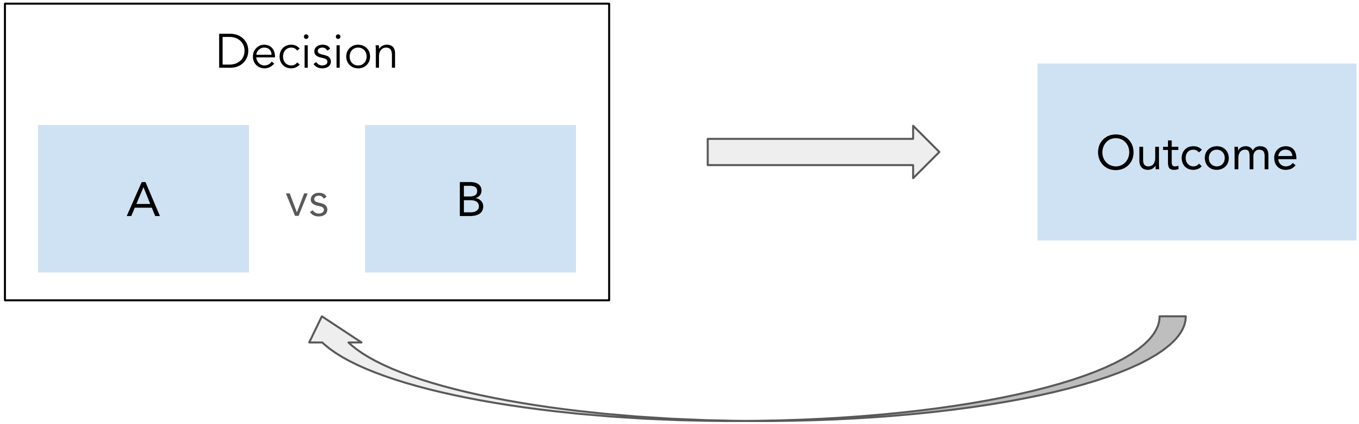
Causal Alignment:
Augmenting Language Models with A/B Tests

Panagiotis Angelopoulos, Persado
Kevin Lee & Sanjog Misra, Chicago Booth

December 6, 2024

(Previous title: Value Aligned Large Language Models)

Extrapolating from A/B tests using generative AI



- Firms conduct A/B tests to optimize: price, product features, ad content, etc
- Want: informative guidance for untested decisions and new contexts
- For price, fit demand and solve, but can't do this for unstructured decisions

Formally

Context *x*, decision *y*, reward *r(x, y)*:

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

If *r* differentiable, gradient ascent.

If *y* is unstructured, guess and check?

Current best practice:

1. Fine-tune θ:
   $$\max_\theta \; \mathrm{E}_{y \sim G(y|x;\theta)}[r(x, y; \phi)]$$
2. Generate $y^* \sim G(y|x;\theta)$

Challenge: Full delegation to AI can be too risky!

BUSINESS

Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

FEB 17, 2024 12:12 PM

Overview

1. We develop a general framework for optimizing the content of marketing communications from A/B test data

2. We provide experimental validation that our method is “effective” and “safe”

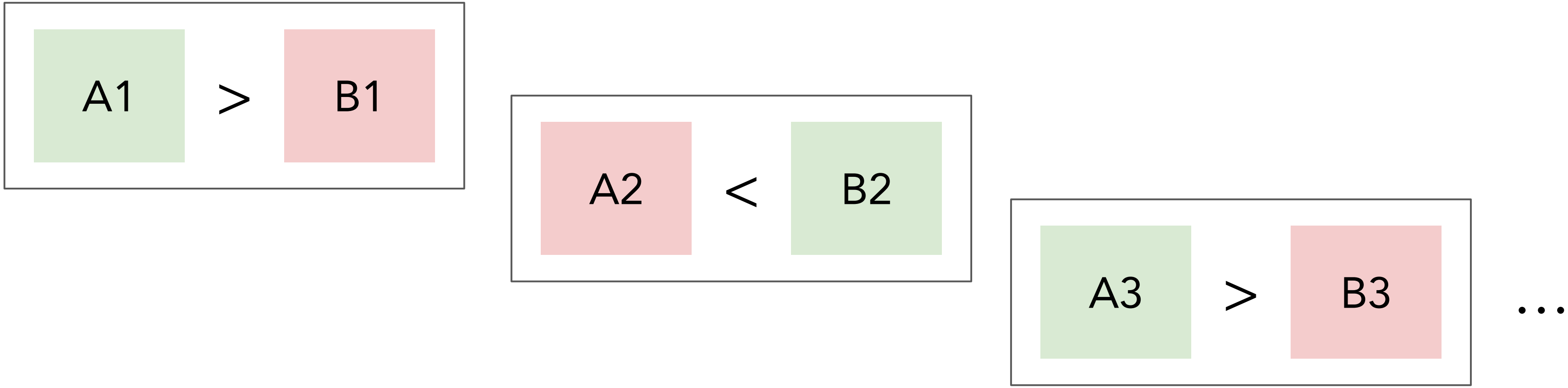
3. Our method uses existing data so can be implemented immediately

Framework: Teach language model to hill-climb on past A/B tests

- **Idea**: If A outperformed B, train language model to convert B to A
- For a new decision, human comes up with a candidate decision, then the language model improves.
- This design reduces risk of harm compared to full delegation to an AI

Broadly applicable to optimization problems over unstructured decision variables

Intuition: Extracting information from multiple A/B tests



- An experienced copywriter can pick out patterns from past A/B tests
- We extract this information using a language model
- We teach the AI to improve humans from ordinal comparisons, which coincides with format of experimental data

Field experiment: Email marketing

Goal: show our framework works in a practical setting

- Email subject lines matter a lot! Affects click-through rate 73%-445%
- Traditionally relies on human experts to craft something catchy and relevant
- Seems like AI could add value! But things could go wrong
    - Don’t want to achieve high open rates by saying false/sensational things

Safety considerations are first-order when deploying AI

Optimizing an LLM to a task creates new issues (Amodei et al. (2016)):

1. Robustness: Will the LLM say something nonsensical that performs poorly?

Solution: instead of generating from scratch, improve on human input

2. Reward hacking: can increase engagement by being inflammatory/offensive.

Solution:

- Impose structure - make emotional valence of output controllable
- Guardrails - learn a model of acceptable output and filter generated output

Training data

- 20,000 campaigns over 10 years from a digital marketing platform

- Diverse industries – retail, e-commerce, fashion, financial services, and insurance – and 337 well-known brands

- Campaigns randomly assign median 800k recipients to 16 subject line variants and record click-through rates

Fine-tuning task for language model

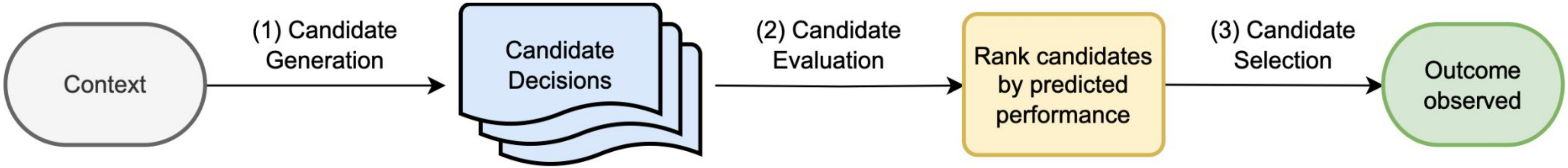
Input	Output
Hot rates are happening now >>> Save on your next getaway during this sale!	>>> Happening Now! You're About To Save Big During This Sale <<<

Impose structure: controlling emotional valence of output

Input	Output
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 <<<

Ref: CTRL, Keskar et al. (2019)

Experiment evaluates our framework against 2 alternatives



Control: human expert creates subject line as usual

Treatment 1: ChatGPT generates improvements to control subject line

Treatment 2: Our tuned language model generates improvements to control

Field experiment results: mean CTR increase of 33%

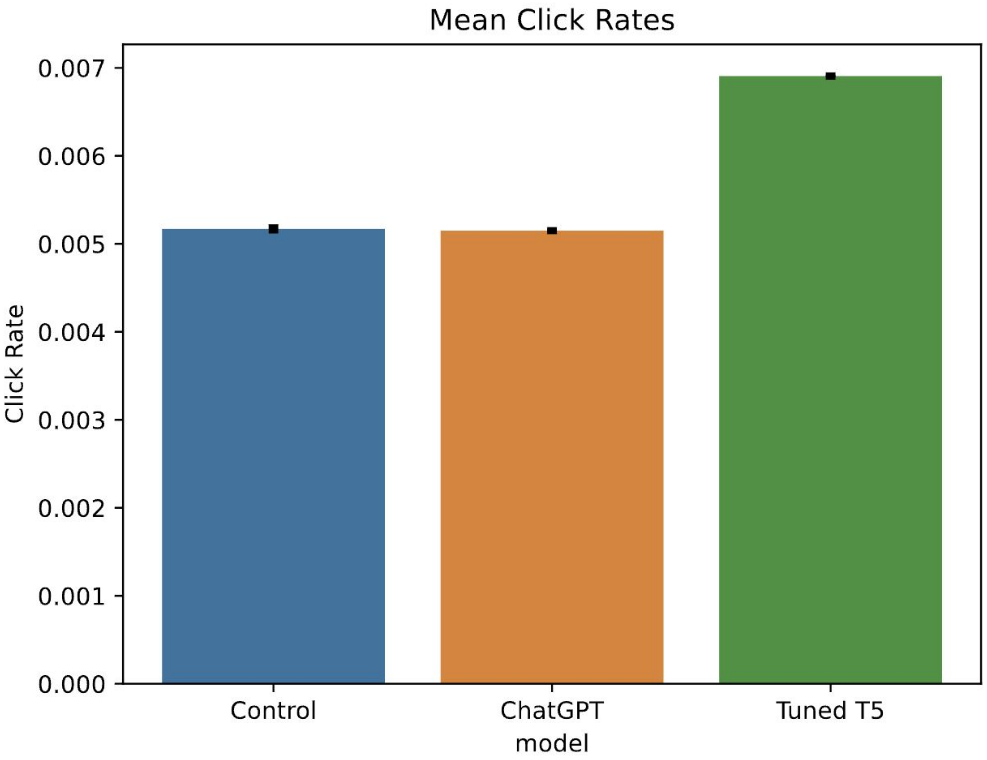
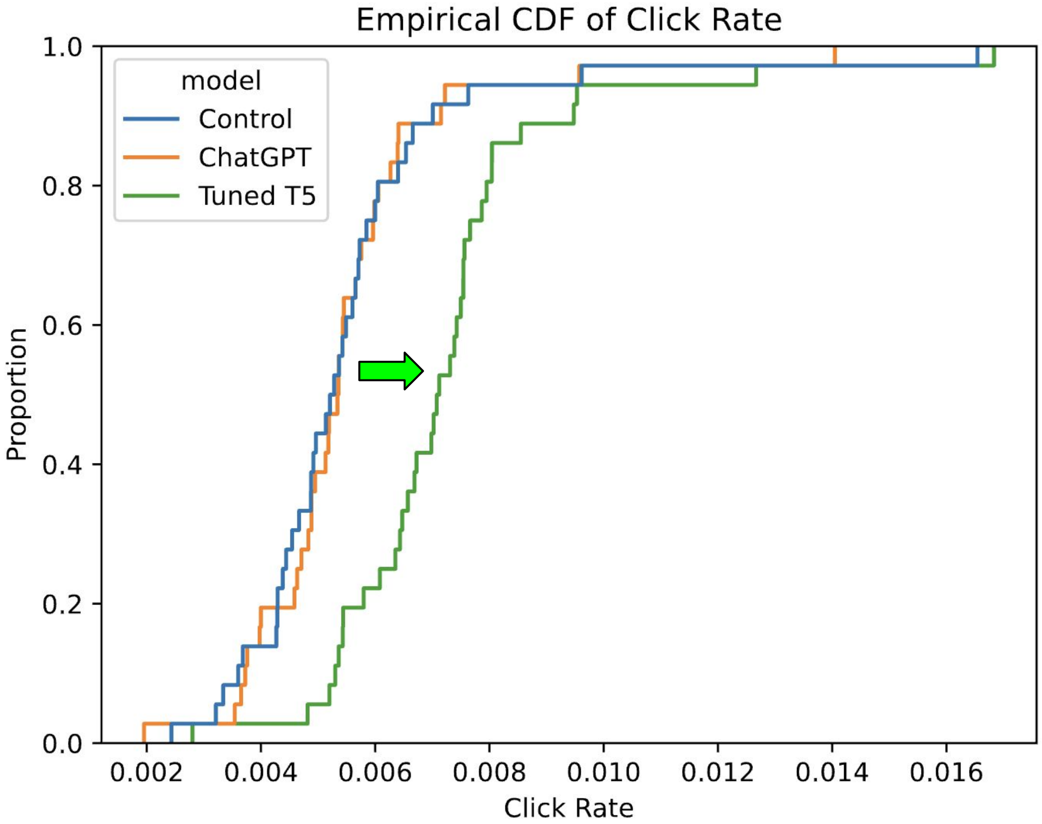


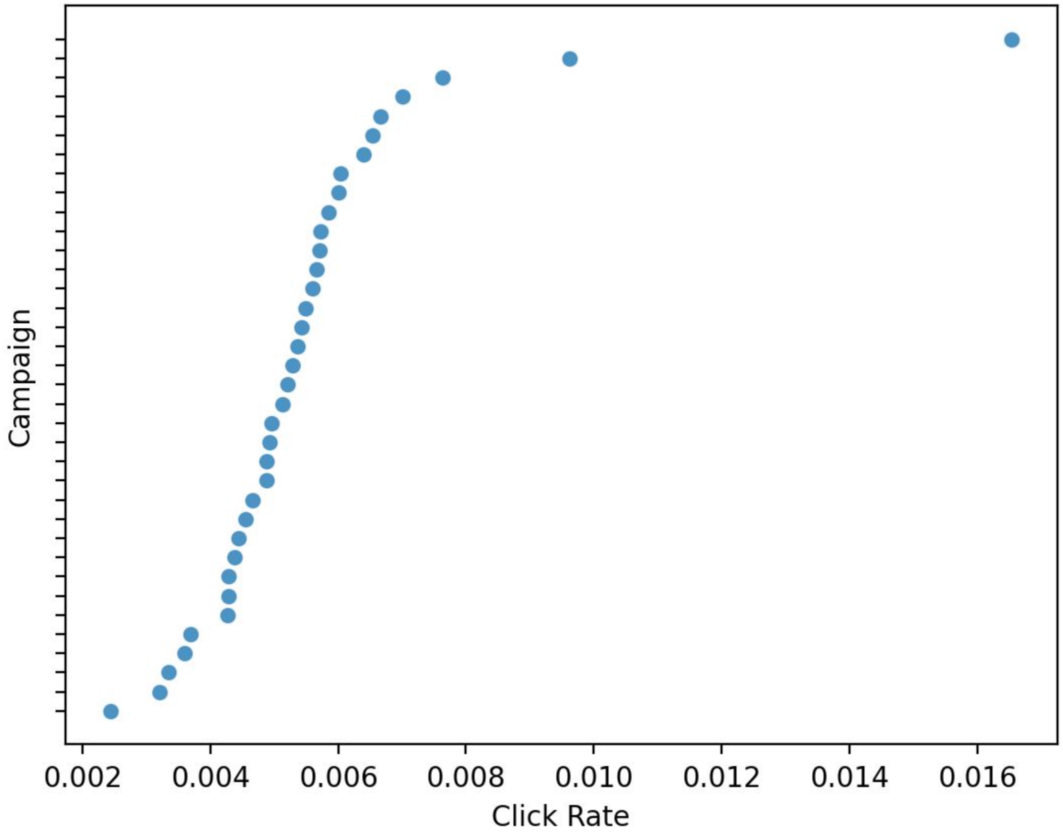
Table 3 **Mean click rates over deployed campaigns, in basis points**

	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

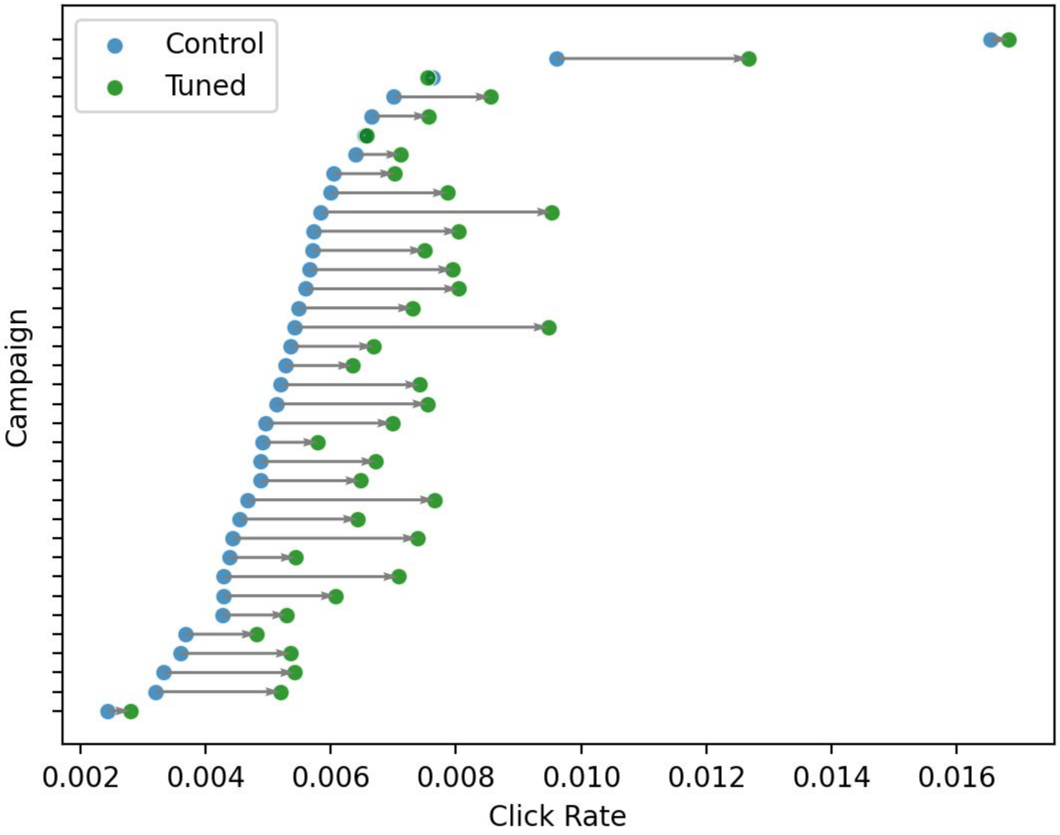
Stochastic dominance: every quantile is better



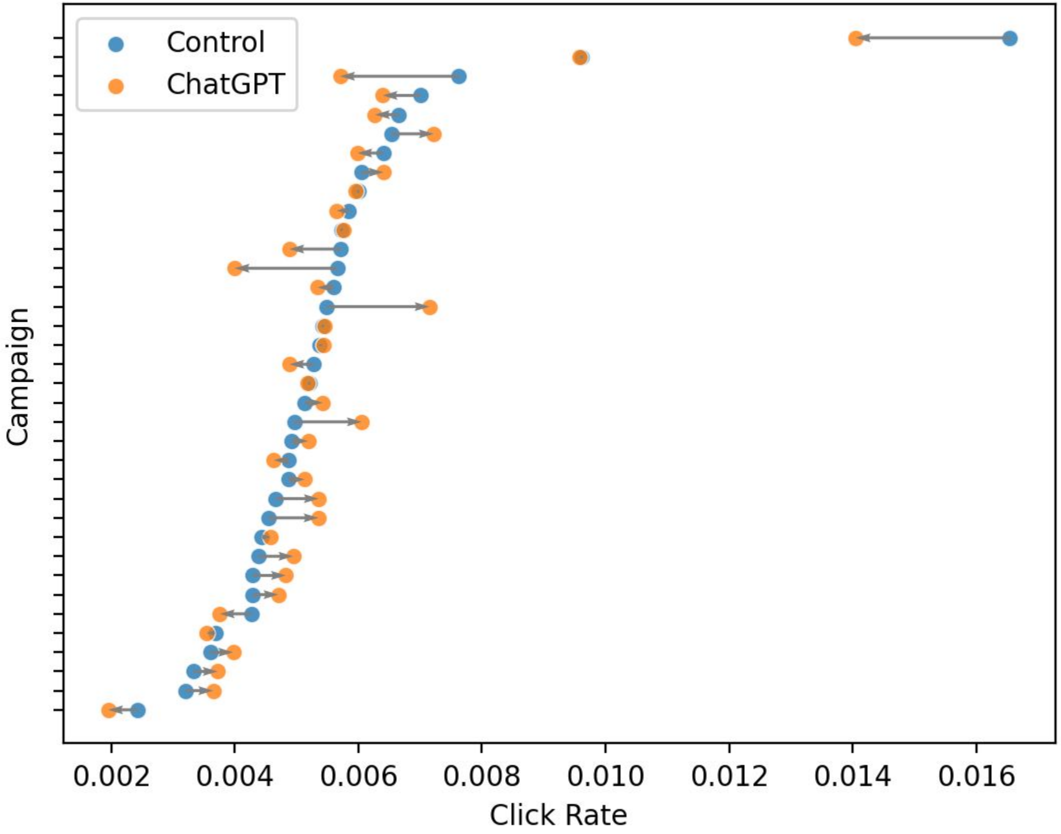
Performance of unassisted human across 36 campaigns



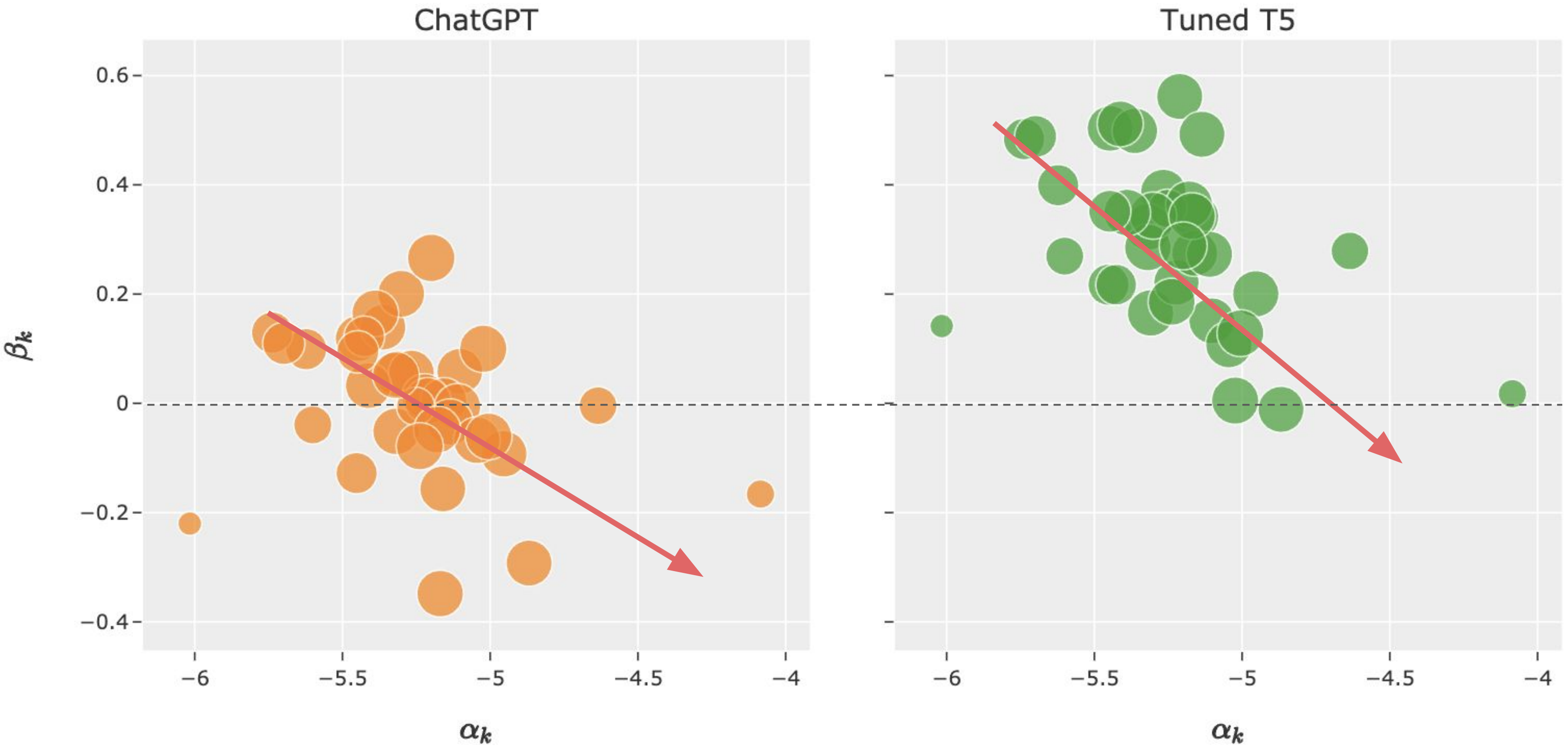
Assistance from our tuned model improves performance



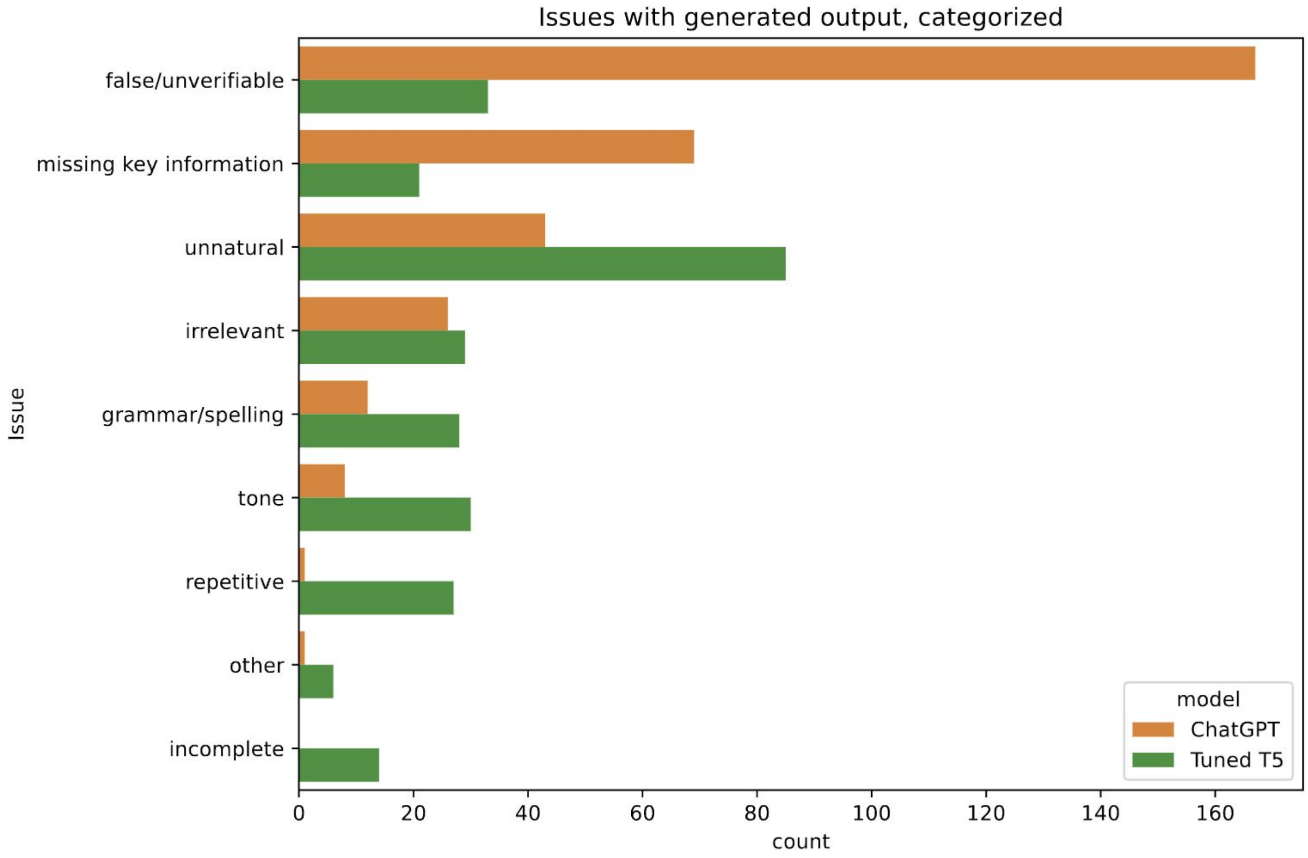
ChatGPT doesn't improve (but doesn't harm!) performance



Treatment effect of AI assistance (β_k) vs control performance (α_k) across campaigns



Better accuracy at the cost of some fluency



Mechanism: Change in feature activations

Most amplified ⬆

1. Phrases emphasizing choice and decision-making
2. References to collaboration and collective effort
3. References to the pronoun "you"

Most suppressed ⬇

1. Statements related to social media interactions
2. Emojis representing emotions or food
3. Numeric values and percentages

Note: These are differences in loadings on features extracted by Gemma Scope, a pretrained sparse autoencoder.

Examples of “what to do” and “what not to do”

Most amplified

1. You've been selected to shop sunny-day styles for less
2. We're happy to announce up to 70% off select tabletop & home décor
3. We're treating you! You're getting up to 70% off Easter essentials

Most suppressed

1. Weekend plans = shopping! 🛍️ Add up to 75% off Daily Deals to your cart now
2. Don't worry, be hoppy! 👯 There's still time to save up to 75% on Easter must-haves
3. Ready to redecorate? Save up to 70% on home must-haves

Note: These are actual data points that maximally activate each feature.

Discussion of results

For AI to improve performance:

- Domain-specific data linking text to outcomes is necessary
- *Small* language model is sufficient (T5-base is 30x smaller than gpt-3.5-turbo)

To safely deploy AI:

- Design task to complement human
- Impose structure
- Filter out undesirable output

Conclusion

- We develop and validate a framework that teaches language models to improve marketing content using past A/B tests

- Framework enables firms to move beyond comparing alternatives to optimizing the content itself, which was previously intractable

- Training on available data + low cost ($50, 20 hours in 2023) + experimental validation → shows how firms can deploy AI to deliver value immediately

Future work

- Other modalities (images), objectives (nudges), and types of causal data
- Heterogeneity (targeting + personalization), experimental design, task design
- Lots to be done! Possible to extend predictive models to prescriptive ones

Thank you!

kevin.lee@chicagobooth.edu