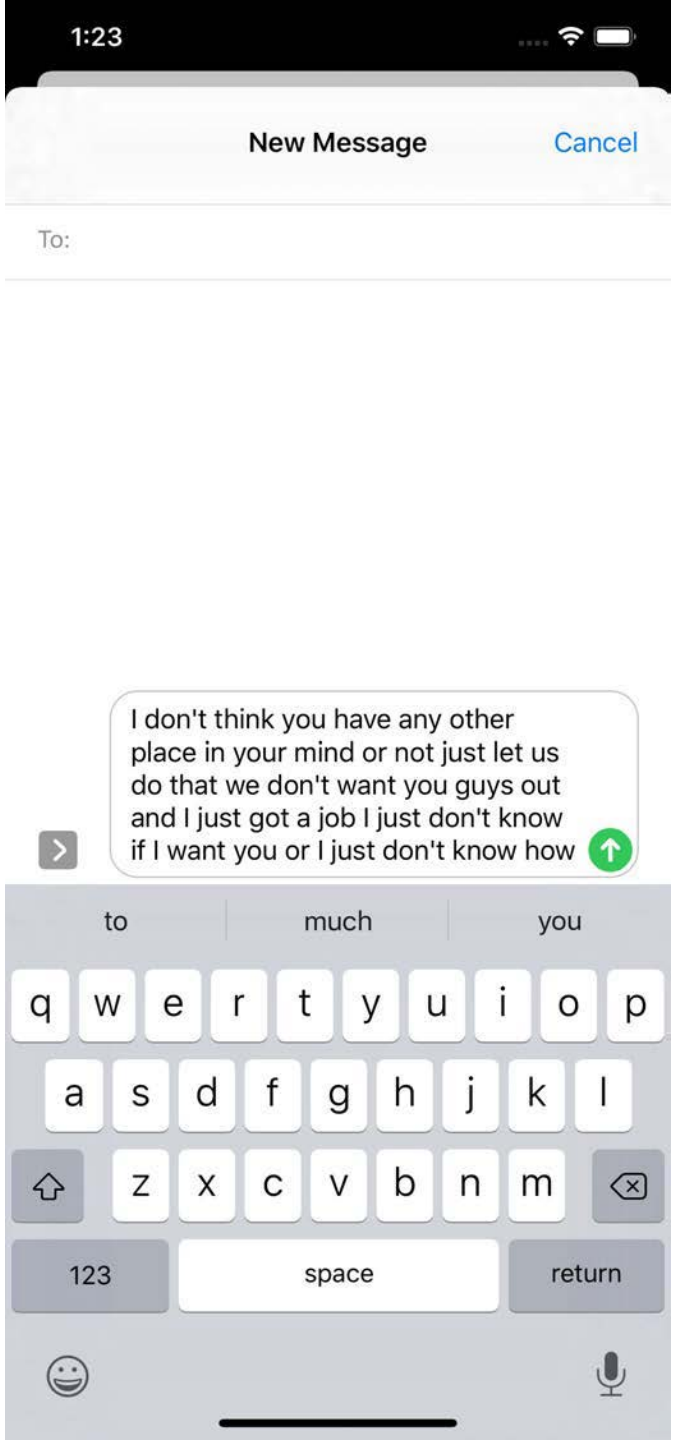
Predictive Text Biases Writers

Ken Arnold


1:23
New Message
Cancel
To:
I don't think you have any other place in your mind or not just let us do that we don't want you guys out and I just got a job I just don't know if I want you or I just don't know how
to
much
you
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
123
space
return

MOVIE QUOTES

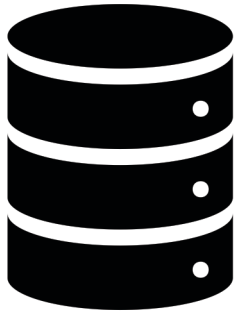
ACCORDING TO iOS 8 KEYBOARD PREDICTIONS

https://xkcd.com/1427/

https://www.blog.google/products/gmail/subject-write-emails-faster-smart-compose-gmail/

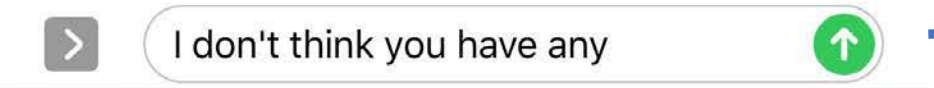
Training **Data**

Language Model



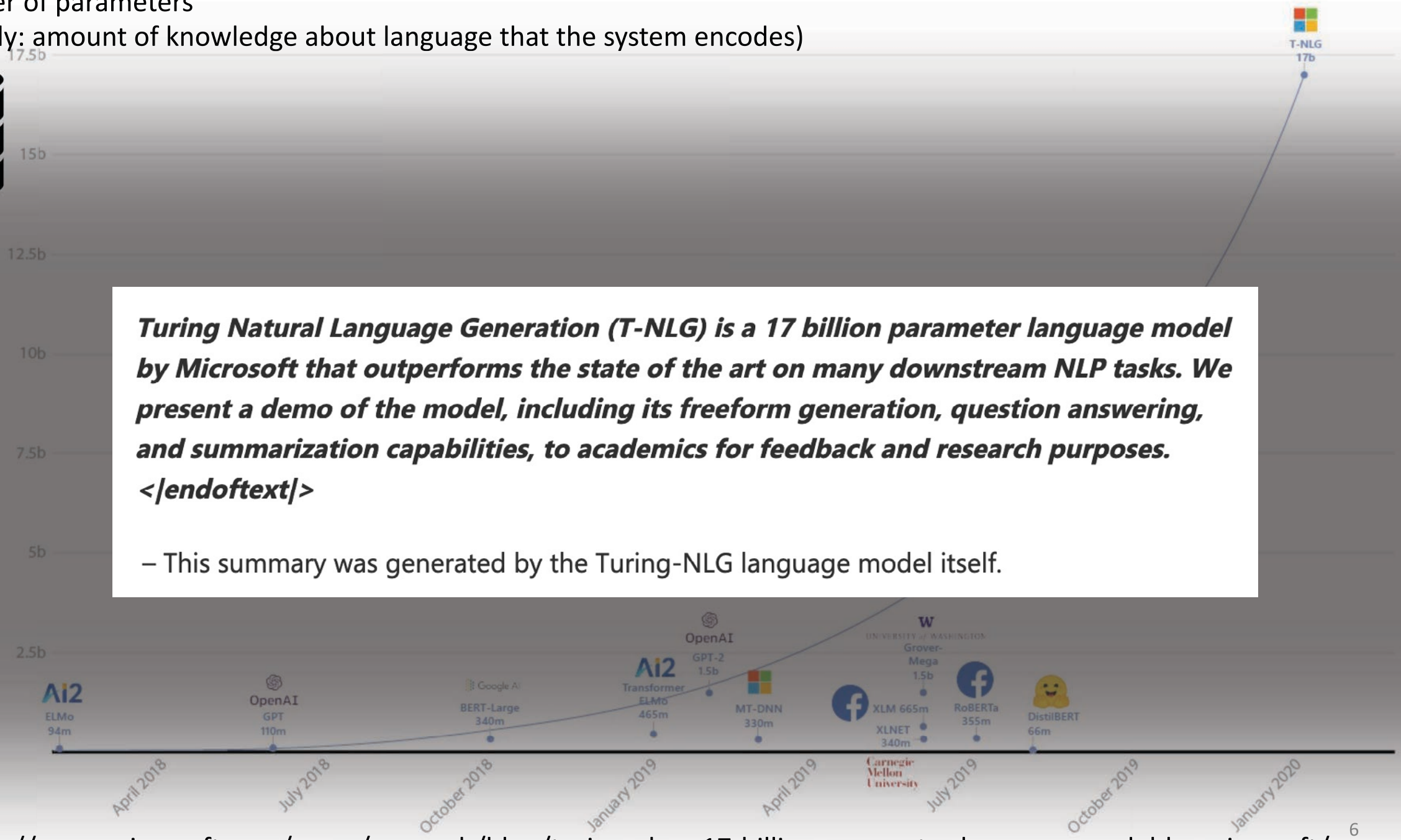
Predictions

Suggestions



Writing in progress

Number of parameters
(roughly: amount of knowledge about language that the system encodes)



Turing Natural Language Generation (T-NLG) is a 17 billion parameter language model by Microsoft that outperforms the state of the art on many downstream NLP tasks. We present a demo of the model, including its freeform generation, question answering, and summarization capabilities, to academics for feedback and research purposes. <|endoftext|>

– This summary was generated by the Turing-NLG language model itself.

https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Prior Evaluations Ignore Content

- Transcription tasks: “Type this”
- Speed and accuracy measures

But:

- Writing choices are made in-the-moment
- Suggestions are offered after *each character typed* — several per second

How do predictive suggestions shape content?

The **text that people write**
using predictive text entry systems
reflects **biases** of these systems.

“I have a phrase in mind and use suggestions that match.”

Bias *noun*. disproportionate weight in favor of or against an idea or thing

Biases

Inherent in interaction design

Emergent from data + algorithms

Intentional

Avoided?

Experiments

1. Fewer **unexpected** words
2. Stronger effect for **phrases**
3. **Sentiment** bias propagates
4. **Interaction data** enables intentional manipulation
5. **Predict questions** to guide without manipulating

Word Suggestions Discourage the Unexpected

Predictive Text Suggests what's Most Expected

This weekend I plan to	I want to go for a	My kids are so
take	**walk**	**excited**
go	**run**	**happy**
attend	**swim**	**much**
visit	ride	good
be	long	young
do	few	scared
make	stroll	smart
spend	little	small
travel	hike	different
play	quick	proud
write	drive	busy
start	drink	cute
meet	break	lucky
have	spin	used
return	shot	tired
bring	beer	close
get	bit	very
head	jog	beautiful
run	second	special

Bias against these words

Suggestions favor certain kinds of words at the expense of others: a **bias by design** against the unexpected

Does writing with suggestions result in more of those kinds of words?

Experiment 1: Does writing conform to suggested words?

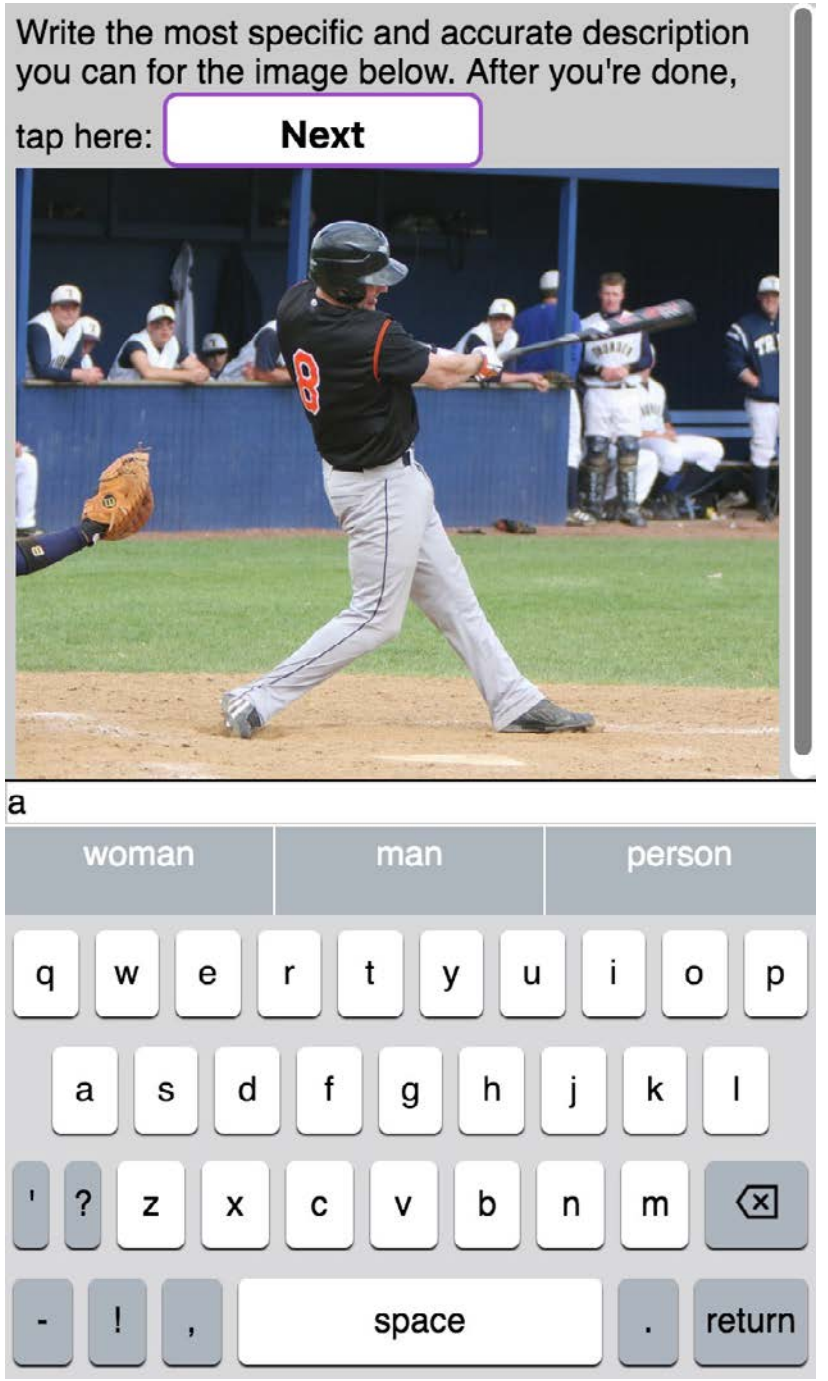
Vary suggestion **visibility**

- Suggestions **never** visible
- Suggestions **always** visible
- Suggestions visible when system is **confident**
  - Tuned so that full-word suggestions are shown for ~50% of words

Hypothesis: Longer, more predictable words when suggestions visible

Image Captioning Task

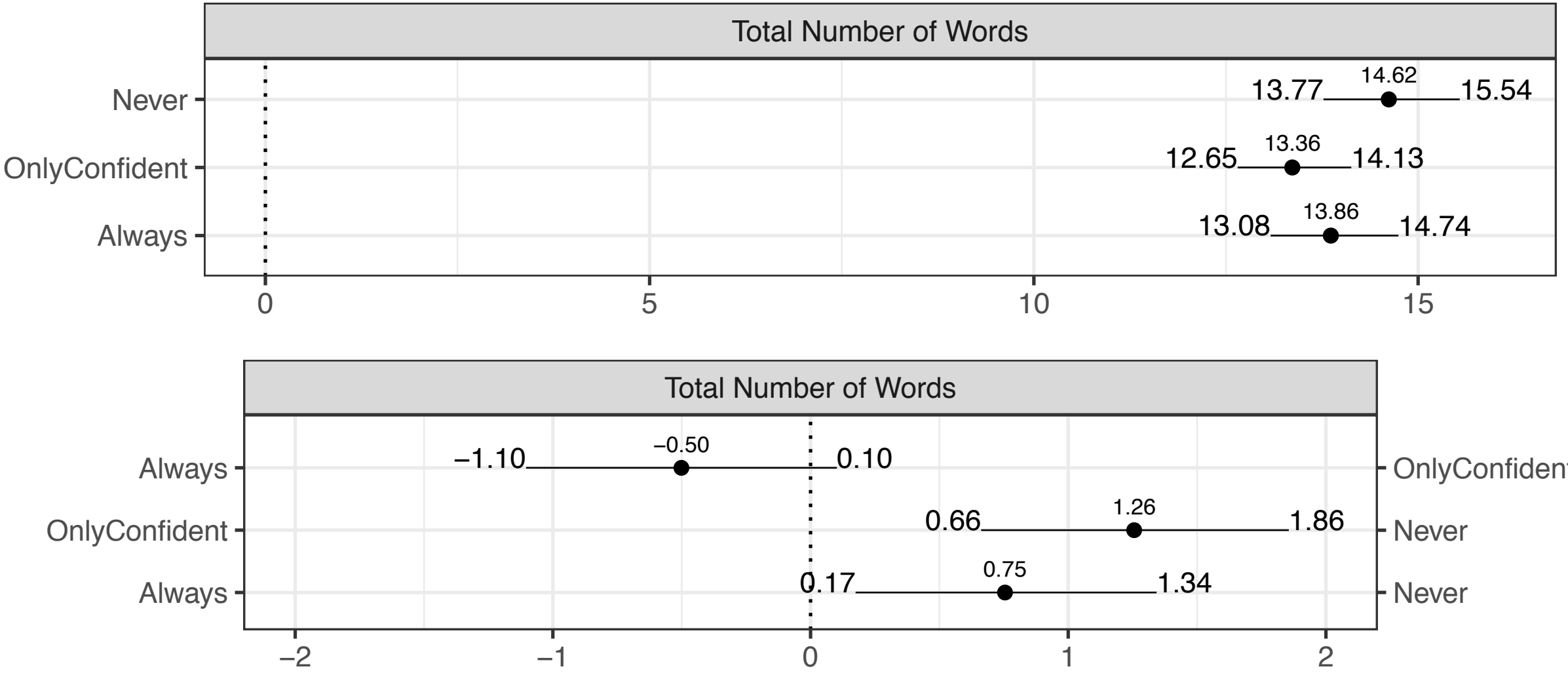
- **Open-ended:** suggestions can affect content (or not)
- **Repeatable**: one participant can write many times



Study Design

- 12 images, fixed order
- Counterbalanced suggestion visibility

Results: **Shorter** Captions

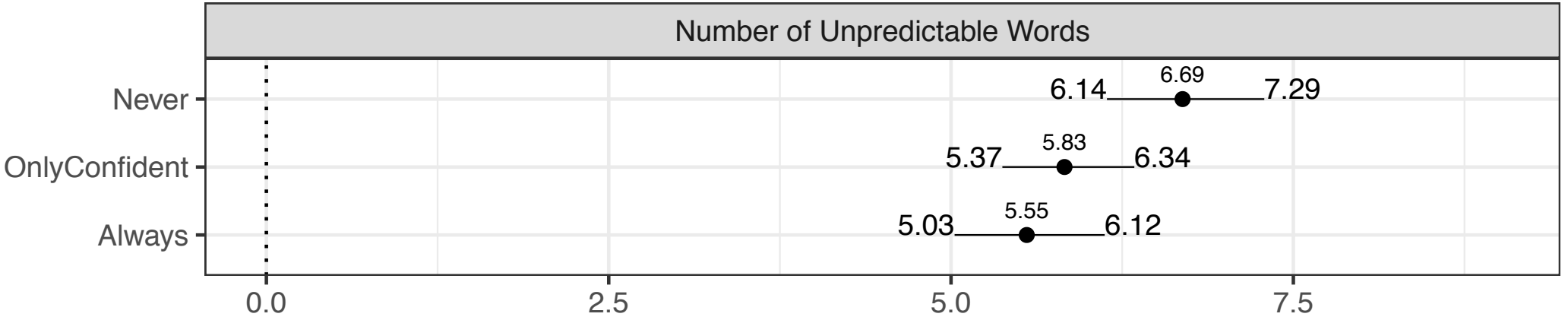


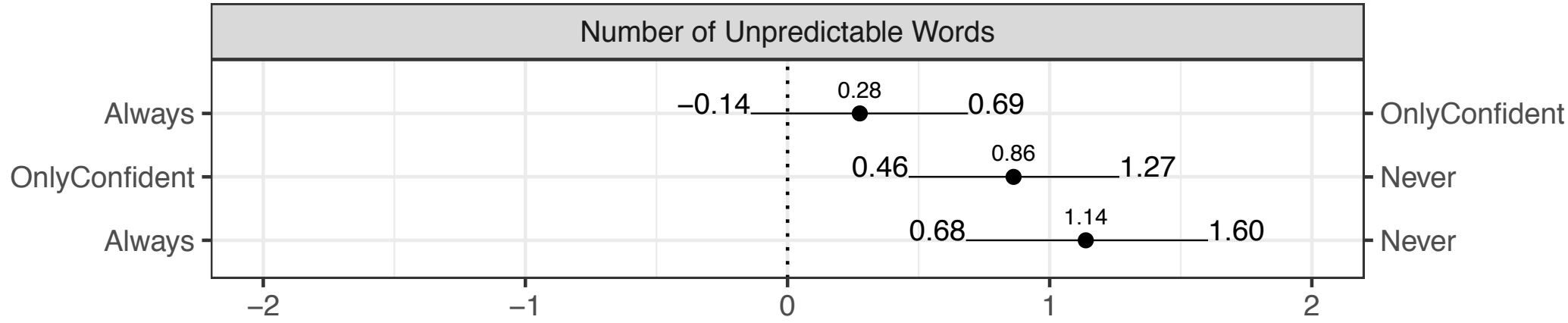
Predictability Measure

a	image	photo	is	parked	a	station	station	station
two	old	man	car	on	the	train	train	area
the	elephant	fashioned	with	traveling	an	cruise	river	train

an	old	train	is	approaching	a	quaint	outdoor	station

Results: **Fewer Unexpected** Words





Substitution Nudge

- Writer takes a suggested word *instead of* the counterfactual word
- Outcome: more predictable, fewer unpredictable

a	image	photo	is	parked	a	station	station	station
two	old	man	car	on	the	train	train	area
the	elephant	fashioned	with	traveling	an	cruise	river	train
an	old	train	is	approaching	a	quaint	outdoor	station

Skip-Nudge

- Suggests the word that would have been written *after* the next word

  ➔ nudge to *skip* a word

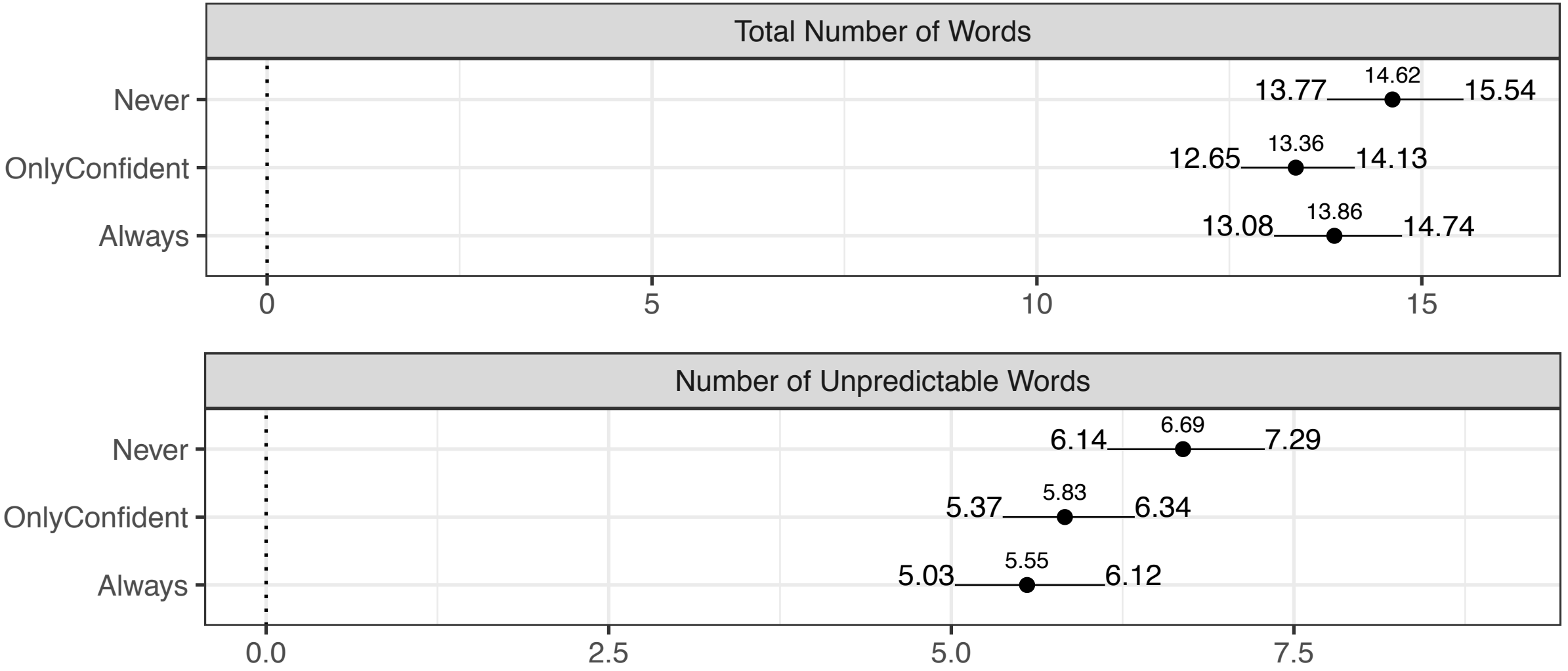
- Occurred at least once in 69% of captions

  Words most often skip-nudged:

  red, white, a, is, wedding, tennis, to, tree, sitting, small

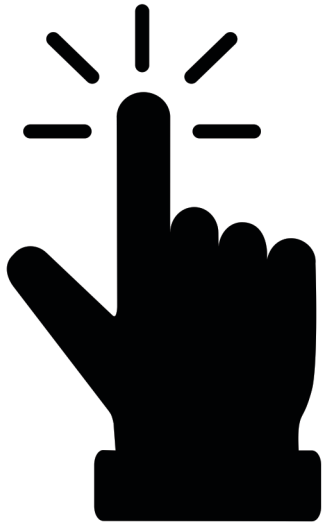
a	image	photo	is	parked	a	station	station	station
two	old	man	car	on	the	train	train	area
the	elephant	fashioned	with	traveling	an	cruise	river	train
an	old	train	is	approaching	a	quaint	outdoor	station

Overall: **Shorter** Captions, Fewer Unexpected



Suggestions may discourage thoughtfulness

Thinking about what to write



Accepting a suggestion

moderate by Erik Arndt from the Noun Project, easy by Tomi Triyana from the Noun Project

Discussion

- Suggestions affect word choice

- Effects of predictive text may be **subconscious:**
  - Individual writers can’t clearly perceive effects.
  - Aggregate analysis is necessary

Content Effects of Phrase Suggestions?

Designing Phrase Suggestions

Design goals

- Don’t interfere with word suggestions
- Should be able to accept as much or as little of phrase as desired
- Use minimal extra screen real estate

Phrase Preview Design

- Text below main suggestion shows **preview of upcoming words** in same slot
- Tap-tap-tap in same slot inserts words from that phrase

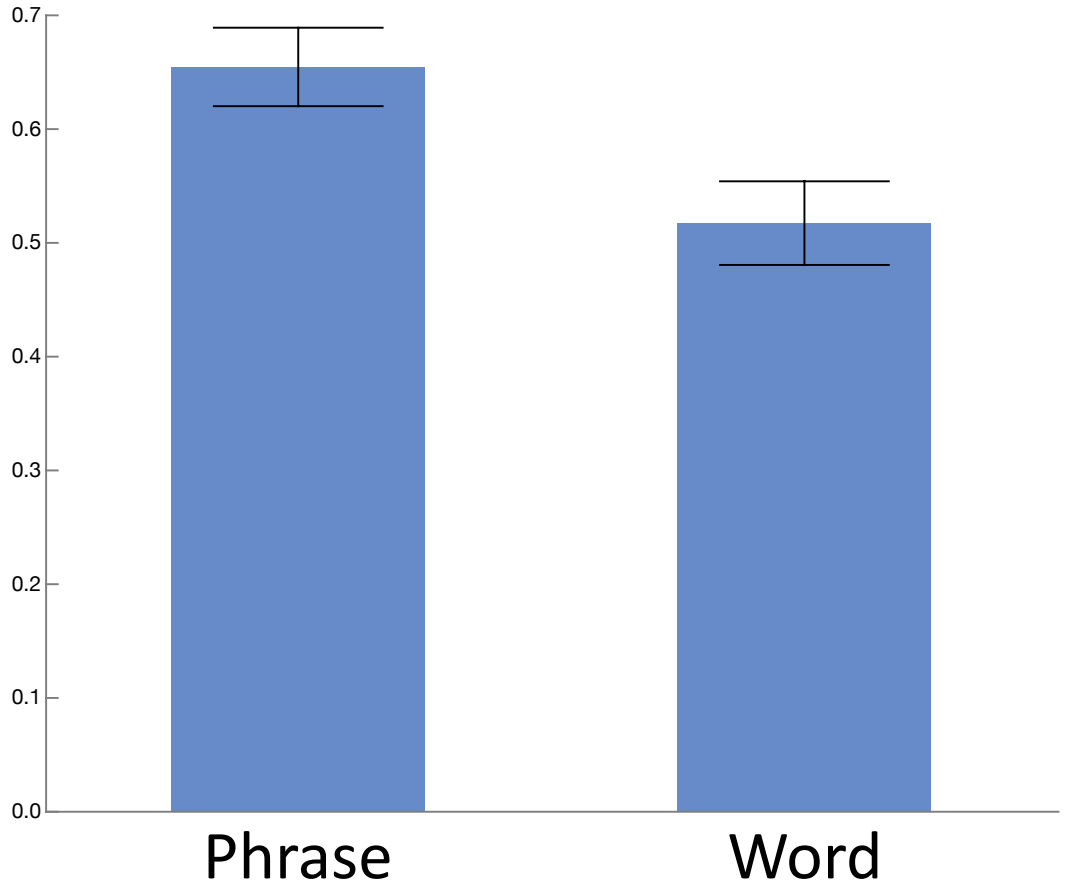
Experiment 2: Do phrase suggestions affect writers differently than words?

- 20 students each wrote reviews for 4 restaurants visited recently

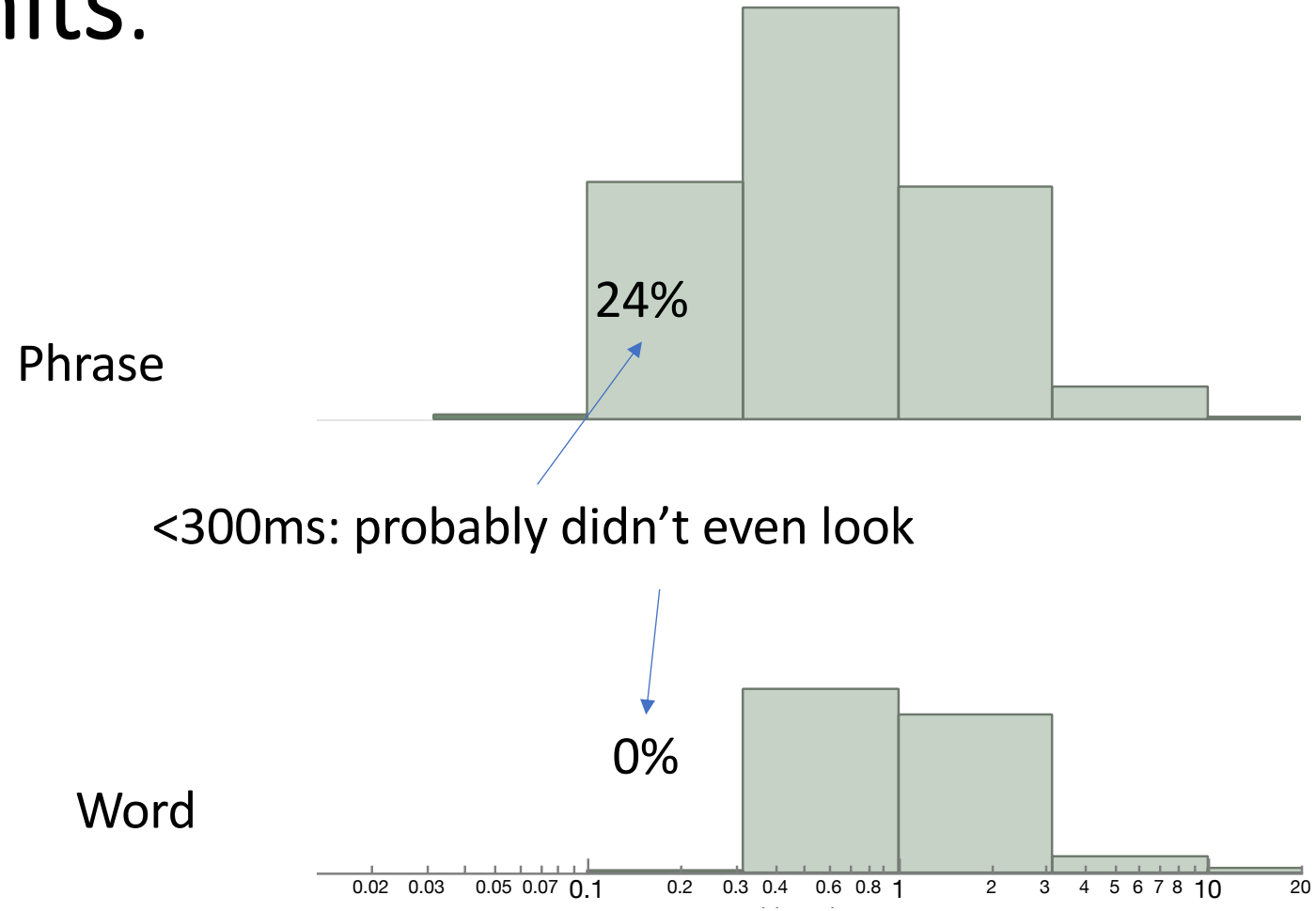
- Alternating conditions (within-participant):
  - **Phrase**: phrase previews visible
  - **Word**: phrase previews hidden (but identical behavior)

Phrase predictions affect content more strongly than word

proportion of words used that had been offered as suggestions



People used multi-tap gestures to insert phrases as units.

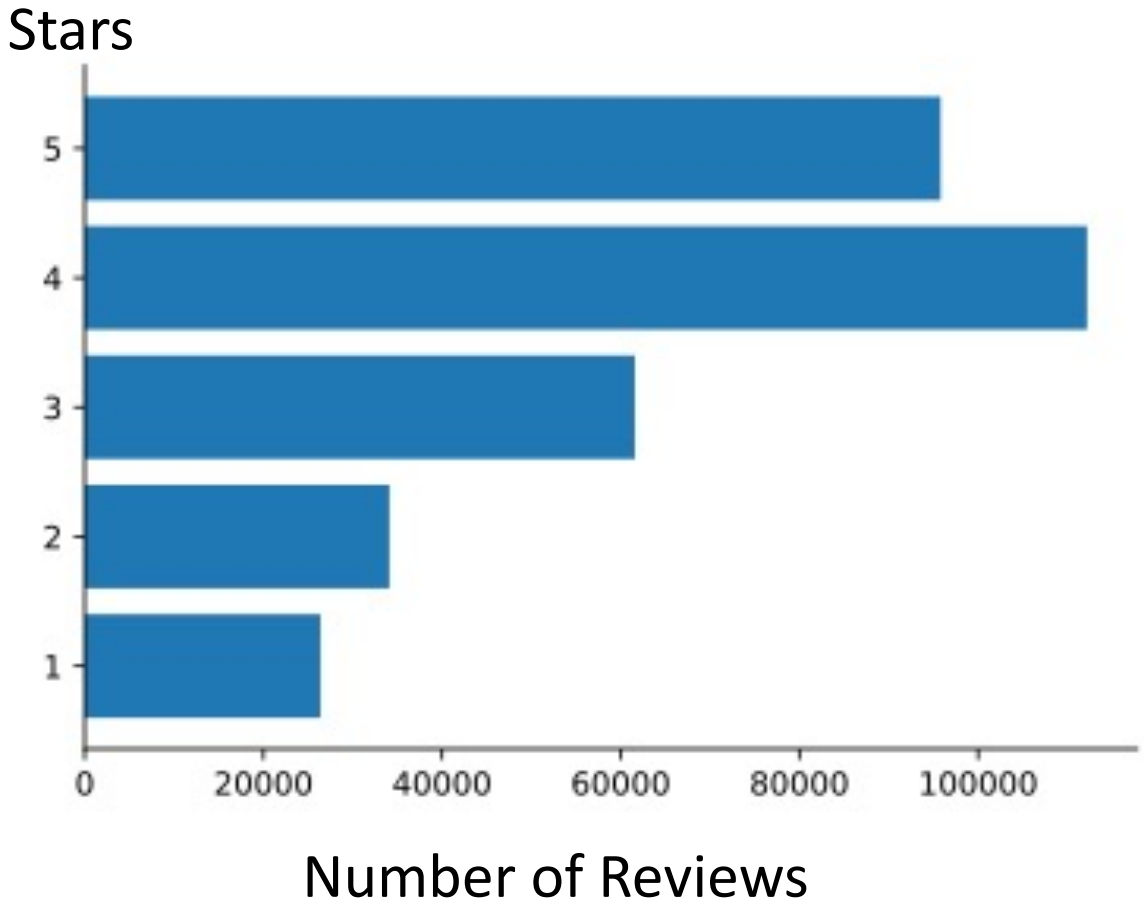


Do suggestions just affect wording or also meaning?

Maybe the writer had chosen what to say
but just accepted a different wording for convenience…

Sentiment Bias:
Data ➔ Suggestions ➔ Writing

Many review datasets are **biased positive**



Yelp Academic Dataset (3861 restaurants, 196k reviews)

Are predictions also biased?

- Use equal number of reviews of each star rating
- Train 5-gram language model (Kneser-Ney)

Example (from 2-star review)

The food at the OG truck is delicious. My GF and I tried both the turkey and the steak dishes and the flavors were excellent. Honestly, quite good.

prediction: It was my first time

original: The problem is that portions

Experiment 3a: How does dataset bias affect predictions?

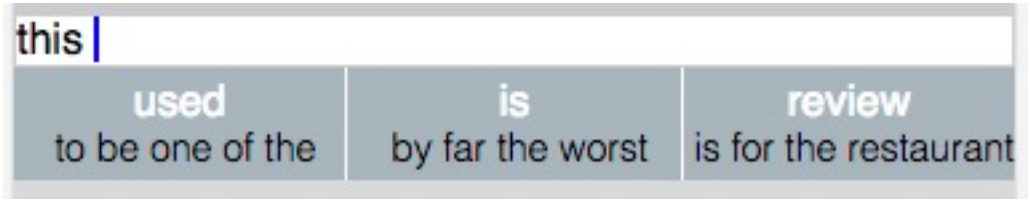
- Generate phrase predictions while “retyping” held-out reviews
- Annotators (MTurk) given pairs of phrases in context: “which is more positive”?
- Annotators rated predicted phrases more positive
  - Especially at sentence beginnings
  - No clear difference for single-word predictions

Biased datasets ➔ biased system predictions

Experiment 3b: Do positively-biased suggestions bias writing?

- N=38 participants, each named 4 restaurants (both good and bad experiences)
- Wrote reviews; system suggested sentiment-manipulated phrases
  - **Positive slant**

  - **Negative slant**

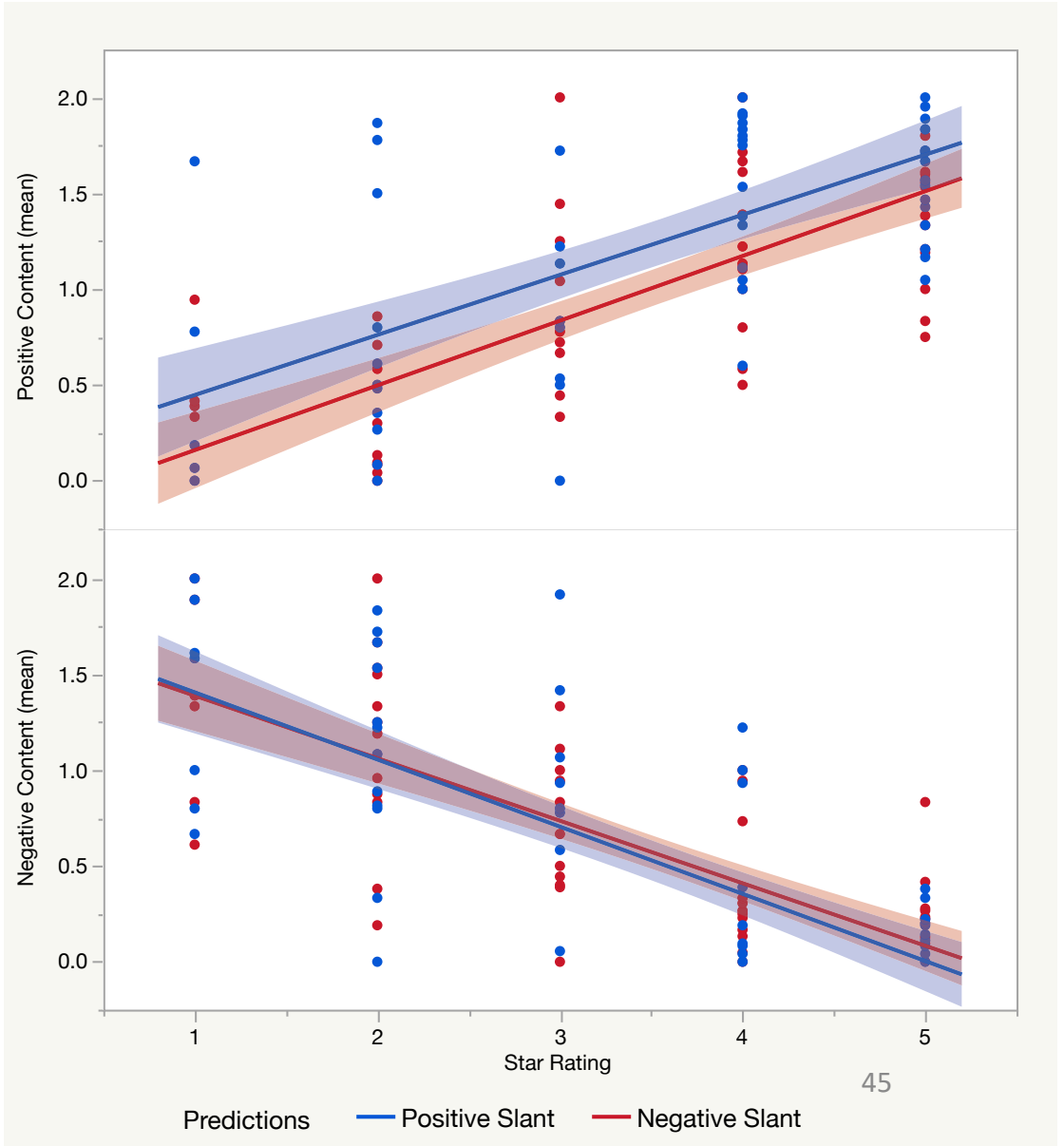


- **Measure**: For each sentence in each review, 3 annotators (MTurk) rate positive content, negative content

Results: Positive-Slanted Suggestions ➔ Positive-Slanted Reviews

- Effect size: about 1/3 star rating
- No measurable effect on negative content

- **Subjective preference towards positive slant system**
  - Negative slant viewed as less relevant



Excluding 3 participants who wrote without tapping keys at all

Discussion

- **Chain of bias**: data, suggestions, writing
- Bias affected **meaning** of what was typed, not just wording
- Writers preferred the system that amplified biases in the data

Intentional Biases

Platforms face pressure to influence online discourse

Various applications, variously ethical, including:

- Promote civil discourse and community norms
- Discourage bullying, hate speech
- Encourage writing that promotes engagement

Is intentional bias feasible?

- Sentiment manipulation was only considered *relevant* when it *amplified existing bias*.

- Fundamental trade-off:
  - Manipulated suggestions may seem less relevant.
  - Writers ignore or disable suggestions they perceive as irrelevant.

Data Collection

- System offers a *controlled variety* of suggestions
- Consistent target: a single review of a single restaurant (Chipotle)
- Collected *interaction data*

i hate spicy food but for some reason i love the flavor of chipotle chiles in any form, so i loooove chipotle. i have been nervous about eating here lately because of the food poisoning scandals but thankfully i have not had any problems! i alway order the burrito bowls and the portions are huge! service is just mediocre but not bad and you cant expect too much from a chain restaurant. overall i would give chipotle four stars.

i love this place. the food is good. it's a little expensive, but the food is so much more than that. and i love the people that work there. the burritos are huge and packed with flavor. i got one with chicken and beef and extra quac. it tasted fresh and i couldn't even finish it all as it was huge! i love the location and the atmosphere was great. i will definitely come back to try something different!

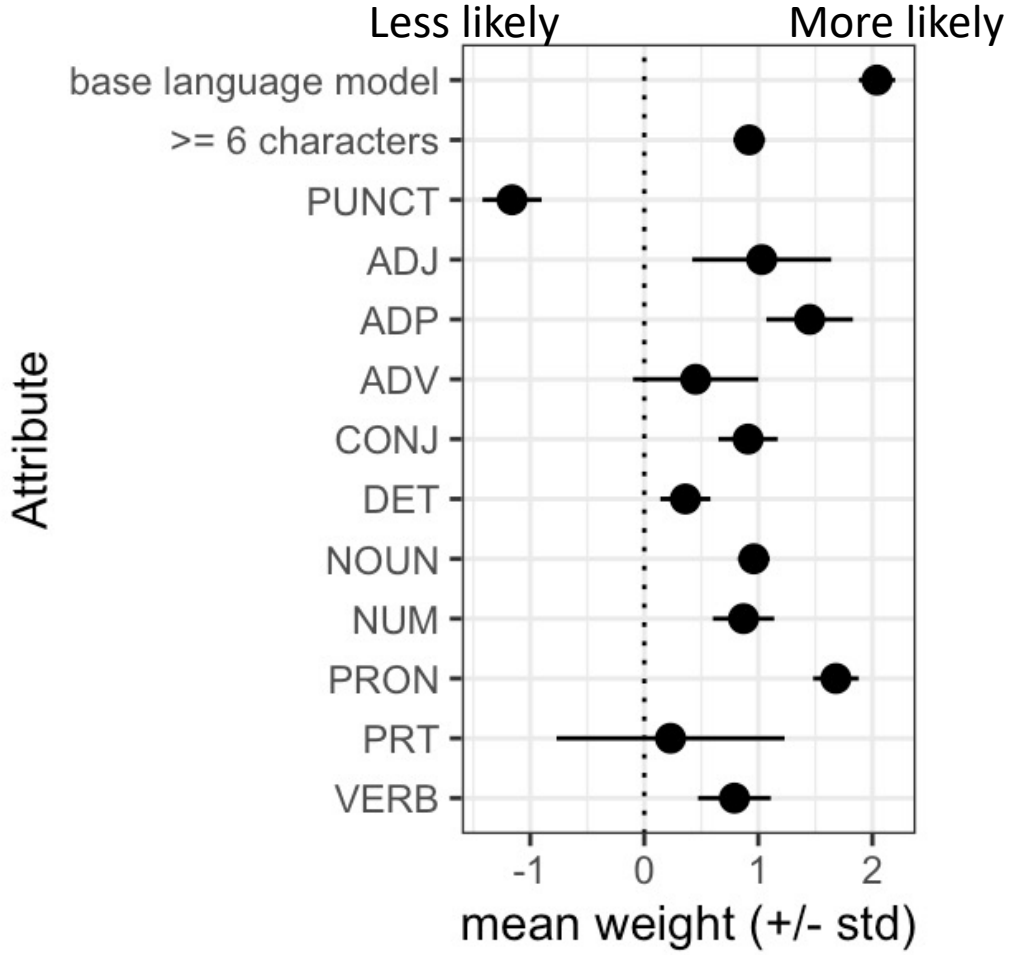
chipotle is one of the best things i've had in a long time. i was surprised at how much i enjoyed the food and the service. there's lots of options for people with dietary restrictions and variety of choices for the rest. the price is a bit much but is worth it. the food is of great quality since. they use a lot of local ingredients. i had the chicken burrito bowl and it was excellent. i would highly recommend this place to anyone.

Experiment 4 (**simulation**): Can we learn generalizable patterns of suggestion acceptance?

- Interaction log had a wide range of suggestions
  - Some suggestions were accepted, others not
- Suggestion system tends to offer certain kinds of suggestions
  - Alternative system: some suggestions more likely, others less likely
- Is there an alternative system that makes *accepted* suggestions more likely?

Results: Optimized Generation Policy

Context	sour cream yummy yummy yummy
Suggestion	yummy in my tummy .
Reward	4.00
Logprobs	orig: -7.13, new: -3.20



Implications

- *Interaction data* enables manipulation
- Interaction data is unequally distributed
  - **Abundant** for platforms and keyboard providers (Google, SwiftKey/Microsoft)
  - Besides this project, **no public data** on suggestion acceptance available

Transparency and accountability are needed!

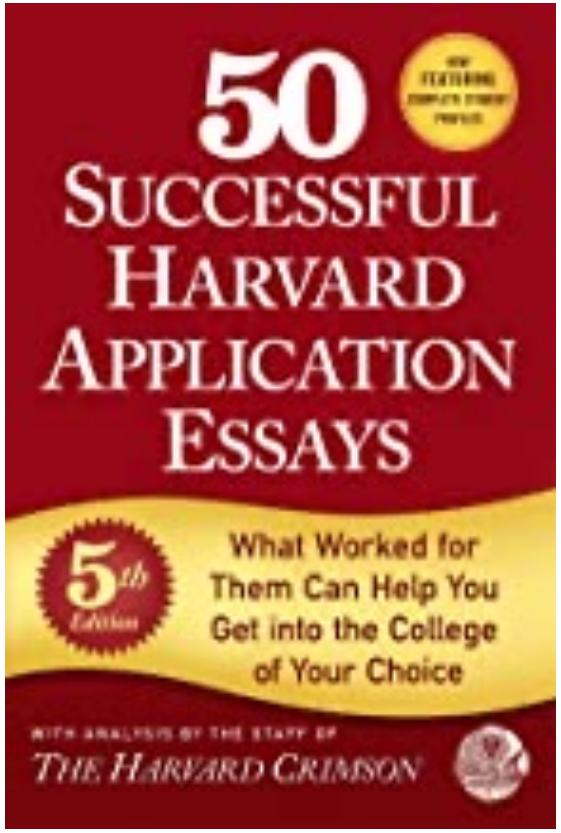
Opportunities for Technological Guidance of Content Creation

Next-word suggestion **risks bias.**

Need a different approach.

Applying to Harvard?

high-quality examples



expert guidance

How to Write the Perfect Harvard Essay: 3 Expert Tips

Tips for Answering This Prompt

- **Instead of just listing the titles of books you've read, you might want to include a short sentence or two commenting on your reaction to the book, your analysis of it, why you enjoyed or didn't enjoy it, etc., after each title.** Be sure to vary up your comments so that

PrepScholar

https://blog.prepscholar.com/harvard-essay-prompt-examples

Experiment 5: Design Study

- **Task**: sentence-level informational writing
- Varied prompt **presentation**:
  - Snippets (of Featured Articles)
  - Questions (corresponding to those snippets)
  - No prompts

Suppose you're writing a Wikipedia article about a sci-fi drama.

Snippet

"Blade Runner" initially underperformed in North American theaters and polarized critics; some praised its thematic complexity and visuals, while others were displeased with its slow pacing and lack of action.

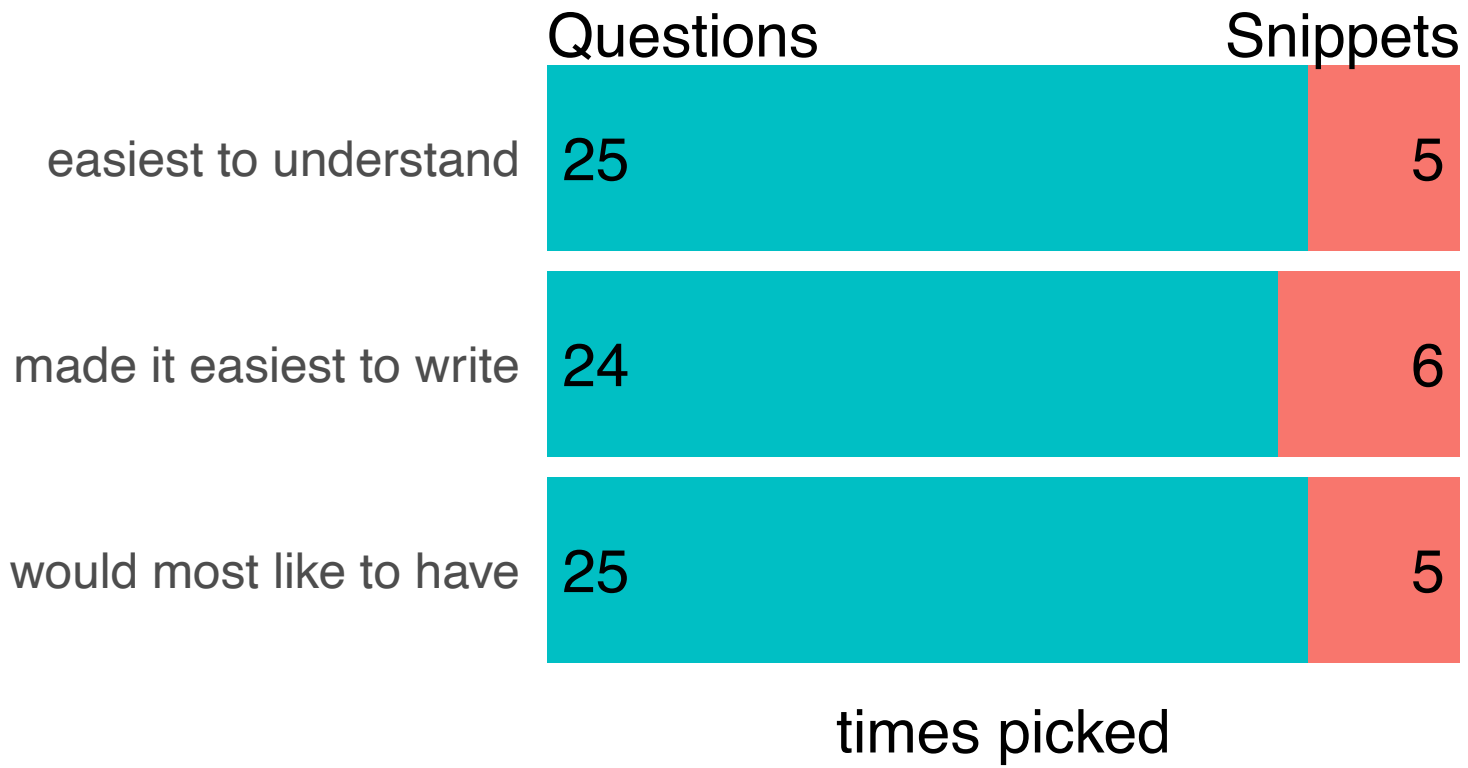
Questions

How did it initially perform? How did critics react? What aspects did critics praise? What aspects did critics condemn?

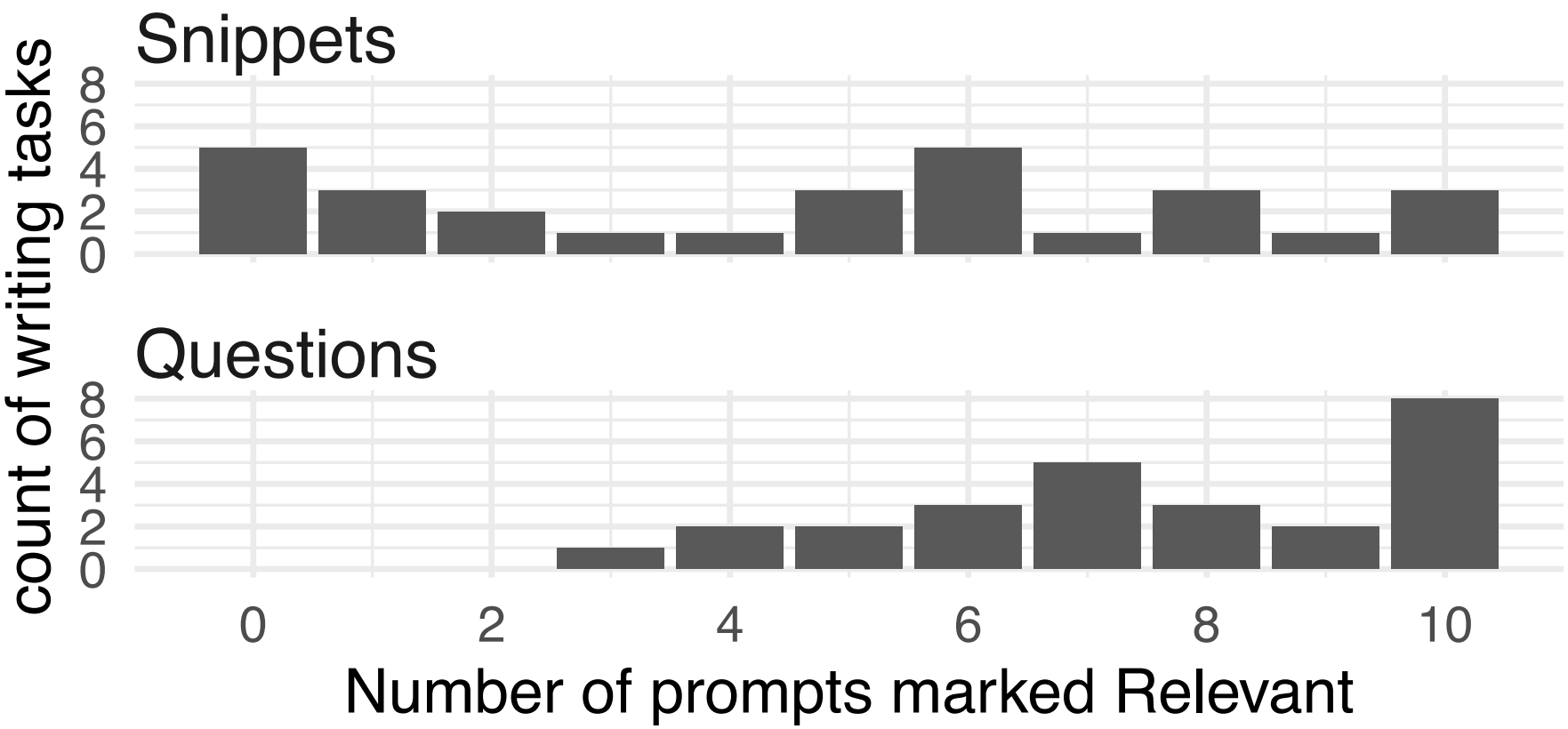
Procedure

- 30 MTurk participants
- Each named a film, a book, and a travel destination they knew well
- For each item, asked to write 10 sentences
  - One prompt for each sentence
  - Same prompts for everyone (not adaptive; only testing **presentation**)
- Marked whether prompt was relevant; if so, wrote a sentence
  - “write sentences that would belong in an encyclopedia article on ___”

Writers strongly preferred Questions



Questions more often perceived as useful



Is generating questions feasible?

- Human computation would require domain knowledge
- Existing question generation methods require answer texts
- Language models could generate plausible sentences, but
  - Might not generate uncommon topics
  - Still need to translate into questions

Hybrid Approach

- Algorithm finds clusters of sentences in training set
- Humans write questions for each cluster
- Algorithm predicts which clusters are relevant

Questions: Did the film premiere at a film festival? How did it perform?

- The film premiered at the 2012 Sundance Film Festival and was screened out of competition at the 62nd Berlin International Film Festival in February 2012.
- The film premiered on January 21, 2013 at the 2013 Sundance Film Festival and was screened in competition at the 63rd Berlin International Film Festival.
- The film had its world premiere at the Sundance Film Festival on January 22, 2017, and later screened at the Berlin International Film Festival.
- The film premiered at the International Documentary Film Festival Amsterdam in November 2010 and has screened at several international festivals.

Questions: Who wrote the screenplay? Who directed the film? Who starred in the film?

- ...is a 1968 American comedy film directed by Ron Winston and written by Charles Williams and starring Robert Wagner and Mary Tyler Moore.
- ...is a 1987 American action comedy film directed by Tony Scott, written by Larry Ferguson and Warren Skaaren and starring Eddie Murphy.
- The movie co-starred Hollywood actors Jim Davis, Allison Hayes and John Hart along with Canadian actors Austin Willis and Tony Brown.

Discussion

- Alternatives to next-word suggestion are **desirable** and **feasible**

- Question generation can be
  - more **forgiving** of error than next-word prediction
  - **natural,** like getting interviewed

- Question-**asking** is a ripe opportunity for intelligent systems
  - Most work focuses on question **answering**.
  - …or question generation for assessing reading comprehension.

Could writing feel more like conversation?

“Interesting, tell me more!”

“What were they doing before that happened?”

“Have you thought about this perspective?”

The text that people write using predictive text entry systems reflects biases of these systems.

Inherent in interaction design

Emergent from data + algorithms

Intentional

Avoided?

1. Fewer **unexpected** words
2. Stronger effect for **phrases**
3. **Sentiment** bias propagates
4. **Interaction data** enables intentional manipulation
5. **Predict questions** to guide without manipulating

Key take-away

Researchers and platforms should measure and document how intelligent technologies may manipulate what we create.

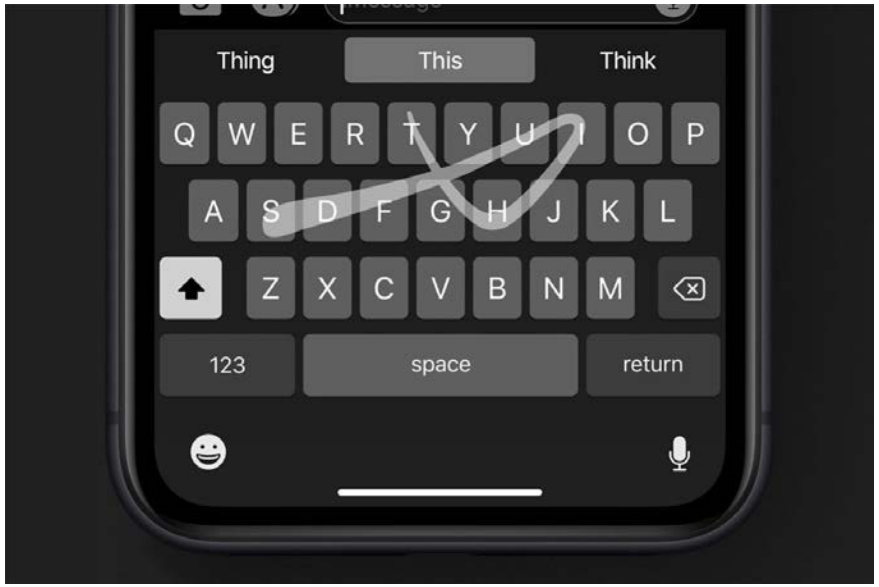
System capabilities are increasing rapidly.

My results suggest that **content effects will become more prevalent in the wild**.

Some Perspectives

Text Entry

- Don’t suggest a word until first letter typed
- Use gesture typing
  - Writer, not system, takes initiative
  - Still might influence content (e.g., writer avoids words that get misrecognized)



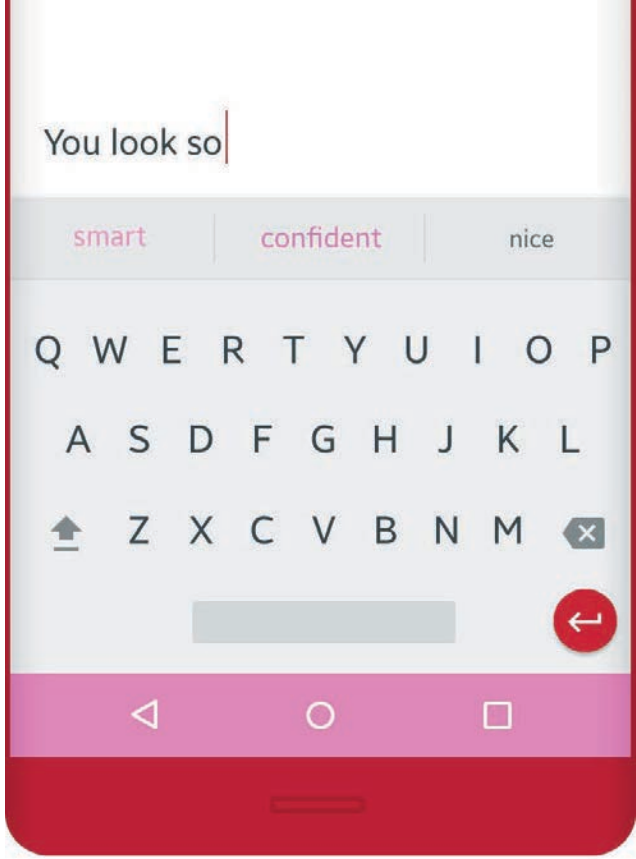
Intelligent Systems

- Need to test on authentic tasks
- Need diverse measures: *content*, not just *speed*

Writing

- Opportunities for L2 writers, writing education
- Opportunities for intentional bias
  - Encourage
  - Discourage (inject errors)

- Online communities could embed their values into predictive text systems for their contributors
  - Make *constructive* thoughts easier to type than *hateful* thoughts
  - *Scaffold* newcomers towards valuable participation



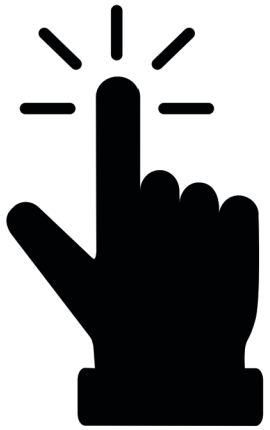
Transparency and Accountability

- **Platforms, share your data!**
  - suggestion usage data (personal and aggregate)
  - … content measures (word predictability, others?)
  - … estimates of content effects (e.g., from A/B tests)

Everyday Creativity

- Convergent and divergent thinking
- Authorship?

Thinking about what to write



Accepting a suggestion

Ethical Questions

- Privacy
- Autonomy, Authorship
- Risk
- Diversity
- Accessibility
- Platform Responsibility

xkcd 2169

Future of Reading and Writing

- Written word is technology for thinking
- Time-tested but being reinvented by abundance of:
  - Data
  - Computation
  - Communication

Imagine what thinking together could look like.

Lessons I’ve Learned

Simplify

“I need to make sure this is complicated enough to be interesting”

vs

“How can I simplify this?”

Example: bias-by-design experiment

Celebrate

“that didn’t work; I need to do something completely different”

vs

“This isn’t what I was expecting, but it’s still valuable…how?”

Example: so many projects I started and abandoned…

Ask for Help
(and “how can I help?”)

“This isn’t good enough to get feedback on yet”

vs

“I could really use some help figuring out what I’m trying to do here”