Causal Alignment: Augmenting Language Models with A/B Tests

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

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

Abstract

We develop a general framework for optimizing the content of marketing communications by adapting a language model to past A/B test results. We fine-tune a language model to transform lower-performing content into higher-performing variants, teaching it to perform hill-climbing in the space of text. When deployed, the model generates improvements to marketing content proposed by humans. This design ensures that AI assistance is unlikely to harm performance, which mitigates implementation risks and facilitates organizational adoption. We validate our framework through a large-scale field experiment in email marketing. In 36 campaigns covering 283 million total impressions, subject lines created with assistance from our tuned model achieve click-through rates 33% higher than those created by unassisted human experts. These performance gains are causally attributable to improvements in content quality, confirming the effectiveness of our framework. Additionally, a general-purpose language model with 30x the parameters fails to improve outcomes while our smaller fine-tuned model succeeds, demonstrating that domain-specific adaptation is necessary and relatively small language models are sufficient. Our findings provide experimental evidence that language models can extract generalizable insights from A/B tests, enabling systematic optimization of unstructured marketing decisions from copywriting to creative design.

Keywords: NLP, Analytics, Decision Support, Field Experiment, 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 conduct A/B tests to optimize the content of their marketing communications, from promotional emails to social media ads. While these experiments identify the best performer among tested variants, they provide limited guidance for creating new content. We develop and validate a novel approach that extracts generalizable insights from historical A/B test results to systematically improve marketing content creation.

Optimizing unstructured content is a fundamentally difficult problem. For structured decision variables like price, firms can estimate predictive models (e.g. a demand curve) and solve for the optimal action. However, this approach does not work for unstructured decision variables like the text of an advertisement or the visual appearance of a product. A predictive model for the click-through rate of advertising text does not reveal what the optimal text is. Typical practice instead uses the predictive model to rank human-generated ideas in an ad hoc fashion. We show how to leverage generative AI to improve this process.

The core insight of our approach is that A/B tests provide rich causal evidence about what makes marketing content effective. Individual A/B tests establish causal relationships between content changes and outcomes through randomization. Multiple A/B tests together should contain common features that distinguish high-performing content from low-performing content. We extract this information by fine-tuning a language model on results from past experiments, training it to transform lower-performing content into higher-performing variants. Essentially, we teach the language model to perform hill-climbing in the space of marketing text. When deployed, the model generates improvements to human-proposed content. This human-proposing/AI-improving design ensures AI assistance is unlikely to harm performance, facilitating organizational adoption.

We validate our framework through a large-scale field experiment in email marketing. Using a language model fine-tuned on subject lines from 20,000 past campaigns, we test its effectiveness in 36 new campaigns totaling 283 million impressions. Subject lines created with assistance from our tuned model achieve click-through rates 33% higher than those created by unassisted human experts. Moreover, assistance from a general-purpose language model with 30 times the parameters fails to improve outcomes. This demonstrates that domain-specific data linking text to business outcomes is necessary and relatively small models are sufficient. We also find evidence of complementarity between human expertise and AI, with the gain from AI assistance largest in instances where unassisted humans perform worse. While our experiment focuses on email marketing, our approach extends naturally to other domains with unstructured decisions like advertising copywriting and product design.

Our primary contributions are:

1. We develop a practical framework for extracting generalizable insights from previous A/B tests: "A better than B" means "turn B into A".
2. We provide experimental evidence that language models fine-tuned with our approach can systematically improve marketing decisions in new contexts.

We also contribute to broader questions about human-AI collaboration in decision-making tasks. Our framework enables AI assistance that significantly improves human decision quality while maintaining appropriate safeguards, as we train a language model to optimize engagement metrics while preserving secondary objectives like factual accuracy. The framework also allows for explicit structure to be imposed on the generated content. In our experiment, we show how to control and optimize the emotional valence of generated text, blending traditional marketing principles with generative AI capabilities. These results suggest promising directions for developing AI systems that complement rather than replace human expertise.

There is an active literature on predicting the performance of marketing content, e.g. Yang and Zhai (2022) review techniques for predicting the click-through rate of online advertisements. We show how to combine such predictive models with a generative model to obtain prescriptive recommendations. We also add to work on estimating causal effects to guide decision-making. There are many papers on this topic; a few examples are 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; we consider a similar problem but for a high-dimensional treatment. Ellickson et al. (2023) estimate the effects of a high-dimensional treatment also in the setting of email subject lines. They project the estimated treatment effects onto manually defined features of the text, like indicator variables for the presence of key words. Their approach yields an interpretable decomposition of treatment effects as an intermediate step to optimizing policies. In comparison, we implicitly represent the treatment within a language model and directly solve for a better policy. By avoiding the need to explicitly codify features, we operationalize information from the data that is difficult to articulate (Polanyi (1966)). Additionally, our framework allows for explicit structure to be imposed but does not require it. We do not address optimal targeting or personalization of treatments in this work, instead focusing on optimizing the contents of a blanket treatment.

Finally, we add to an empirical line of work on the alignment of language models with different kinds of objectives. The usefulness of adapting a pretrained language model to a task was first shown in Dai and Le (2015) and popularized in Devlin et al. (2018). For linguistics tasks like sentiment classification, they found that training a next-word prediction model on unlabeled text before updating its parameters on task-specific data increased accuracy. The intuition is that teaching a model to "understand" language is an informative prior. Recently this approach has been extended to more complex tasks. Wei et al. (2021) and Ouyang et al. (2022) show the remarkable result that given examples of instructions and responses, sufficiently large models can learn to follow previously unseen instructions. Reisenbichler et al. (2022), which our work builds on, show the first direct adaptation of language models to maximizing a business outcome, fine-tuning GPT-2 (Radford et al. (2019)) to generate performant SEO content. Similarly to their setup, we design our model to optimize text generation for a business outcome. But instead of using signals from an external algorithm (features of existing high-ranking results on Google), we extract informative signals from firms' historical data, further broadening the scope of applications for language models to generic business objectives.

The remainder of this paper is organized as follows: Section 2 introduces our general framework, and Section 3 gives details on the email marketing task and dataset. Section 4 describes our training and inference procedures. Section 5 covers the results from our validation experiment, discusses human-AI complementarity, and analyzes the quality of AI-generated output. Section 6 concludes. Implementation details not critical to understanding the main idea are covered in the appendices.

2 General Framework

Many business decisions can be expressed as optimization problems 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. Letting ϕ denote additional parameters of the reward, like the elasticity of demand, a typical decision rule is to maximize the predicted reward:

$$y^*(x) = \arg\max_y \ r(x, y; \phi).$$

ϕ can be estimated from historical data of the form $D = \{(x_i, y_i, r_i)\}_{i=1}^N$.

When r is differentiable in y, e.g. if y is a real-valued quantity like a price, $r(x, y; \phi)$ 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.

A heuristic solution is to rely on a domain expert who generates what they believe are promising candidate decisions $y_1, y_2, y_3, \ldots$ and uses the predictive model $r(x, y; \phi)$ to rank them. Yet this approach is limited, as it does not provide constructive guidance on what decision the expert should choose. That process is left to a mix of intuition and ad hoc interactions between the expert and the predictive model.

Our procedure for optimizing decisions in unstructured action spaces leverages the capabilities of generative models. While we focus on decisions that take the form of text, the same approach should work for other modalities like images.

Consider the following procedure for maximizing the reward:

1. Generate $y^* \sim G(y|x; \theta)$, where G is the output distribution of a language model with parameters θ.
2. Fine-tune θ so that G generates actions with high predicted reward:

$$\max_\theta \ E_{x \sim D, y \sim G(y|x;\theta)}[r(x, y; \phi)]. \tag{1}$$

Solving this optimization problem, however, creates three new issues. First, estimation error in the reward model will lead $G(y|x; \theta)$ to overfit to the estimated reward and generate unrealistic text (Christiano et al. (2017)). Second, specification error where the reward does not perfectly capture real-world objectives will lead to undesirable text (Amodei et al. (2016)). Lastly, the distribution

$G(y|x;\theta)$ could perform much worse than predicted on new data (Amodei et al. (2016)). For these reasons, fully delegating decisions to an AI is infeasible in many practical settings.

We prevent unrealistic output by regularizing $G(y|x;\theta)$. After initializing θ at the parameters θ_0 of a pre-trained language model known to generate realistic text, we terminate the optimization procedure early to limit the drift of the fine-tuned model from $G(y|x;\theta_0)$.

We address undesirable output from misspecified reward by collecting additional data on which outputs humans consider acceptable. We filter generated output based on a model learned from this data. The most common issues with AI-generated output are related to factual accuracy, and our learned model is successfully able to identify these instances.

Finally, in our main methodological innovation, we ensure robust performance by designing our language model to improve rather than replace humans. Given a new context x, instead of directly generating a decision y from a language model, we first solicit a decision y from a human expert then ask the model to generate a *better* decision y'. In a happy coincidence, this suggests a structure for the fine-tuning task that is readily compatible with data from past experiments. In these experiments, multiple decisions were taken for each context, so we fine-tune the language model to convert worse decisions into better ones. This procedure can be viewed as teaching the language model to do hill-climbing in text space. The rationale for this design is that if the human decision is safe but suboptimal, while a language model acting autonomously could be catastrophically bad, anchoring the model on the human decision limits the downside risk. This design is practically useful as well, since at least in the near term, businesses may be reluctant to fully delegate decision-making tasks to AI.

We confirm in a field experiment that our framework performs as intended, with decisions assisted by our tuned language model outperform a human expert. This experimental evaluation ensures that the observed performance improvements are from genuine improvements in decision quality and not artifacts of the estimated reward model. In the next sections, we describe the setting in which we evaluate our framework.

3 Task and Data Description

Email marketing is one of the most common and effective ways for firms to communicate with their customers. 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 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.

We label subject lines with a rich set of semantic tags that describe content and emotion. We will use these tags to steer output from a language model. Details for the descriptive tagging are in Appendix A and for the emotion tagging are in Appendix B.

4 Model

We first estimate a model on existing data, which in the machine learning literature is called "training". Next, we use the trained model to make predictions on new data, which is called "inference". Note that inference here is different from traditional statistical inference.

4.1 Training

At a high level, we want a language model to take downstream business objectives into account when generating. We do this via supervised fine-tuning, which entails taking a pretrained language model, a dataset with (input, output) pairs of the desired behavior, and updating the parameters of the pretrained model on the domain-specific data.

Suppose we ran an A/B test, found that A performed better than B, and had some information about the context in which the experiment was run. We can convert this result into two types of data for fine tuning:

- Input is the decision context, output is the good decision A
- Input is the worse decision B, output is the better decision A

This teaches the model to do two things. The first is a decision suggester – for a given scenario, what is a good decision? The second is a decision improver – given a decision, what is a better decision? The former capability is standard but the latter is novel and practically useful. Observational data for which analogous conclusions can be drawn can be used here as well.

Our A/B test data comes from previous email marketing campaigns. Subject line variations are the treatment groups, and campaigns have up to 16 of them. The topic of the campaign is the context. For fine-tuning, we form training examples in three ways:

- Type 1: Input is a lower performing subject line, output is a higher performing subject line from the same campaign.
- Type 2: Input is a list of keyphrases describing the topic, output is a subject line that outperforms at least 50% of variants in its campaign.

Table 1 Fine-tuning data examples

Type	Input	Output
1	Hot rates are happening now >>> Save on your next getaway during this sale!	>>> Happening Now! You're About To Save Big During This Sale <<<
2	new special offer every week | styles you want | 60% off	To You: Confirming Up to 60% Off the Styles You Want + Stay Tuned for a New Special Offer Every Week
3	Have you tried a lacroix-tail? | coupon inside | lacroix	(1) New notification: Open for drink recipes featuring LaCroix

- Type 3: Input is a subject line and keyphrases, output is a subject line that outperforms at least 50% of variants in its campaign.

Type 1 is decision improver, and Types 2 and 3 are variants of decision suggester. Examples of each type of training data 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. Compared to the GPT family, T5 (Text-to-Text Transfer Transformer) performs especially well when fine-tuned.

Next, we make the emotional valence of the generated output controllable. Past work has shown that the emotional valence of content is a key driver of engagement (Teixeira et al. (2012)), and picking the wrong tone/emotion in the subject line is a common reason for inferior performance. Adapting research in psychology (Gobe (2010)), we develop an emotional ontology which has five categories: PRIDE, TRUST, JOY, ANTICIPATION, and FEAR. Each top-level category is further divided into 3 sub categories (see Appendix B for details). Each subject line in our historical dataset was tagged with the applicable categories. Then, when fine-tuning, the emotional tags of the *output* are appended to the *input*, as shown in Table 2. At inference time, the user can specify the desired emotion tags in the input, which will steer the emotional tone of the output. This fine-tuning technique is called conditional training and was previously used to control style and content in linguistics (Keskar et al. (2019)). This emotion steering step is not required, but it illustrates how explicit structure can be imposed on the blackbox language model if desired.

To assist the language model, we train a predictive model that takes a subject line as input and returns the predicted click rate. We find that linear regression on embeddings from BERT works well (details are in Appendix C). This will be used at inference time to rank outputs from the language model.

Finally, some guardrails must be implemented before deploying our model in the field. The main risk is that the language model could say false, irrelevant, or inappropriate things (Bender et al. (2021)). To address these issues, we train a model of what human experts consider acceptable and apply a filtering step at inference time. We define a binary metric, which we call "accuracy", to take value 1 if the subject line can be used without edits and 0 otherwise. On model outputs

Table 2 Emotion tagging example

Type	Input	Output
1	Hot rates are happening now >>> Save on your next getaway during this sale! | _CURIOSITY_ | _GRATIFICATION_	>>> Happening Now! You're About To Save Big During This Sale <<<
2	new special offer every week | styles you want | 60% off | _ENCOURAGEMENT_ | _SAFETY_ | _EXCLUSIVITY_	To You: Confirming Up to 60% Off the Styles You Want + Stay Tuned for a New Special Offer Every Week
3	Have you tried a lacroix-tail? | coupon inside | lacroix | _CURIOSITY_	(1) New notification: Open for drink recipes featuring LaCroix

Emotion of output is appended to input.

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
  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 accuracy=1 given subject line text. Rather than train a separate model, we add accuracy classification to the pool of tasks on which our T5 model is fine-tuned. For fine-tuning, the input is the subject line prefixed with "SL_Rating", and the output is 0 or 1. For inference, we normalize the relative confidence scores the model places on output tokens of 0 or 1 to obtain a predicted probability that accuracy=1. A few additional guardrails were implemented, details for which are available in Appendix D.

Overall training costs were modest. The T5-base model was fine-tuned on a V100 GPU ($2.50 per hour on Google Cloud) running for 20 hours.

4.2 Inference

For a new email campaign, the user enters a subject line, topic keyphrases, or both. Optionally, emotion tags can be provided. If they are not, we generate candidates for multiple emotion tags and

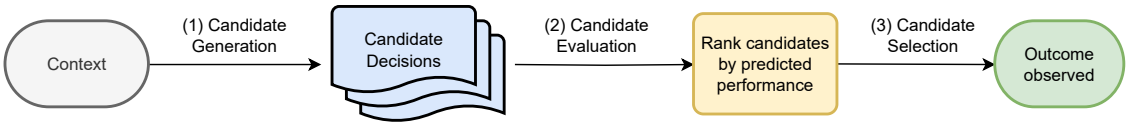

Fig. 1 Inference pipeline: Given a context, a manager makes a decision that induces an outcome. The three key steps are (1) generating, (2) evaluating, and (3) selecting candidate decisions. (1) is done by a language model, (2) is done by a predictive model, and (3) is done by a human. Traditionally, (1) was done by a human, which is challenging in the email marketing copywriting setting.

pass all candidates to the next step, thereby inferring the best emotion for the context. The fine-tuned model generates many (50-100) candidates, from which candidates with predicted probability of accuracy=1 below 0.4 are removed. The predictive model ranks the remaining candidates. The interface presents the top 10 candidates to the user, who makes the final selection. This workflow is summarized in Figure 1.

Inference costs were also modest. The model is hosted on one p2.xlarge server on AWS ($0.90 per hour), and the server is shared by other production models.

Performance may be improved by running the inference pipeline for multiple iterations. The output from one iteration can be the input to the next iteration, which will then return a subject line that is expected to perform better. Doing too many iterations, however, increases the risk of over-optimizing for performance in undesirable ways. We regularize against this by stopping after one iteration, so the results from our tuned language model should be viewed as a lower bound on performance. We leave optimizing the configuration of the pipeline to future work.

5 Results

5.1 Click rates

We evaluate our framework in a field experiment consisting of real email marketing campaigns. The experiment is conducted by a private marketing platform that offers marketing content optimization services to clients from a variety of industries. Our models are deployed through a simple user interface where a Brand Content Strategist (BCS), a non-technical user, can type inputs, click a "generate" button, and receive model outputs.

For each email campaign in the experiment, the control subject line is the non-optimized, human-authored subject line that the brand would usually send. The experimental arms measure the value-added by providing two forms of AI assistance to a human expert. In the first arm, the campaign topic and control subject line are entered as part of a prompt to ChatGPT, which uses a general-purpose language model (gpt-3.5-turbo). Details on the prompt are in Appendix E. The BCS chooses from among the 10 options that ChatGPT suggests. In the second arm, the BCS enters the control subject line and any additional keyphrases as input to the inference pipeline described in Section 4.2. Candidate subject lines are generated by our tuned T5 model, filtered based on the accuracy model, and reranked by a predictive model. The BCS selects a message from the outputs with a predicted performance higher than the control's message.

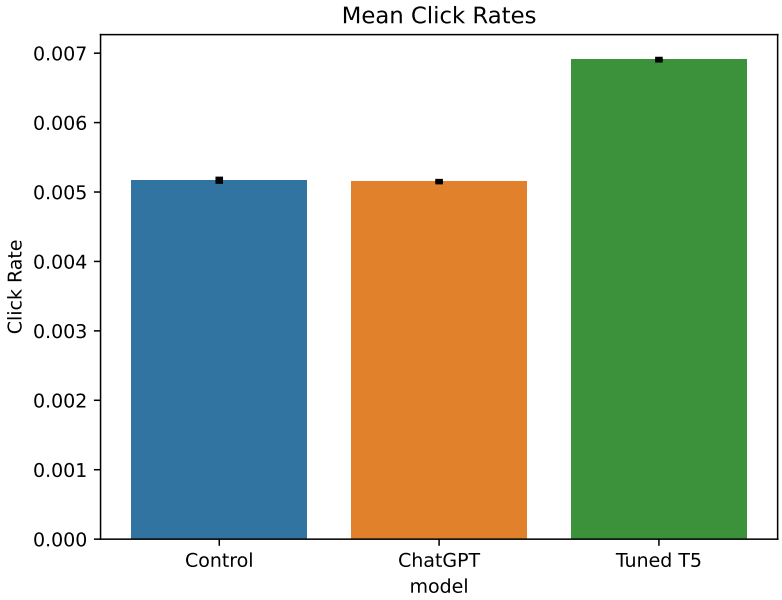

Fig. 2 Mean click rates over deployed campaigns. Assistance from ChatGPT does not outperform an unassisted human, while assistance from the tuned model does. Numerical values for means and standard errors are in Table 3.

Table 3 Mean click rates over deployed campaigns in basis points

	Click Rate (bp)		Count	
Model	mean	s.e.	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

In each email campaign, recipients are randomized over the three arms, and whether the recipient clicked on a link inside the email is recorded. The primary outcome metric is the click rate achieved in each treatment arm.

The experiment is designed to study complementarity: how much is performance improved when an AI helps a human? Our design makes it unlikely that either of the experimental arms will perform worse than the control. This is helpful for feasibility of the experiment; reducing the downside risk of a new technology makes business partners more willing to agree to test it. The design is also realistic – at least in the early days, we expect that AI assistance in decision-making tasks will be deployed with a human expert having oversight. We leave questions of substitutability (i.e. to what extent AI can replace humans) to future work.

In 36 campaigns totalling 283 million impressions, we find that on average, the ChatGPT-assisted arm does not improve on unassisted humans while assistance from our model achieves an average lift of 33%. Average outcomes from aggregating over campaigns are shown in Figure 2, with raw numbers in Table 3.

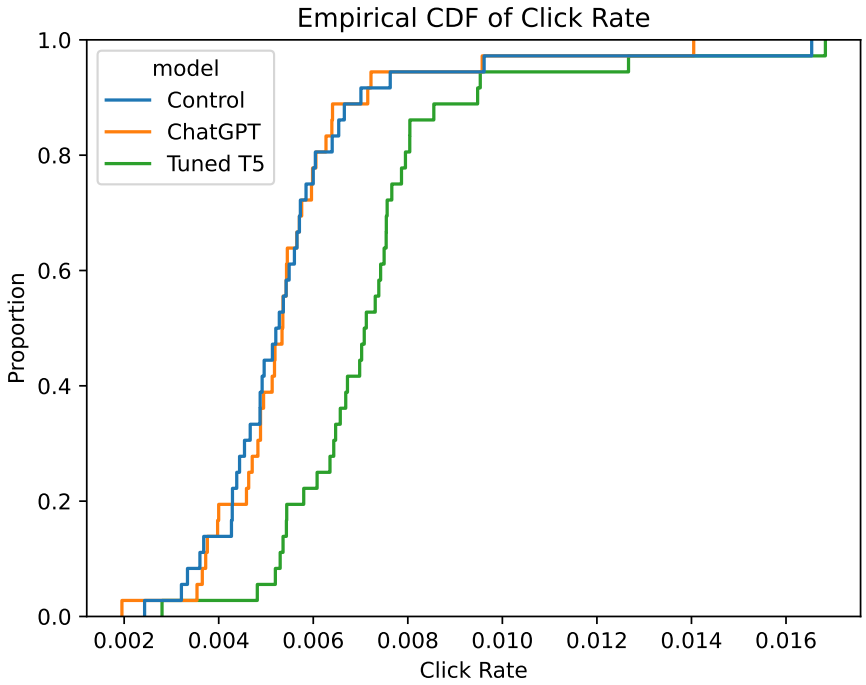

Fig. 3 Marginal distributions of click rates over campaigns. The outcome distribution from our tuned model first order stochastically dominates the distribution from the control and ChatGPT. Its CDF is shifted to the right, which means that every percentile is larger. ChatGPT assistance does not help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

For the ChatGPT-assisted treatment arm, Figure 3 compares the full distribution of outcomes to the control's. While ChatGPT assistance does not lead to better outcomes, it does not systematically harm performance, which shows the robustness of our human-proposing/AI-improving paradigm. The left panel of Figure 4, which plots joint outcomes from individual campaigns, shows that ChatGPT tends to improve outcomes in campaigns where humans perform worse while worsening outcomes where humans perform better. This implies there could be value in selectively allocating assistance from ChatGPT.

For our tuned model, Figure 3 shows that it first order stochastically dominates both the control and ChatGPT. That is, every quantile of the tuned model's performance is larger than that of the control and ChatGPT. Equivalently, any decision maker with nondecreasing utility (i.e. everyone) prefers the distribution of outcomes under our tuned model. In the right panel of Figure 4, we see the strongest comparison of relative performance: for 35 of 36 email campaigns, our tuned model outperforms the control.

Assistance from our fine-tuned language model unambiguously adds value but simply prompting a general-purpose model for high-performing text does not (yet). At least for now, fine-tuning or some form of domain-specific adaptation is necessary to unlock the full value of A/B test data.

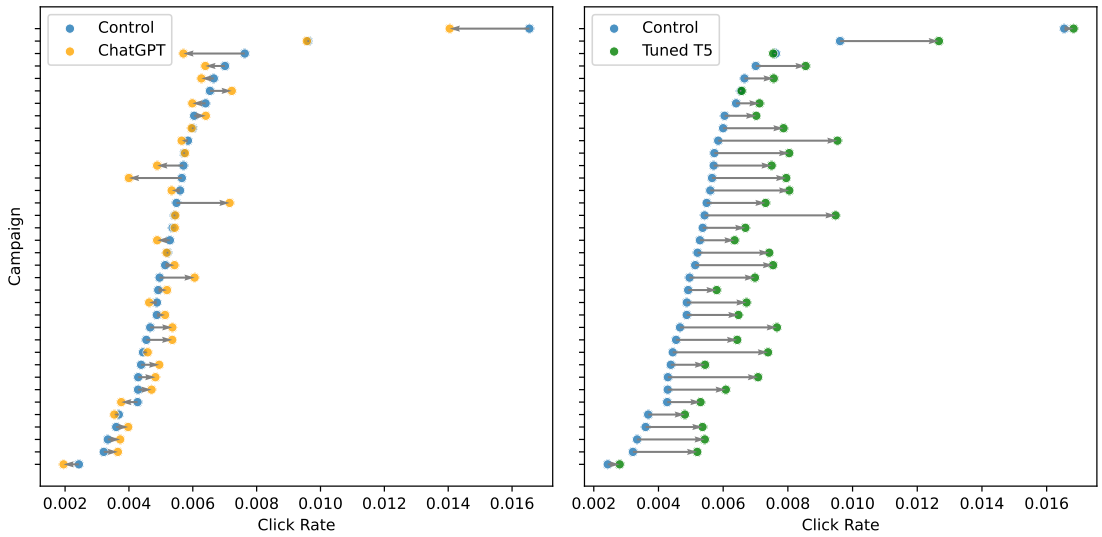

Fig. 4 Joint distribution of click rates, by campaign. Relative to the control, our tuned model improves outcomes (shifts points to the right) for 35 out of 36 campaigns. ChatGPT shows mixed results, improving on 19 out of 36 campaigns, but it has a positive effect in campaigns where humans perform worse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

5.2 Complementarity

Next, we investigate how the gains from AI assistance vary across campaigns. Consider the causal logistic regression model

$$Pr(Y_i = 1|T_i) = \frac{1}{1 + \exp\{-(\alpha_k + \beta_k T_i)\}},$$

where Y_i is an indicator for whether recipient i clicked on a link inside an email, T_i is an indicator for whether recipient i was assigned to treatment, and (α_k, β_k) are campaign-specific coefficients for campaign k. Denoting the log odds ratio with $\text{logit}(p) := \frac{p}{1-p}$, we can rewrite the logistic regression model as

$$\text{logit}(Pr(Y_i = 1|T_i)) = \alpha_k + \beta_k T_i,$$

which makes apparent that β_k is the treatment effect of AI assistance on the log odds ratio of clicking on a marketing email in campaign k, while α_k is the log odds ratio in the control group. $\text{logit}(\cdot)$ is monotonic and increasing, so α_k can be interpreted as a measure of the click rate performance of the control group.

We compute the odds ratios for each campaign and treatment arm and compare the resulting values of β_k against α_k. For campaigns where the control performs worse, which is consistent with high difficulty or low effort/ability by the human, the language models are more helpful. This is true for both the general-purpose and tuned models and is illustrated by the negative slopes in Figure 5.

We quantify this relationship by running the following (descriptive) regression across campaigns:

$$\beta_k = \gamma_0 + \gamma_1 \alpha_k + \varepsilon_k$$

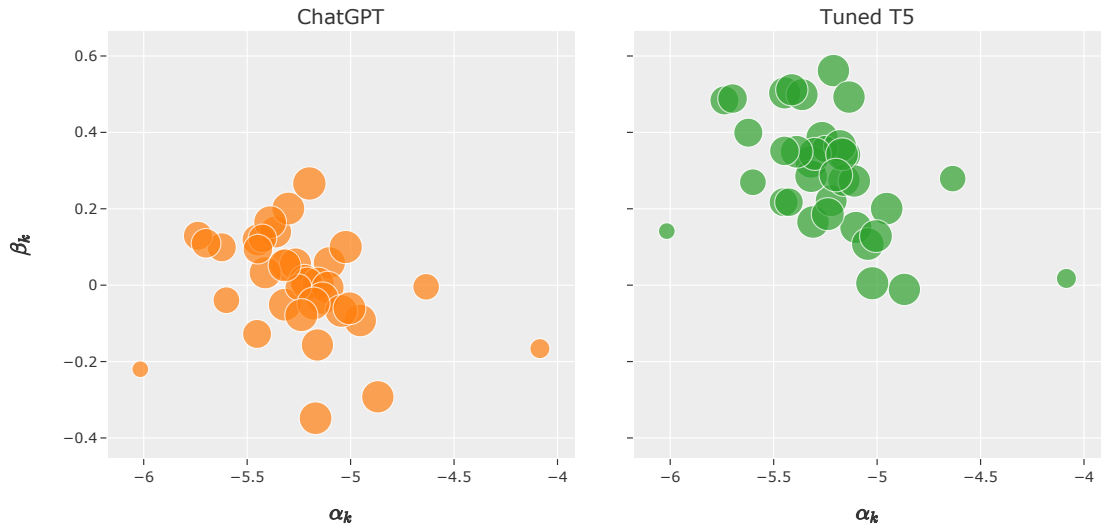

Fig. 5 Treatment effect from model assistance (β_k) vs. baseline performance (α_k), measured in log odds ratios. Email campaigns from our field experiment have a negative correlation between the treatment effect from model assistance (β_k) and the performance of an unassisted human (α_k). This is consistent 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.

Table 4 Treatment Effect vs Control Click Rate for each model across test campaigns

	Dependent variable: β_k	
Model	ChatGPT	Tuned T5
Intercept	-0.987^{**}	-1.247^{***}
	(0.421)	(0.435)
α_k	-0.190^{**}	-0.295^{***}
	(0.080)	(0.083)
Observations	36	36
Adjusted R^2	0.116	0.249
F Statistic	5.599^{**}	12.618^{***}
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$	

Separate regressions are estimated for the general-purpose and tuned model treatments, and standard errors are calculated by weighting each campaign by its number of impressions. Results in Table 4 confirm the negative relationship. The negative regression coefficient is robust to the inclusion of brand fixed effects.

5.3 Content quality

Lastly, we investigate the quality of model-generated content. For each of 192 inputs, we generate 5 inputs from both ChatGPT and our tuned model. The 10 results were shuffled and eight expert evaluators were asked to rate them as 1 or 0 based on the accuracy criteria described in Section 4.1.

Table 5 Quality Metrics for ChatGPT and Tuned T5

Model	N	Accuracy	BERTScore	Self-BLEU
ChatGPT	960	65.6%	0.687	0.707
Tuned T5	960	**71%**	**0.726**	**0.556**

Accuracy scores are obtained directly from human experts. BERTScore measures relevance of output to input (0-1), higher is more relevant. Self-BLEU (0-1) measures diversity of output, lower means higher diversity.

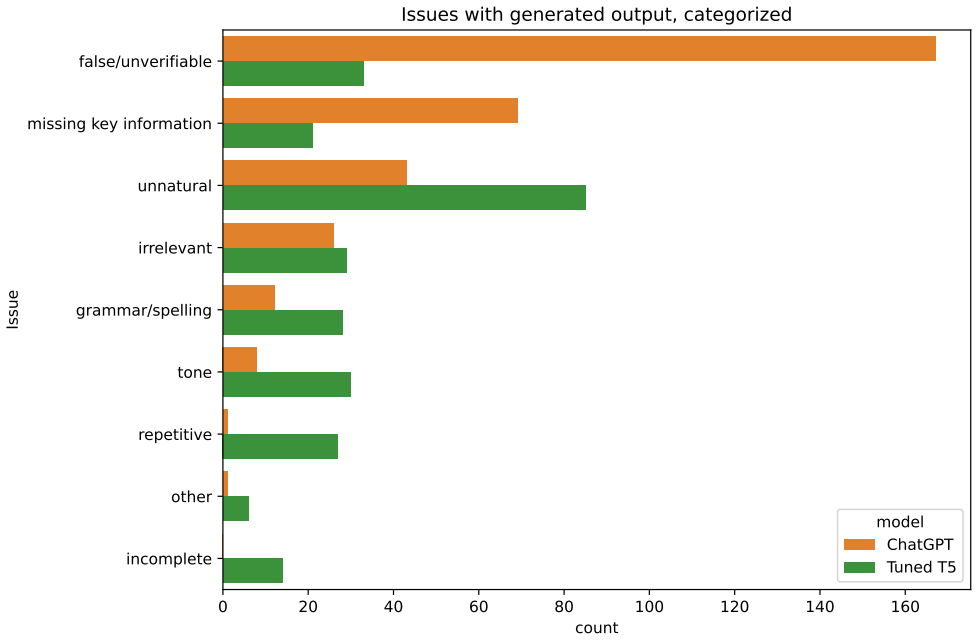

Fig. 6 Issues with AI-generated output

Our model performs slightly better than ChatGPT, achieving overall accuracy of 71% compared to 65.6%.

We also calculate two automated metrics. BERTScore (Zhang et al. (2019)) measures relevance of the output to the input on a scale of 0 to 1 and has been shown to correlate well with human judgments of quality. Higher values mean the output is more relevant. Self-BLEU (Zhu et al. (2018)) measures the diversity of generated output on a scale of 0 to 1. Lower values mean that outputs are less similar to each other, i.e. higher diversity. In our setting, diversity is desirable because it increases the chances of choosing the correct emotional tone. Eventually we expect there will be a tradeoff between relevance and diversity; additional diversity after a point will come at the cost of relevance. We find that our model generates more relevant (higher BERTScore) and more diverse (lower self-BLEU) compared to ChatGPT – it attains a more efficient point on the diversity-relevance frontier. Table 5 shows the specific numerical values.

For each output with 0 accuracy rating, we asked evaluators to categorize into a further set of categories as shown in Figure 6, with example instances shown in Tables 6 and 7. The main

Table 6 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
false/unverifiable	Input: get personalized quote | up to 15% off or more on car insurance | just 15 minutes Output: Get Your Free Quote Today and Save More than 15% on Car Insurance	"Free" 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 7 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
grammar/spelling	Input: documentary films Output: We think you'll really like this documentary films.	"this" is grammatically incorrect, should be "these"
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

issues for ChatGPT are "false/unverifiable" and "missing key information". This shows that our model successfully learned to include factual information only. Conversely, the main issues with our model are stylistic rather than semantic, the biggest being "unnatural". This is largely due to the limitations of a smaller language model, as T5-base has 30x fewer parameters than ChatGPT. Even with a smaller size, our fine-tuned model outperforms ChatGPT on content quality overall.

6 Conclusion

We developed and validated a framework that teaches language models to improve marketing decisions using past A/B tests, providing a solution to the otherwise intractable problem of optimizing unstructured content. Through a large-scale field experiment in email marketing, we established two key findings about AI deployment in this domain: first, that small language models are sufficiently powerful for improving outcomes when properly adapted to the specific task, and second, that having humans propose initial content while AI suggests improvements provides an effective safety mechanism. This framework enables firms to extract generalizable insights from A/B tests while maintaining appropriate controls over AI-generated content.

Our results imply that firms should value A/B tests not just as a tool for making decisions in the moment but also as a strategic asset for improving future decisions. In any setting where firms conduct A/B tests to make decisions over unstructured decision variables, like the content of advertisements or the design of a website, our framework enables firms to move beyond comparing alternatives to optimizing the decisions directly. While we focus on text for marketing communications, the framework extends to other types of content like images and to any measurable objective. A compelling application in the public sector is optimizing the content of nudges, such as reminders to renew student financial aid or SNAP benefits, which can have substantial social impact.

Our framework takes experimental data as given and shows how to extract additional value from it by generating optimized treatments. Modifying the experimental design to take these generated treatments into account could be promising, perhaps by adapting the methods in Dew (2024) or Ye et al. (2024).

Given our finding that AI assistance improved outcomes the most in instances where humans performed worse, future work should investigate effective ways to combine human judgment and AI assistance. In a radiology diagnosis setting, Agarwal et al. (2023) showed that the best combination of humans and statistical models varies significantly with the task due to nuances in human information processing. Similarly, in risk assessments, Green and Chen (2019) found that the presentation format of algorithmic suggestions influences the accuracy and bias of human judgment. Extending this research beyond binary classification tasks to more complex tasks like text generation would provide valuable insights into enhancing decision-making processes.

In this study, our focus was on treatment heterogeneity rather than individual-level heterogeneity. While not directly comparable, the magnitude of improvement from modifying the treatment is similar to that from optimizing targeting. Our optimization of a blanket treatment yielded a 33% increase in top-of-funnel engagement, while studies by Hitsch et al. (2023) and Ellickson et al. (2023) report about an 11% gain in profit from targeting. An advantage of working with blanket treatments is that they do not require personal data collection, a growing concern amid evolving privacy laws. Future work could investigate whether combining our content optimization framework with targeting yields additional gains.

On a technical level, our formulation resembles actor-critic methods in reinforcement learning, where a policy network (in our case, a language model) generates actions and a value network (our

predictive model) evaluates them. While we employed supervised fine-tuning to train our language model, reinforcement learning methods like RLHF (Christiano et al. (2017)) or DPO (Rafailov et al. (2023)) might prove more effective by directly optimizing for the desired objective, though they are more complex to implement.

To close, the success of our empirical application positions marketing decision-making in high-dimensional action spaces as the latest domain where inductive, data-driven methods prove to be effective. By showing how to safely align generative AI with business objectives while preserving human agency and oversight, we provide actionable guidance for firms seeking to optimize unstructured decisions across marketing and beyond.

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.

Christiano, P.F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, Volume 30.

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.

Dai, A.M. and Q.V. Le 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, Volume 28.

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

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

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.

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 .

Hitsch, G.J., S. Misra, and W. Zhang. 2023. Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957* .

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.

Polanyi, M. 1966. *The Tacit Dimension*. Routledge.

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 .

Teixeira, T., M. Wedel, and R. Pieters. 2012. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research 49*(2): 144–159 .

Wei, J., M. Bosma, V.Y. Zhao, K. Guu, A.W. Yu, B. Lester, N. Du, A.M. Dai, and Q.V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* .

Yang, Y. and P. Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management 59*(2): 102853 .

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* .

Zhang, T., V. Kishore, F. Wu, K.Q. Weinberger, and Y. Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* .

Zhu, Y., S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100.

Appendix A Keyphrase extraction

We use descriptive semantic tags for subject lines, i.e. keyphrases, to steer generated content. A keyphrase is a group of words that represents the main topic of an email campaign, such as products being promoted (e.g. bags, Nike Air VaporMax 2021), special offers (e.g. 20% off, free shipping), or holiday events. A good keyphrase has the following characteristics:

- Relevance: accurately describes the main topic
- Clarity: not confusing or misleading in any way
- Conciseness: short but conveys all necessary information; "new organic linen sheets" preferable to "linen sheets"
- Specificity: detailed instead of generic, e.g. "The Great Room Event" instead of "furniture sale"

Some subject lines in our data have been tagged by humans. For untagged subject lines, we trained a keyphrase extraction model to automate the tagging using a separate fine-tuned T5-base model. Inputs are subject lines and outputs are human-tagged keyphrases. The training parameters are the same as in Section 4.1.

Appendix B Emotion extraction

We use an emotional ontology with 15 categories, grouped into 5 top-level categories, as shown in Table B1. By making the emotion of the language model's output controllable, we can enumerate over all the emotions, generate output for each, and automatically select the best one for the topic.

As was the case for keyphrases, some subject lines in our data have been tagged by humans. For untagged subject lines, we fine-tuned a RoBERTa-base (Liu et al. (2019)) model using subject

Table B1 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

lines as inputs and emotion tags as outputs. We add an output layer with 15 nodes on top of the RoBERTa model and apply softmax activation to produce the probability distribution over 15 emotion classes. The loss function for fine-tuning is a categorical cross-entropy loss.

Appendix C Predictive model

Factors such as the intrinsic appeal of the product, the preferences of the recipient, or seasonality affect the success of each campaign. To isolate the effects coming from the mechanics of language, we normalize the data to remove overall campaign effects. We tried several model specifications and found that BERT embeddings and linear regression performs well. Adjusting BERT's tokenizer to align with marketing language and terminology led to a 4% increase in overall performance. Hyperparameter tuning was performed on a separate validation set using the Ray library. Full results are in Table C2.

Table C2 Root Mean Squared Error (RMSE) metrics of regression models

Model	Test RMSE
Linear regression (baseline)	0.927
T5 embeddings + Linear regression	0.924
Fasttext + CNN Multigram	0.911
Fasttext + BI-LSTM	0.917
DistilBert + custom vocabulary	**0.883**

Table D3 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 D Additional guardrails and post-processing

In addition to checking generated subject lines for accuracy, 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 D3. Generated subject lines classified into any of these categories with probability more than 0.8 were removed.
2. Diversity check

   To reduce redunadancy, 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.

Appendix E 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"*