative process is not informed by past experiments. In the second treatment arm, candidate subject lines are generated by our

[5]Details of the prompt are provided in Appendix D.

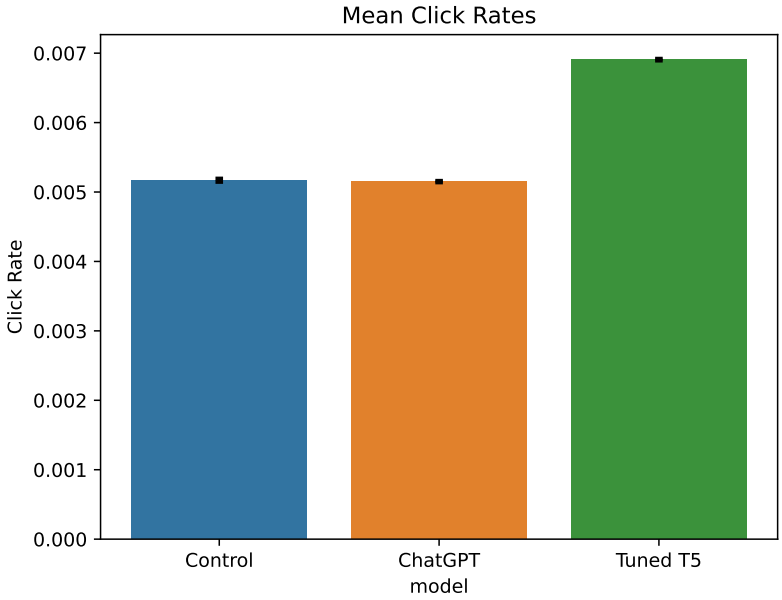

Fig. 2 Mean click rates over deployed campaigns. Assistance from ChatGPT does not outperform an unassisted human, while assistance from the tuned model does. Numerical values for means and standard errors are in Table 2.

causally aligned improver. Outputs from the fine-tuned generative model are re-ranked by the preference model, and the BCS selects a subject line from this ranked list.

Within each campaign, recipients are randomized across the three arms, and the primary outcome is whether the recipient clicks on a link inside the email. We summarize performance using click rates in each treatment arm. Across the 36 campaigns in our study, this design allows us to compare unassisted human judgment, assistance from a general-purpose model, and assistance from a model aligned to the causal signal contained in prior A/B tests.

The experiment is designed to evaluate AI assistance rather than full automation. In both treatment arms, the human remains involved in proposing the initial subject line and selecting the final output. This reflects how such systems are likely to be deployed in practice and reduces downside risk in a customer-facing setting. We therefore interpret the experiment as measuring the value of combining human judgment with AI improvement, and leave questions about full substitution of AI for human experts to future work.

5 Results

5.1 Click Rates

In 36 new email marketing campaigns totaling 283 million impressions, assistance from our causally aligned model increases click rates by 33% on average relative to unassisted human experts, while assistance from ChatGPT does not improve outcomes. Figure 2 reports average click rates across campaigns, and Table 2 provides the corresponding numerical values.

Table 2 Mean click rates over deployed campaigns in basis points

Model	Click Rate (bp) mean	s.e.	Count Campaigns	Impressions
Control	51.69	0.127	36	31.5m
ChatGPT	51.49	0.063	36	126m
Tuned T5	**69.06**	0.073	36	126m

The comparison to the general-purpose model is informative because the two treatment arms use the same workflow but differ in whether the generative process is aligned to prior A/B tests. In the ChatGPT-assisted arm, average performance is essentially unchanged relative to the control. As shown in Figure 3 and the left panel of Figure 4, outcomes under ChatGPT are mixed across campaigns. The model does not improve performance on average, but it does not systematically worsen it either. This provides evidence that the human-proposing/AI-improving workflow remains practically robust even when the model is not informed by the causal signal in past experiments.

In contrast, the causally aligned model improves outcomes broadly rather than only in expectation. Figure 3 shows that its campaign-level outcome distribution is shifted to the right of both the control and ChatGPT, and the right panel of Figure 4 shows that it outperforms the control in 35 of 36 campaigns. The gains are not driven by a small number of outliers; the improver raises performance across nearly the full distribution of campaigns.

Taken together, these results show that the gains from our framework do not arise from model scale alone. The general-purpose model, despite having roughly 30 times as many parameters, does not improve outcomes on average, whereas the smaller model trained on within-experiment comparisons delivers substantial gains across the full distribution of outcomes. The evidence thus points to causal alignment to past experiments, rather than general language ability alone, as the source of performance improvement.

5.2 Sources of Performance Gains

The field experiment compares two conditions: assistance from a general-purpose model that is not aligned with historical A/B test data, and assistance from our causally aligned system. This comparison establishes the value of alignment as a whole, but it does not isolate which components are most responsible for the observed gains. We now turn to decomposing the total effect into its constituent component effects.[6]

Since (re)running a comparable experiment on a new set of email campaigns is costly, we conduct an offline evaluation[7] on held-out data, using the scores from the preference model to proxy for real outcomes. For each initial subject line y_0, we generate candidate alternatives y_i under different policies and evaluate them using the preference model's predicted probability of improvement, $Pr(y_i > y_0)$. This is the estimated probability that a candidate subject line outperforms the original. One could think of this improvement probability as a "win rate". We note here that this probability does not correspond directly to the outcome of interest (e.g. click-through rates). Rather, it reflects

[6]We thank the reviewers for suggesting this analysis.
[7]More details about this evaluation are available from the authors upon request.

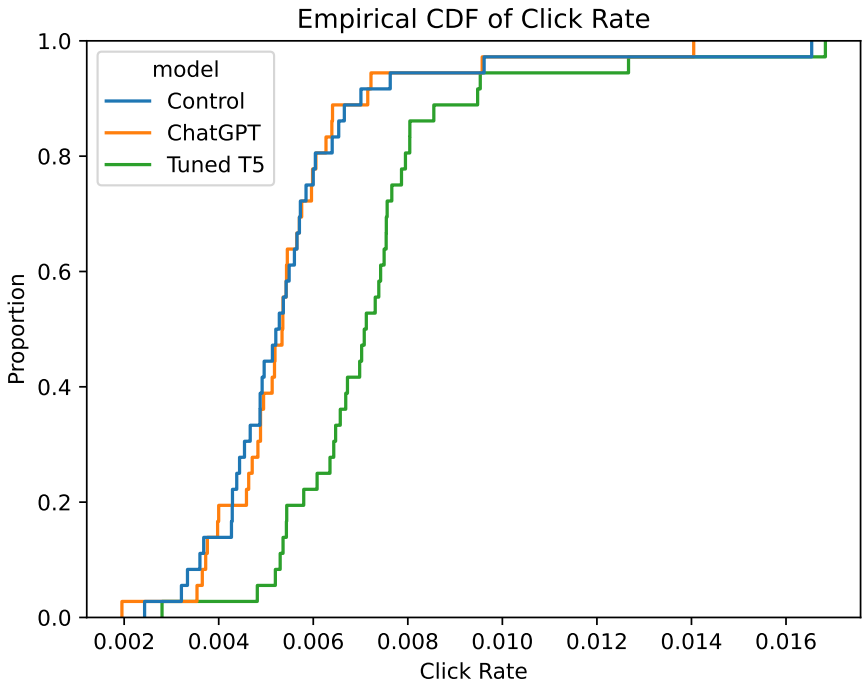

Fig. 3 Marginal distributions of click rates over campaigns. The outcome distribution from our tuned model first order stochastically dominates the distribution from the control and ChatGPT. Its CDF is shifted to the right, which means that every percentile is larger. ChatGPT assistance does not help performance but does not hurt it either, which shows the robustness of our human-initialized/AI-improved design.

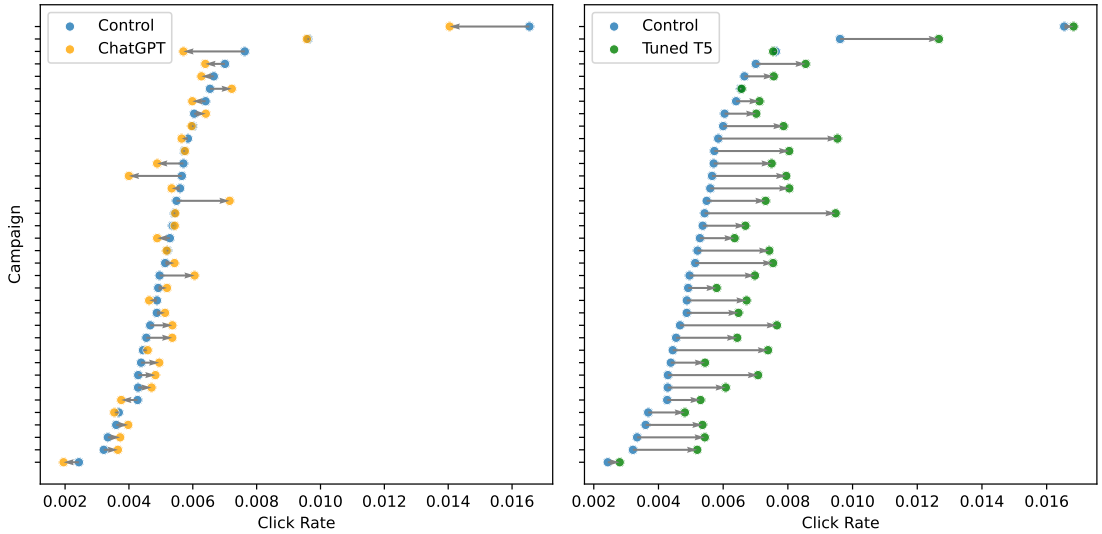

Fig. 4 Joint distribution of click rates, by campaign. Relative to the control, our tuned model improves outcomes (shifts points to the right) for 35 out of 36 campaigns. ChatGPT shows mixed results, improving on 19 out of 36 campaigns, but it has a positive effect in campaigns where humans perform worse and a negative effect in campaigns where humans perform better.

the probabilistic assessment of the scoring model (one trained on historical A/B tests) that one message is "better" than the other. If the preference model is indifferent about the two subject lines (i.e. it predicts they will perform similarly), the improvement probability will be close to 50%, so the relevant range of improvement probabilities is 50% to 100%.

We want to compare the same base language model with and without fine-tuning. The original T5 model is now dated and not suitable for this task without fine-tuning, so we use a more up-to-date model: Llama3.3-70B (Grattafiori et al., 2024). On held out campaigns, prompting the base model gives an improvement probability of 52.8%, which is close to random and consistent with the field-experimental result that assistance from a general-purpose model does not systematically improve the message. By contrast, draws from the fine-tuned model improve on the input with probability 71.4%, indicating that fine-tuning shifts the generative distribution toward better subject lines. Selecting the best of five draws from the fine-tuned model further raises the improvement probability to 99.8%. We note here that these results are conditional on the preference model, and we are implicitly assuming that the human expert will always pick the best option. Humans may choose to pick a sub-optimal option in which case one should view this probability as an upper bound. The overall decomposition is represented succinctly below.

$$\underbrace{P_{\text{IMP}}(\text{FT+Sel}) - P_{\text{IMP}}(\text{Base})}_{\text{Total effect}=0.998-0.528=0.470} = \underbrace{P_{\text{IMP}}(\text{FT}) - P_{\text{IMP}}(\text{Base})}_{\text{Fine-tuning effect}=0.714-0.528=0.186} + \underbrace{P_{\text{IMP}}(\text{FT+Sel}) - P_{\text{IMP}}(\text{FT})}_{\text{Selection effect}=0.998-0.714=0.284}$$

Here, $P_{\text{IMP}}(\pi)$ denotes the improvement probability under policy π, i.e. the probability, according to the preference model, that a candidate subject line generated under policy π outperforms the original subject line y_0. We use Base to denote the untuned base language model, FT the fine-tuned model, and FT+Sel the policy that generates multiple draws from the fine-tuned model and selects the highest-scoring candidate.

These results suggest that fine-tuning is an important component of the gains from the fully aligned system (contributing around 40% of the total effect), while re-ranking further improves performance by selecting among higher-quality candidates (about 60%). The reader should note that this result is based on the data, language models being used, as well as the methodological choices in the fine-tuning procedure. Even so, we feel comfortable claiming that the alignment of the language model contributes in a significant manner to the performance of the overall procedure.[8]

More broadly, these results speak to the question of how best to guide the generative process toward downstream objectives. Fine-tuning is one approach, but it sits alongside other levers such as prompt and context engineering and inference-time computation (e.g. recently developed "reasoning" models). Our evidence does not adjudicate among these approaches in general. It shows that post-training on causal A/B test data is an effective mechanism in this setting, and that it contributes meaningfully within the fully aligned system we evaluate in the field.

[8]Performance also varies with the amount of training data. In additional offline analyses, gains from fine-tuning increase with the number of training experiments, with strong returns from increasing the number of experiments in the training set up to 370 and diminishing returns afterwards. Full results are in Appendix E.

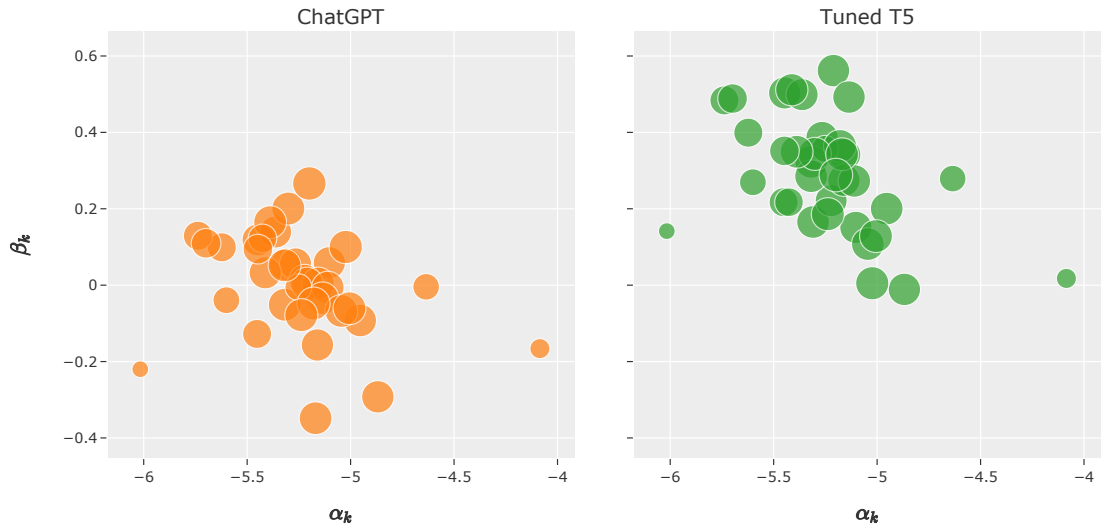

Fig. 5 Treatment effect from model assistance (β_k) vs. baseline performance (α_k), measured in log odds ratios. Email campaigns from our field experiment have a negative correlation between the treatment effect from model assistance (β_k) and the performance of an unassisted human (α_k). This is consistent with AI assistance being more helpful in instances that are more difficult for humans. The size of each marker is proportional to the number of impressions in the respective campaign.

5.3 Complementarity

We next examine how the gains from AI assistance vary across campaigns. Figure 5 plots the treatment effect of model assistance against baseline campaign performance under the unassisted human control. For both the general-purpose model and the causally aligned model, treatment effects are larger in campaigns where the control performs worse. This pattern is consistent with AI assistance being most valuable in settings that are more difficult for human experts.

We interpret this result as descriptive rather than causal. Several mechanisms could generate the same pattern, including differences in task difficulty, differences in human effort or ability, or simple ceiling effects. Still, the finding is informative for practice because it suggests that the value of AI assistance is not uniform across campaigns. Understanding when human judgment and AI improvement are most complementary is an important direction for future work.

5.4 Content Quality

Another important question is whether the performance gains from causal alignment come at the cost of content quality. To assess this, we compare outputs from ChatGPT and our aligned generation process using expert evaluations of whether a generated subject line is acceptable for deployment without edits.

For each of 192 inputs, we generate five candidate subject lines from ChatGPT and five from our aligned model. The resulting outputs are pooled and shuffled, and eight expert evaluators rate each one as either acceptable or unacceptable for use in an email campaign according to the criteria described in Appendix B. A rating of 1 indicates that the output could be used in a campaign

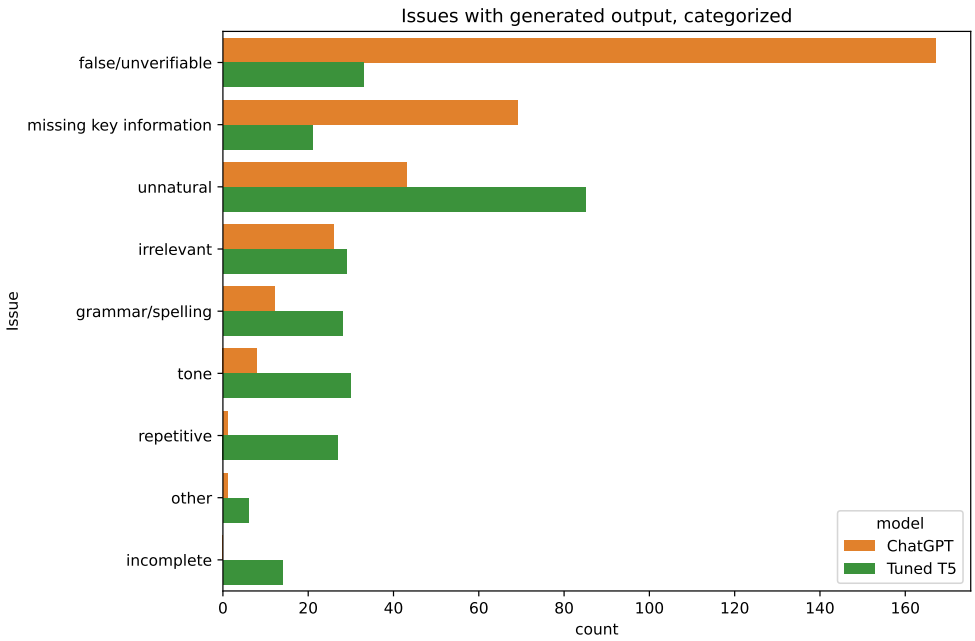

Fig. 6 Issues with AI-generated output

without edits. Our aligned generation process receives a higher acceptance rate than ChatGPT: 71.0% versus 65.6%.

The two models fail in different ways. As shown in Figure 6, ChatGPT more often produces outputs that omit key information or introduce claims that are false or unverifiable. By contrast, the aligned model's failures are more often stylistic than semantic, with unnatural phrasing as the most common issue. Example outputs are shown in Tables 3 and 4. This pattern is consistent with ChatGPT being trained to produce broadly fluent and natural responses, whereas our aligned model is trained to generate content shaped by the causal signal contained in past A/B tests.

Overall, these results suggest that the gains from causal alignment do not come at the expense of lower-quality content. Relative to ChatGPT, the aligned workflow performs better on both observed margins: click performance in the field and expert-rated acceptability of generated subject lines.

6 Conclusion

We develop and validate a new way to use randomized experiments to improve unstructured decisions. Rather than treating A/B tests only as a tool for selecting the best performer among a small set of alternatives, we show how they can be used to train a language model to learn *directions of improvement*. Within each experiment, randomization identifies which message causally outperforms another in a fixed context. Aggregated across many experiments, these within-context ordinal comparisons provide supervision for a model that transforms a reasonable decision into a better one.

In our application to email marketing, we implement this idea through an improver architecture that combines a human baseline with a model trained on historical A/B tests. A human expert proposes an initial subject line, the model generates improved alternatives, and a preference model

Table 3 Example issues with ChatGPT output

Issue	ChatGPT content	Expert comment
false/unverifiable	Input: Shop now >>> everything is extra 35% off Output: Last Chance to Save Big! Show Now and Get Extra 35% off	"Last Chance" is unverifiable
missing key information	Input: 90 day trial of Digital Advisor Output: Get Ahead of the Game: Sign up for 90 Days of Digital Advisor Guidance	"trial" is missing in the input
missing key information	Input: open a CD | limited-time 13-month CD | see offer Output: Maximize your savings potential with this exclusive CD opportunity.	"limited-time 13-month" is missing in the output

Table 4 Example issues with Tuned T5 output

Issue	Tuned T5 content	Expert comment
unnatural	Input: take an additional 40% off | last day | free shipping | 40% Output: Go on, you've scored: Last day to take an additional 40% OFF + free shipping!	"Go on, you've scored:" sounds unnatural as an intro
irrelevant	Input: BOGO free on all shorts. Let those temps rise. Output: You're getting BOGO free on ALL shorts. Warm yourself up!	"Warm yourself up!" is irrelevant
repetitive	Input: you're pre-approved | offer for an additional line of credit Output: You'll love this offer for an additional line of credit. You're pre-approved for this offer!	"You'll love this offer" and "You're pre-approved for this offer!" sound repetitive

re-ranks those alternatives using within-experiment comparisons. This design is important substantively as well as practically. It allows the system to benefit from experimental data while remaining anchored to a reasonable starting point, which makes deployment in customer-facing settings more robust.

We evaluate the framework in a large-scale field experiment spanning 36 campaigns and 283 million impressions. Subject lines created with assistance from our causally aligned system achieve click-through rates 33% higher than those from unassisted human experts. The gains are not confined to a small number of campaigns: the tuned model improves outcomes in 35 of 36 campaigns and shifts the full distribution of campaign-level performance upward. By contrast, a much larger general-purpose language model does not improve outcomes on average. This comparison is central to the paper's interpretation. The results show that the value comes not from model scale alone, but from aligning generation to the causal signal contained in past experiments.

These findings have broader implications for both AI deployment and experimentation. For AI deployment, they suggest that in decision settings where safety and robustness matter, a human-proposing/AI-improving architecture can be a practical alternative to full automation. For experimentation, they imply that firms should view A/B tests not only as a way to choose among current options, but also as a strategic asset for building future decision-making capability. The directional information revealed by randomization is not available in ordinary observational data, and our results suggest that this causal structure can be turned into systematic performance gains.

Two extensions are especially promising. First, our framework takes experimental data as given, but future work could redesign experimentation itself to generate treatments that are more informative for training improvement models (Dew, 2024). Second, because our results point to gains from combining human judgment with AI improvement, an important next step is to understand how interface design and the allocation of responsibility shape performance in settings with unstructured decisions (Agarwal et al., 2023; Green and Chen, 2019).

Our empirical setting focuses on email subject lines, but the logic extends more broadly to other unstructured treatments, including advertising copy, website content, product descriptions, images, and policy communications. The same approach could also be applied to targeted or personalized interventions when organizations can estimate recipient-level heterogeneity. More generally, whenever organizations repeatedly experiment over high-dimensional interventions, they possess data that can be used not just to evaluate decisions, but to train systems that make better ones. Integrating generative AI with experimentation thus turns causal evidence into a mechanism for cumulative learning and accelerates the discovery of superior interventions.

References

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz. 2023. Combining human expertise with artificial intelligence: Experimental evidence from radiology. *National Bureau of Economic Research* .

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565* .

Bender, E.M., T. Gebru, A. McMillan-Major, and S. Shmitchell 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bradley, R.A. and M.E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4): 324–345 .

cjadams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, nithum, and W. Cukierski. 2017. Toxic comment classification challenge. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

Dew, R. 2024. Adaptive preference measurement with unstructured data. *Management Science 0*(0) .

Dubé, J.P. and S. Misra. 2023. Personalized pricing and consumer welfare. *Journal of Political Economy 131*(1): 131–189 .

Ellickson, P.B., W. Kar, and J.C. Reeder III. 2023. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science 42*(4): 704–728 .

Gobe, M. 2010. *Emotional branding: The new paradigm for connecting brands to people*. Simon and Schuster.

Gordon, B.R., R. Moakler, and F. Zettelmeyer. 2023. Predictive incrementality by experimentation (pie) for ad measurement. *arXiv preprint arXiv:2304.06828* .

Grattafiori, A., A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* .

Green, B. and Y. Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction 3*(CSCW): 1–24 .

Keskar, N.S., B. McCann, L.R. Varshney, C. Xiong, and R. Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* .

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P.F. Christiano, J. Leike, and R. Lowe 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 27730–27744.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog 1*(8): 9 .

Rafailov, R., A. Sharma, E. Mitchell, S. Ermon, C.D. Manning, and C. Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290* .

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research 21*(1): 5485–5551 .

Reisenbichler, M., T. Reutterer, D.A. Schweidel, and D. Dan. 2022. Frontiers: Supporting content marketing with natural language generation. *Marketing Science 41*(3): 441–452 .

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz 2015. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

Sisodia, A., A. Burnap, and V. Kumar. 2025. Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *Journal of Marketing Research 62*(3): 405–428 .

Ye, Z., H. Yoganarasimhan, and Y. Zheng. 2024. Lola: Llm-assisted online learning algorithm for content experiments. *arXiv preprint arXiv:2406.02611* .

Yoganarasimhan, H., E. Barzegary, and A. Pani. 2020. Design and evaluation of personalized free trials. *arXiv preprint arXiv:2006.13420* .

Appendix A Controlling Emotional Valence

The language model that generates improvements is a black box. It converts raw text into raw text. It is possible to impose structure on this process, e.g. by enabling the human to ask for text with a specific emotional valence. We provide one example of how to do this.

In our A/B test data, the emotional valence of the text, which our data provider had defined using an ontology from (Gobe, 2010), was a key explanatory variable for predicting engagement. As shown in Table A1, the emotional ontology has five categories: PRIDE, TRUST, JOY, ANTICIPATION, and FEAR, and each top-level category is further divided into 3 sub categories. After tagging each subject line in our historical dataset with the applicable categories, the language model fine-tuning procedure can be modified to append the emotional tags of the *output* to the *input*. This teaches the model to take as input an initial subject line and the desired emotional valence of the output, and return an improved subject line with that emotional valence. This fine-tuning technique is called conditional training and was previously used to control style and content in linguistics (Keskar et al., 2019). This step is not required but shows how features that are useful for predicting performance can be used to steer generation.

Appendix B Acceptability Guardrail

We define a binary metric, which we call "acceptability", to take value 1 if a subject line can be used without edits and 0 otherwise. On model outputs generated from 4600 unique inputs, we collected 100k binary ratings from in-house experts. They were instructed to rate model outputs as 1 only if:

- Sounds like natural, correct English and conveys the same meaning as input information
- Includes all of the information provided by the keyphrases

Table A1 Emotional ontology

Top category	Sub-category	Definition
PRIDE	ACHIEVEMENT	to praise or reward for an implied accomplishment
	EXCLUSIVITY	to imply or state one's unique privilege in receiving the message
	LUCK	to point out good fortune in having the chance to enjoy something special
TRUST	SAFETY	to eliminate any worries or doubts, to make one feel secure
	GRATITUDE	to express acknowledgement, appreciation, or affection in a personal way
	INTIMACY	to address or salute in a formal or informal way that implies some sort of relationship
JOY	EXCITEMENT	to deliver positive news or introduce something enthusiastic
	FASCINATION	to stimulate excitement/interest related specifically to a new experience or possession
	GRATIFICATION	to stimulate excitement/interest related specifically to value or a financial gain
ANTICIPATION	ENCOURAGEMENT	to motivate and/or inspire one to take an action by explicitly prompting them to do something
	CURIOSITY	to nudge, intrigue or tease; to stimulate interest by being vague
	CHALLENGE	to provoke a decision or an action by either daring or asking a question
FEAR	ATTENTION	to alert about the importance of a certain message and/or to provide information
	URGENCY	to warn about the importance of a certain message and/or provide information
	REGRET	to encourage a certain action by stressing one's potential fear of missing out

Example: keyphrases = new arrivals | summer sales event. "Here are all the new arrivals for our summer sales event" is acceptable, "Our summer sales event has arrived" is not.

- Is not more specific than the keyphrase information
  Example: keyphrase = apply for a home loan. "You're invited to apply for a home loan" is acceptable, "This is your last chance to apply for a home loan" is not.
- Does not contain any offensive language
- Does not contain grammatical or spelling errors, i.e. avoids excessive repetition and structures with multiple intros or outros unless the keyphrases specifically warrant it. If it feels unnatural, it's unnatural.
- Is not irrelevant to the keyphrases

Using this data, we train a binary classifier that returns the probability that acceptability=1 given subject line text. For inference, we normalize the relative confidence scores the model places on output tokens of 0 or 1 to obtain a predicted probability that acceptability=1. To reduce the load on the human reviewer in the last stage of Figure 1, we filter generated subject lines that have a probability of being acceptable below 0.4.

Table C2 Toxicity types

Type	Definition
Toxic	very bad, unpleasant, or harmful
Severe toxic	extremely bad and offensive
Obscene	(with respect to the portrayal or description of sexual matters) offensive or disgusting by accepted standards of morality and decency
Threat	a statement of an intention to inflict pain, injury, damage, or other hostile action on someone in retribution for something done or not done
Insult	speak to or treat with disrespect or scornful abuse
Identity hate	hatred, hostility, or violence towards members of a race, ethnicity, nation, religion, gender, gender identity, sexual orientation, or any other designated sector of society

Appendix C Additional Guardrails and Post-Processing

In addition to checking generated subject lines for acceptability, we apply the following steps:

1. Toxicity check

   We trained a toxicity prediction model by fine-tuning the 110m parameter BERT-base-uncased model on the public "Toxic Comment Classification Challenge" dataset from Kaggle (cjadams et al. (2017)). The toxicity model detects the toxicity types defined in Table C2. Generated subject lines classified into any of these categories with probability more than 0.8 were removed.
2. Diversity check

   To reduce redundancy, we compute the self-BLEU score for each subject line, where higher values mean more similarity with the rest of the subject lines. We filter out any subject lines with self-BLEU higher than 0.9.
3. Entity replacement

   Sometimes our language model generated additional information not contained in the input. A specific brand or product name not given in the input could be in the output. To fix this issue, we trained an entity extraction model for detecting brand and product. At inference time, we run this model on the output. If it detects a brand and/or product, we check whether such phrases appear in the input. If there is a mismatch, the phrases in the output are replaced by the phrases in the input. The entity extraction model was trained by fine-tuning a RoBERTa-base model on subject lines in our data which already had brand and product tags. The input is the subject line, and the output is the tags.
4. Format correction

   Some hard-coded rules were applied, like making sure the first letter of the output was capitalized and requiring a space before and after an emoji unless it is in the beginning or end of the output.

Num experiments	Win-rate	Accuracy
1464	0.714	0.697
657	0.676	0.687
376	0.673	0.675
154	0.650	0.644
40	0.641	0.624
12	0.618	0.605

Table E3 Performance of language model generating improvements ("win-rate") and of preference model used for reranking ("accuracy") as a function of the number of the number of experiments in the training set. Evaluations are conducted on a hold out.

Appendix D ChatGPT Prompt

We used ChatGPT, specifically "gpt-3.5-turbo", which at the time of our study was the most capable model available from OpenAI. We formulated the following prompt template and applied it to every input:

Generate {num_results} performant email marketing subject lines using the key information delimited by triple backticks: ```{input}```

If we ask the model for 5 results for "best selling | luggage | up to 70% off | family of brands", OpenAI's API returns:

1. *"Up to 70% off best selling luggage from your favorite family of brands"*
2. *"Don't miss out on our luggage sale: up to 70% off top sellers"*
3. *"Travel in style with our best selling luggage, now up to 70% off"*
4. *"Amazing deals on must-have luggage from our family of best selling brands"*
5. *"Pack your bags for less: save up to 70% on luggage from our top selling brands"*

Appendix E Performance Scaling with Sample Size

We analyze how the performance of the language model used for generating improvements and the preference model used for reranking vary as a function of training sample size. The language model is evaluated on its probability of improvement, or win-rate, compared to the input subject line. This is the fraction of generations that are predicted to outperform the input, with evaluations done according to the best performing preference model. The various preference models are evaluated on their accuracy in ranking held out experiments, under the same triplet structure that was used in estimation.

Evaluations are conducted on a held out set of experiments. To ensure against data leakage, we require that the campaigns in the training set and hold out set come from different brands. In Table E3, we show performance against various sizes of the training data.