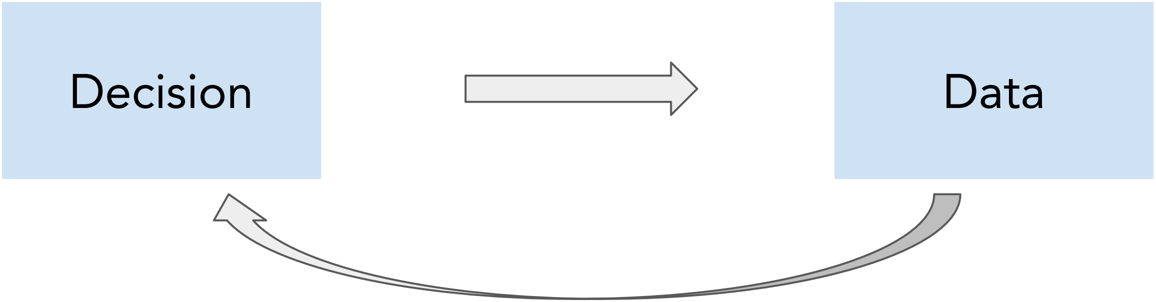
Causal Alignment:
Augmenting Language Models with A/B Tests

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

ESIF-AIML, August 14, 2024

(Previous title: Value Aligned Large Language Models)

Data-driven decisions



Decisions

- Product features
- Price
- Promotion content

Methods

- A/B tests
- Predictive models
- Generative models (!)

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?

Alternative:

1. Generate $y^* \sim G(y|x;\theta)$
2. Fine-tune θ:

$$\max_\theta \; \mathrm{E}_{y\sim G(y|x;\theta)}[r(x, y; \phi)]$$

But full delegation of decision to AI can be too risky!

Framework: Fine-tune language model on A/B tests

- **Idea**: If A outperformed B, train language model to convert input B to output 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

Findings

1. A/B tests are a useful source of feedback for aligning language models

2. Our framework shows how to do this: “A better than B” means “turn B into A”

3. In a field experiment, we show that our framework delivers performance improvements in *new* decision contexts

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

Framework is evaluated against 2 alternatives

- Old way: train a model to predict performance of marketing content. Human comes up with ideas, uses predictive model to sort.
- New way (?): Can we just ask ChatGPT "give me high-performing emails/ads on {topic}"?

Challenges:

- How to leverage data from past marketing campaigns?
- How to ensure factual accuracy/reasonable performance by ChatGPT?

Data

- 20,000 campaigns over 10 years from a marketing platform
- Diverse industries – retail, e-commerce, fashion, financial services, and insurance – and 337 well-known brands
- Campaigns have median 800k recipients:
    - Randomly assigned to 16 subject line variants generated from a template
    - Click-through rates recorded

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 <<<

Controlling emotional valence

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 <<<

Safety considerations

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

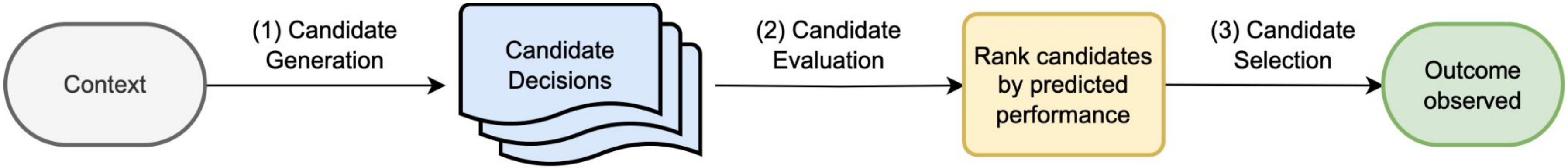
1. Reward hacking: can increase engagement by being inflammatory/offensive.

Solution: learn a model of acceptable output, filter generated output

2. Performance on new data: Will the LLM say something nonsensical?

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

Experiment: Measure effect of AI assistance



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 increase of 30%

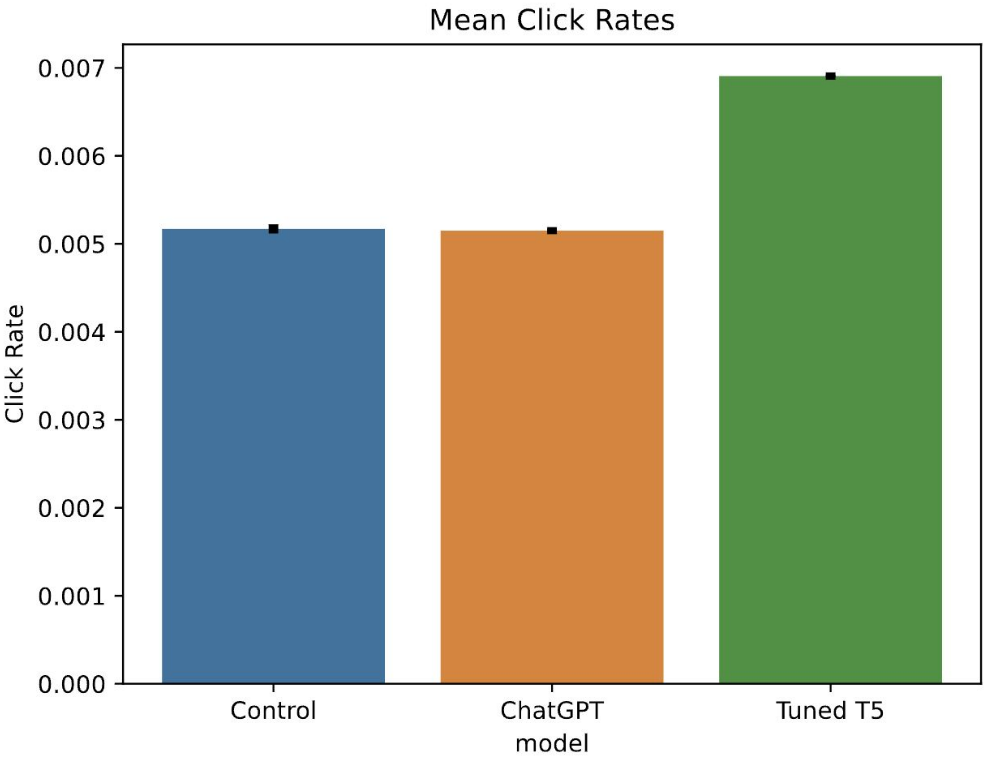
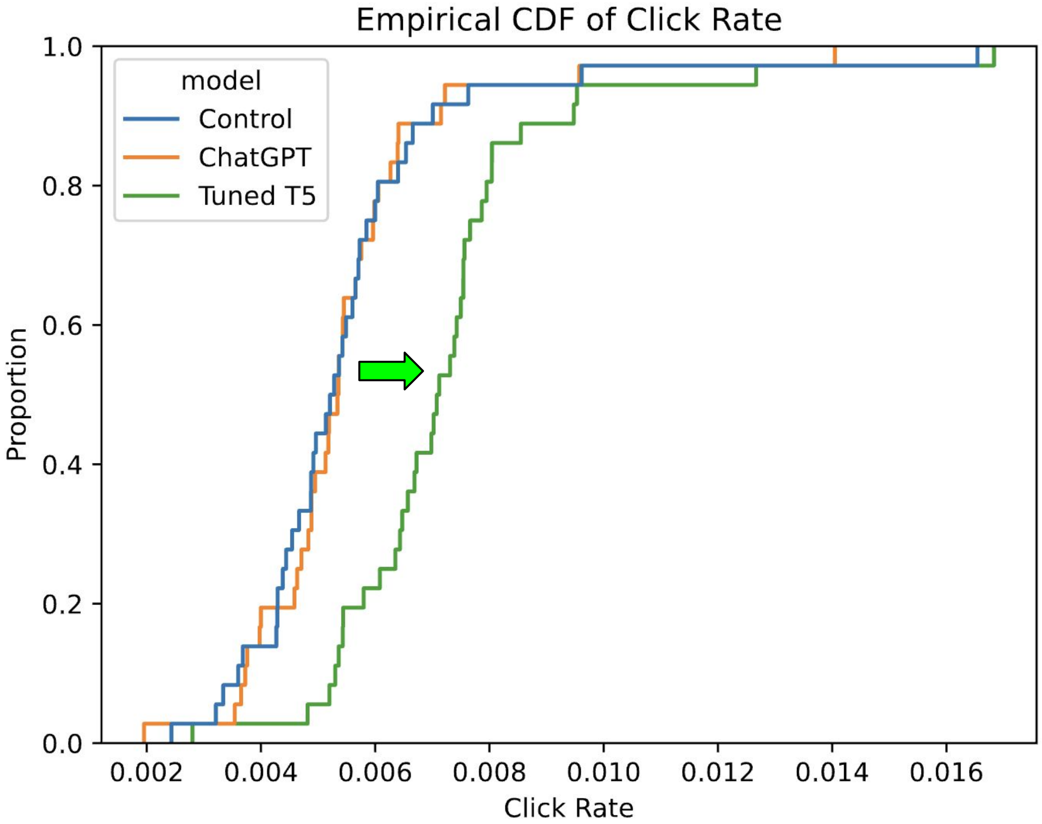


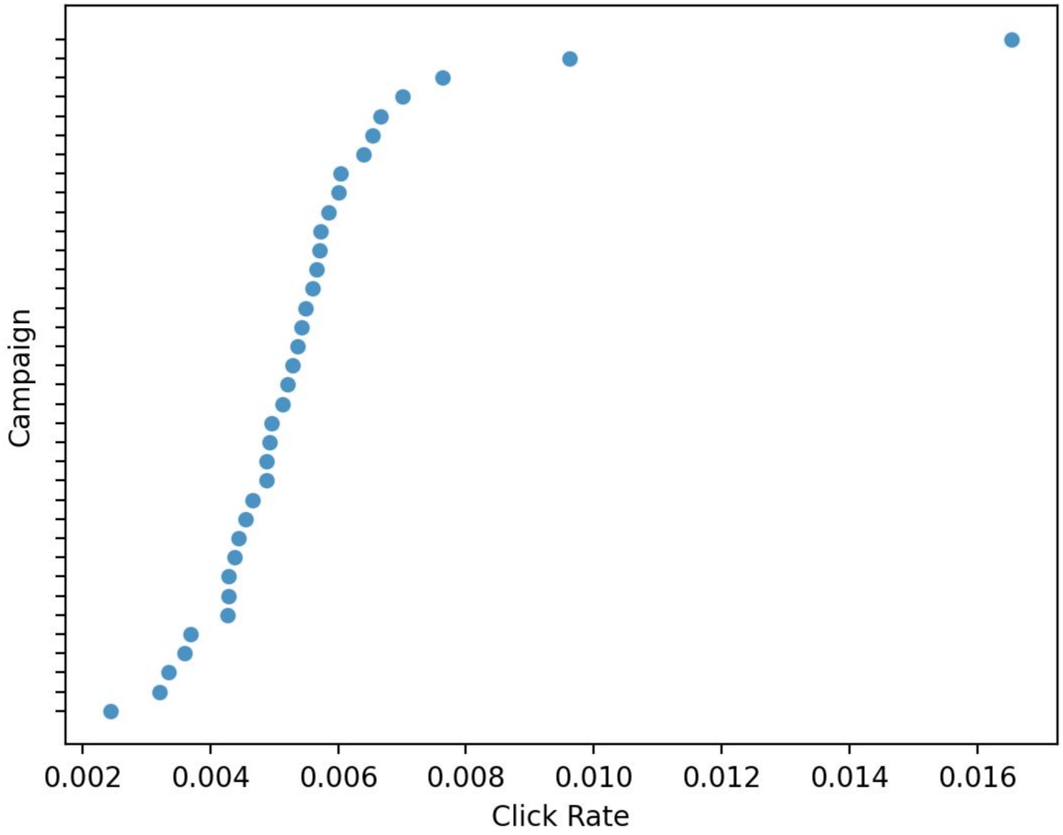
Table 3 **Mean click rates over deployed campaigns, in basis points**

	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

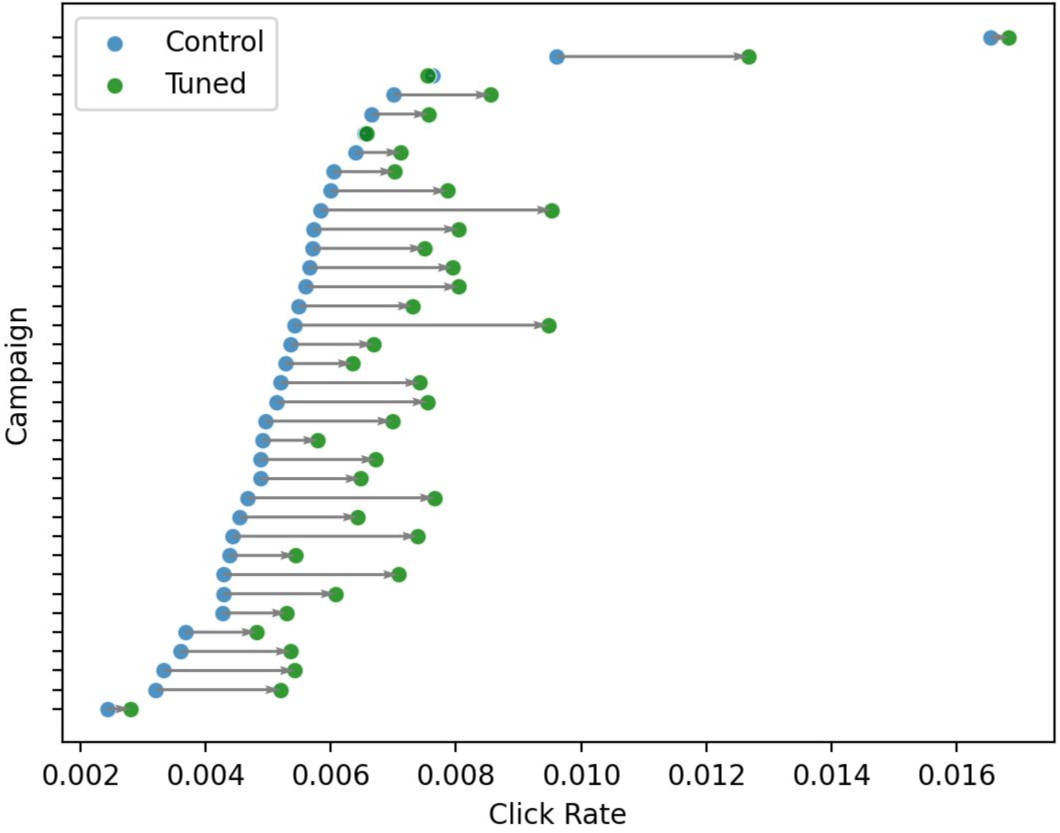
Stochastic dominance: every quantile is better



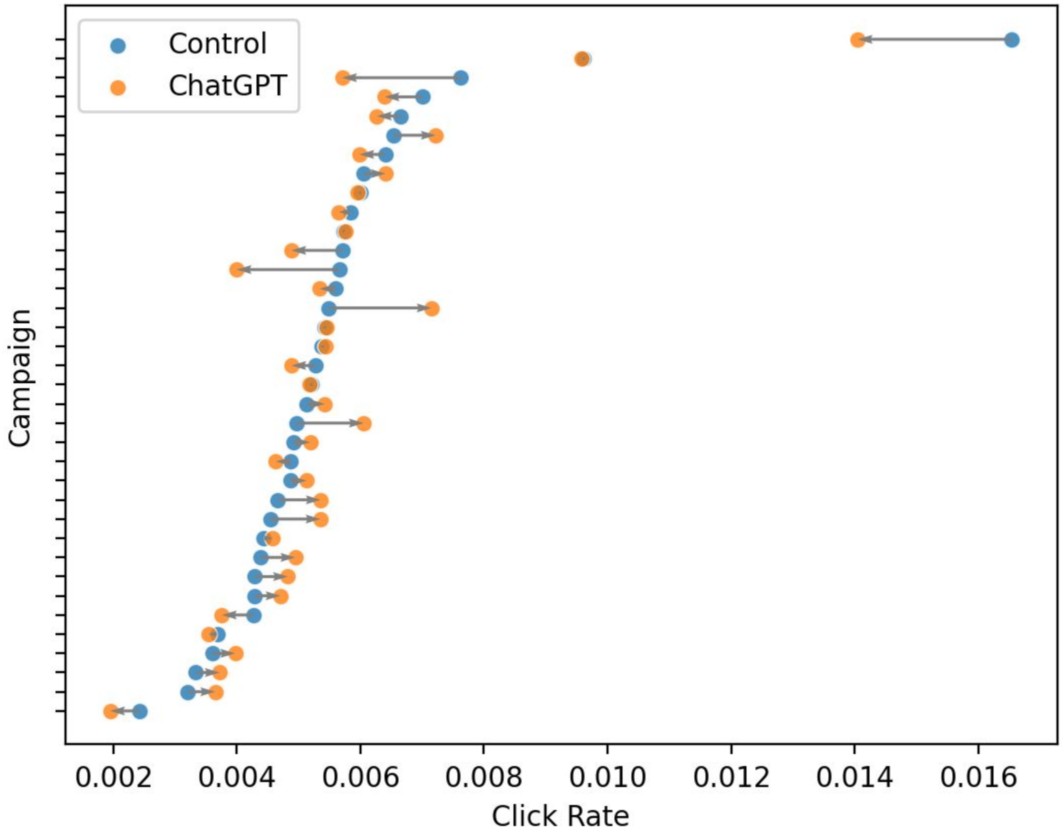
Performance of unassisted human across 36 campaigns



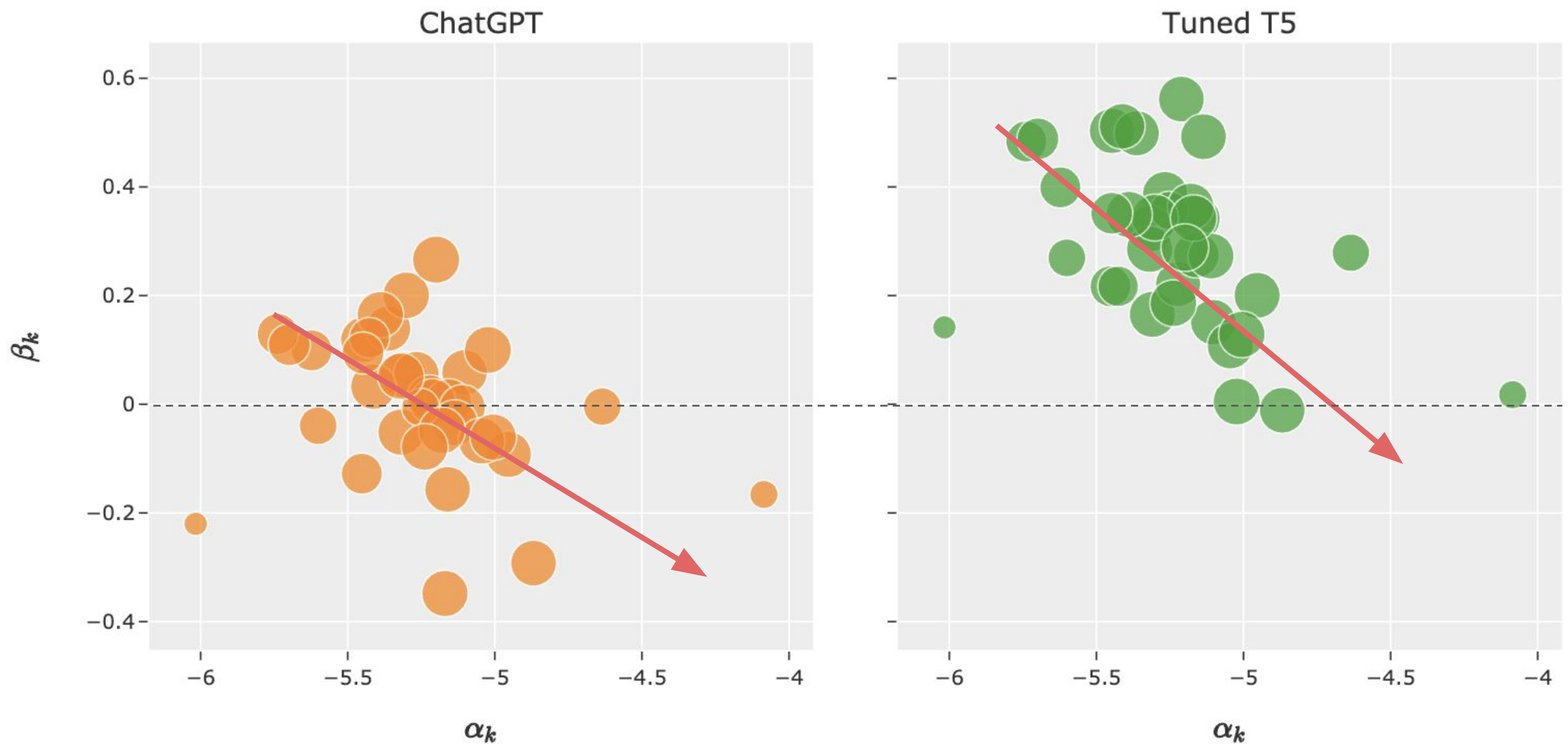
Assistance from our tuned model improves performance



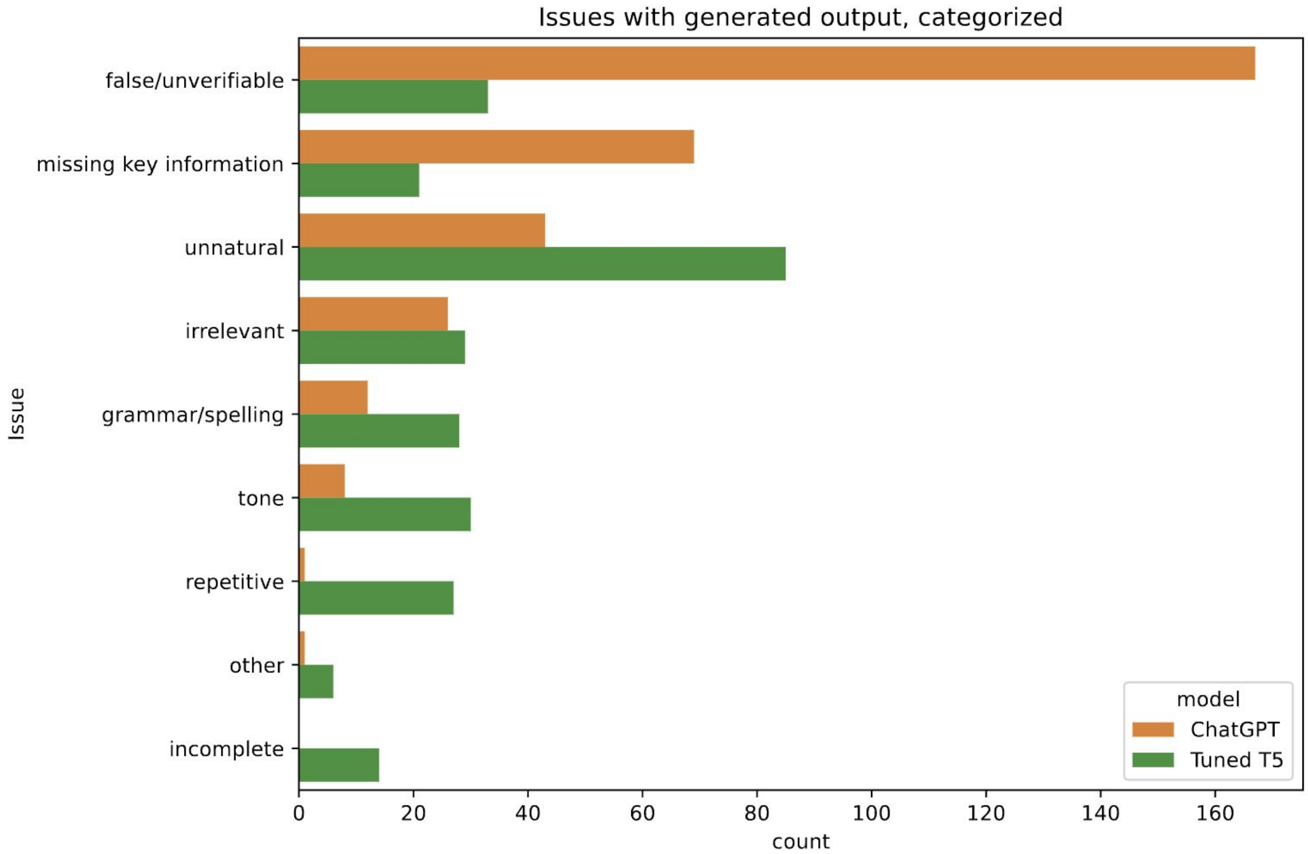
ChatGPT doesn't improve performance



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:

- Fine-tuning is necessary
- *Small* language model is sufficient (T5-base is 30x smaller than gpt-3.5-turbo)

To regulate behavior of AI:

- Design task to complement human
- Filter out undesirable output
- Impose mechanism ex ante and ex post

Conclusion

- Language models are useful for high-dimensional/unstructured decisions
- A/B tests are valuable beyond individual decisions; collectively are a strategic asset for improving future decisions
- Lots to be done! Possible to extend predictive models to prescriptive ones

Thank you!

kevin.lee@chicagobooth.edu