Large language models & human language processing: from variability to prediction

Raquel Fernández

ILFG seminar - 17 Jan 2024

erc European Research Council

UNIVERSITY OF AMSTERDAM

Institute for Logic, Language & Computation

Overview

Variability is an intrinsic property of human language production

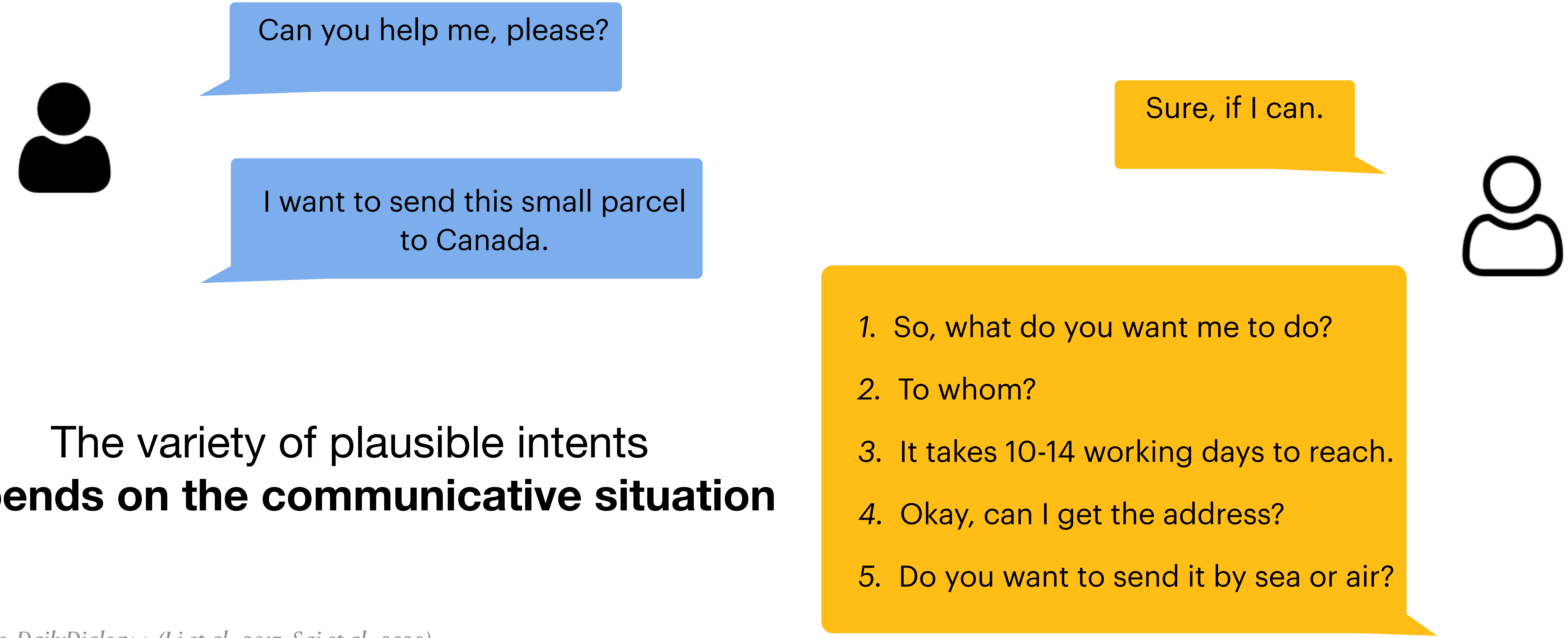
- Part 1: a framework to evaluate neural text generators in terms of their ability to reproduced production variability (i.e., uncertainty) observed in humans.

- Part 2: a proposal to exploit **production variability** to quantify **utterance predictability** in comprehension

  - **information value** quantifies the predictability of an utterance relative to a set of plausible alternatives, generated by neural text generators

In any given context,
speakers may have **variable intents**

(*what* to say)



Can you help me, please?

Sure, if I can.

I want to send this small parcel to Canada.

1. So, what do you want me to do?
2. To whom?
3. It takes 10-14 working days to reach.
4. Okay, can I get the address?
5. Do you want to send it by sea or air?

The variety of plausible intents
depends on the communicative situation

Dialogue: DailyDialog++ (Li et al., 2017; Sai et al., 2020)

Even when context and intent are fixed, speakers’ **linguistic realisations** may vary

(*how* to say it)

Several companies have thus far reacted cautiously when it comes to hiring.

→

Viele Firmen haben bisher vorsichtig reagiert, wenn es um Neuanstellungen geht.

Einige Unternehmen haben bisher bei der Einstellung vorsichtig reagiert.

Mehrere Unternehmen haben bisher zurücckhaltend reagiert, wenn es um die Einstellung von Mitarbeitern geht.

In Bezug auf Neuanstellungen haben diverse Unternehmen bisher mit Vorsicht reagiert.

Einige Unternehmen haben darauf mit Vorsichtsmaßnahmen reagiert wenn es um Neueinstellungen geht.

Source and translations: newstest 2014 (Bojar et al., 2014; Ott et al., 2018)

Even when context and intent are fixed, speakers’ **linguistic realisations** may vary

(*how* to say it)

Several companies have thus far reacted cautiously when it comes to hiring.

→

Viele Firmen haben bisher vorsichtig reagiert, wenn es um Neuanstellungen geht.

Einige Unternehmen haben bisher bei der Einstellung vorsichtig reagiert.

Mehrere Unternehmen haben bisher zurücckhaltend reagiert, wenn es um die Einstellung von Mitarbeitern geht.

In Bezug auf Neuanstellungen haben diverse Unternehmen bisher mit Vorsicht reagiert.

Einige Unternehmen haben darauf mit Vorsichtsmaßnahmen reagiert wenn es um Neueinstellungen geht.

Source and translations: newstest 2014 (Bojar et al., 2014; Ott et al., 2018)

Even when context and intent are fixed, speakers’ **linguistic realisations** may vary

(*how* to say it)

Several companies have thus far reacted cautiously when it comes to hiring.

Viele Firmen haben bisher vorsichtig reagiert, wenn es um Neuanstellungen geht.

Einige Unternehmen haben bisher bei der Einstellung vorsichtig reagiert.

Mehrere Unternehmen haben bisher zurücckhaltend reagiert, wenn es um die Einstellung von Mitarbeitern geht.

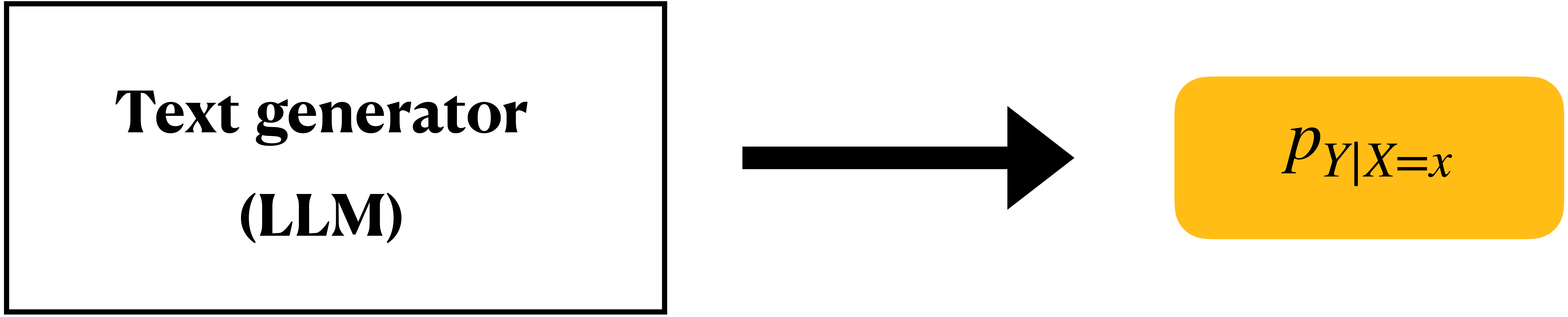
In Bezug auf Neuanstellungen haben diverse Unternehmen bisher mit Vorsicht reagiert.

Einige Unternehmen haben darauf mit Vorsichtsmaßnahmen reagiert wenn es um Neueinstellungen geht.

Source and translations: newstest 2014 (Bojar et al., 2014; Ott et al., 2018)

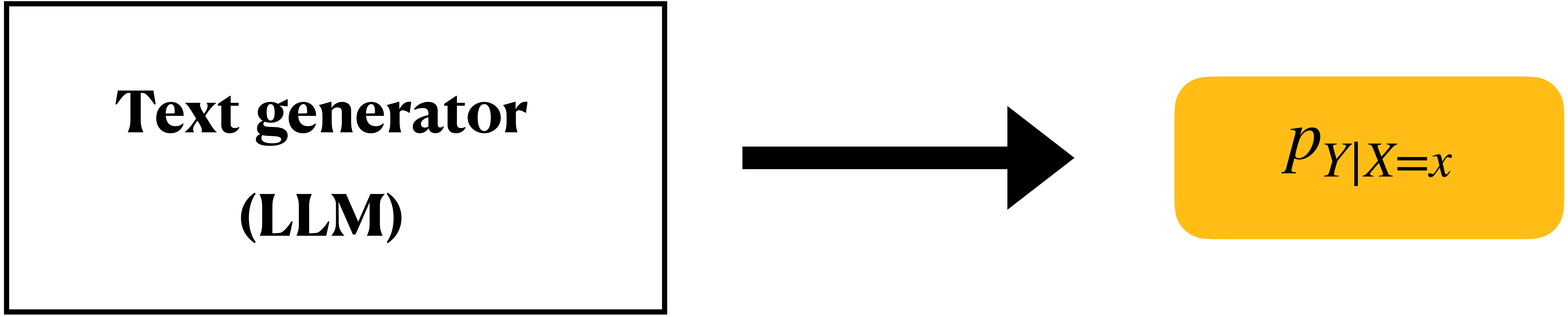
A framework for probing the uncertainty of NLG models

Probability distribution over sequences of tokens can be regarded as a **representation of the generator's uncertainty** (Halpern, 2017) about productions for a given generation context.



A framework for probing the uncertainty of NLG models

Probability distribution over sequences of tokens can be regarded as a **representation of the generator's uncertainty** (Halpern, 2017) about productions for a given generation context.

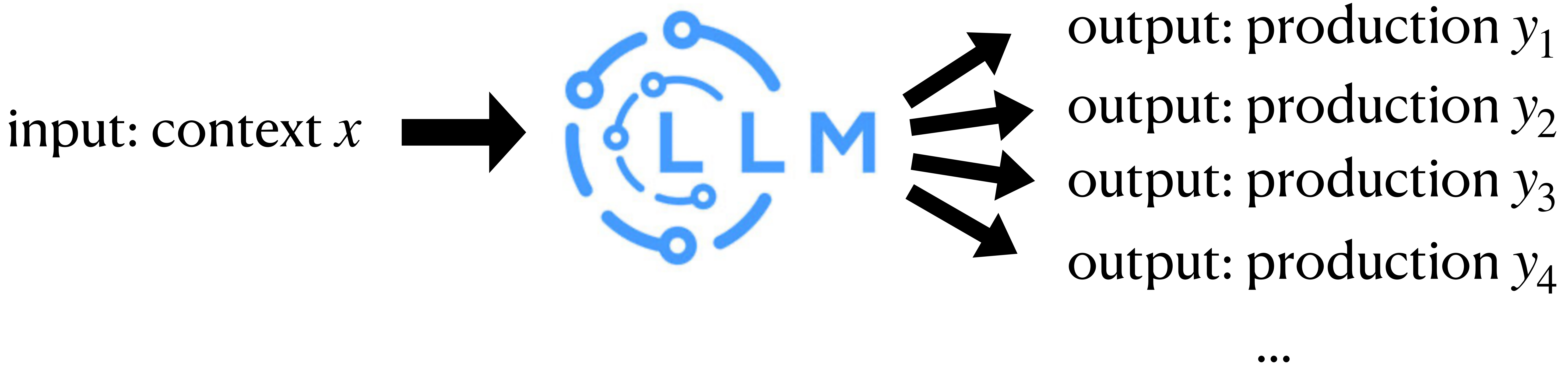


Text generator (LLM)

$p_{Y|X=x}$

Is this representation of uncertainty in compliance with production variability exhibited by a population of humans?

A framework for probing the uncertainty of NLG models



input: context x

output: production y_1

output: production y_2

output: production y_3

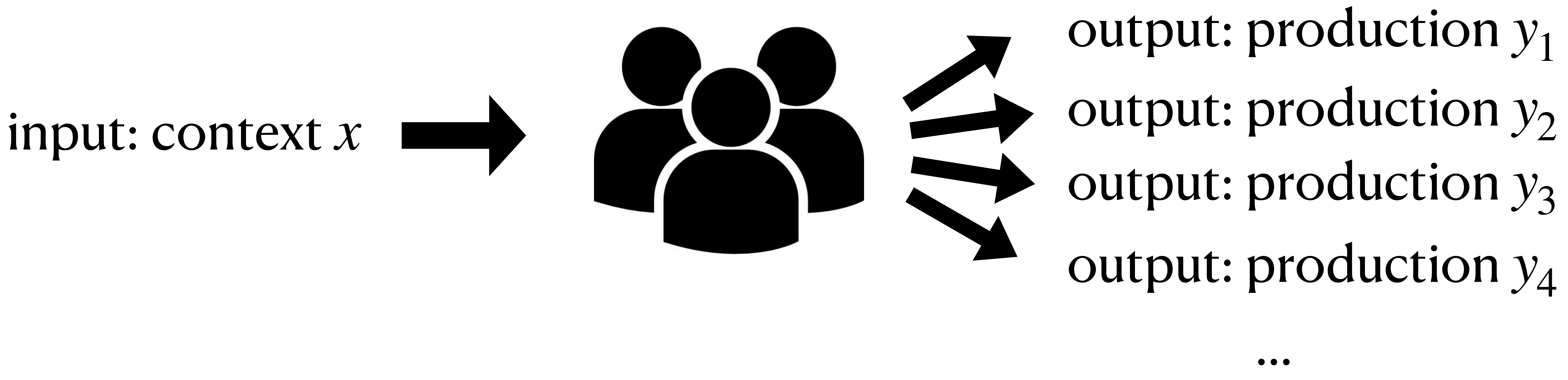
output: production y_4

...

We can quantify variability by measuring pairwise distance for a set of productions, given a distance metric $k(Y, Y) \in \mathbb{R}$. For instance:

- Semantic variability (*what* is said): cosine distance
- Lexical variability (*how* it is said): ratio of common words

A framework for probing the uncertainty of NLG models



input: context x

output: production y_1

output: production y_2

output: production y_3

output: production y_4

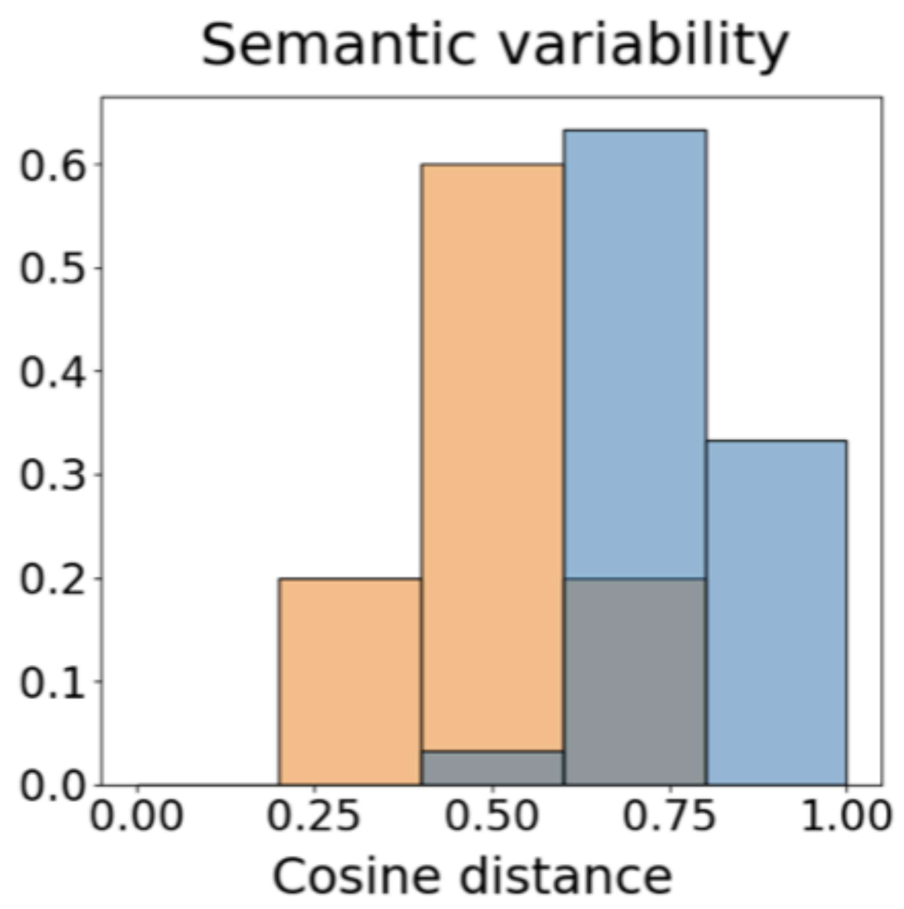
...

We can quantify variability by measuring pairwise distance for a set of productions, given a distance metric $k(Y, Y) \in \mathbb{R}$. For instance:

- Semantic variability (*what* is said): cosine distance
- Lexical variability (*how* it is said): ratio of common words

Dialogue context

It's very dark in here. Will you turn on the light?
Okay. But our baby has fallen asleep.
Then, turn on the lamp, please.
But where's the switch?



Humans

- Don't you know where the switch is?
- Switch is on the left side of the lamp.
- Just press the second switch on the board.
- Lamp is upon the study table and now you know where the switch is.
- I will light up the torch, so you can find the switch and turn on the lamp.

DialoGPT-medium, nucleus $p = 0.9$

- You don't have one.
- Where's the button?
- It’s on the top.
- Well, you'll want to turn it on.
- Turn it on.
- I'm sorry.
- On my chest
- I'm on it!
- Turning on the switch
- I have a few, try and figure it out.

$$\text{Human}_k(x) := k(Y, Y) \longleftrightarrow \text{Model}_k(x) := k(\hat{Y}, \hat{Y})$$

Quantify deviation from plausible human variability via a statistical divergence

Empirical Results

Production probes

Lexical similarity

Fraction of common n-grams in two productions, with $n \in [1,2,3]$ *

Syntactic similarity

Fraction of common POS n-grams in two productions, with $n \in [1,2,3]$ *

Semantic similarity

Cosine similarity between sentence embeddings of two productions **

* Number of matching n-gram occurrences divided by the total number of n-grams in both strings.

** (S-BERT; Reimers and Gurevych, 2019)

Experimental setup: data and models

Translation

Data: 500 sentences from *WMT-14 En-De* (Bojar et al., 2014) with 10 reference translations (Ott et al., 2018)

Models: Helsinki-NLP's Transformer-Align model trained on Opus-MT (Tiedemann & Thottingal, 2020)

Storytelling

Data: 759 story prompts from *WritingPrompts* (Fan et al., 2018) with at least 5 reference stories available

Models: GPT-2 large pre-trained and **finetuned** on *WritingPrompts*

Text simplification

Data: 2000 sentences from *ASSET* w/ 10 simplifications (Xu et al., 2016; Alva-Manchego et al., 2020)

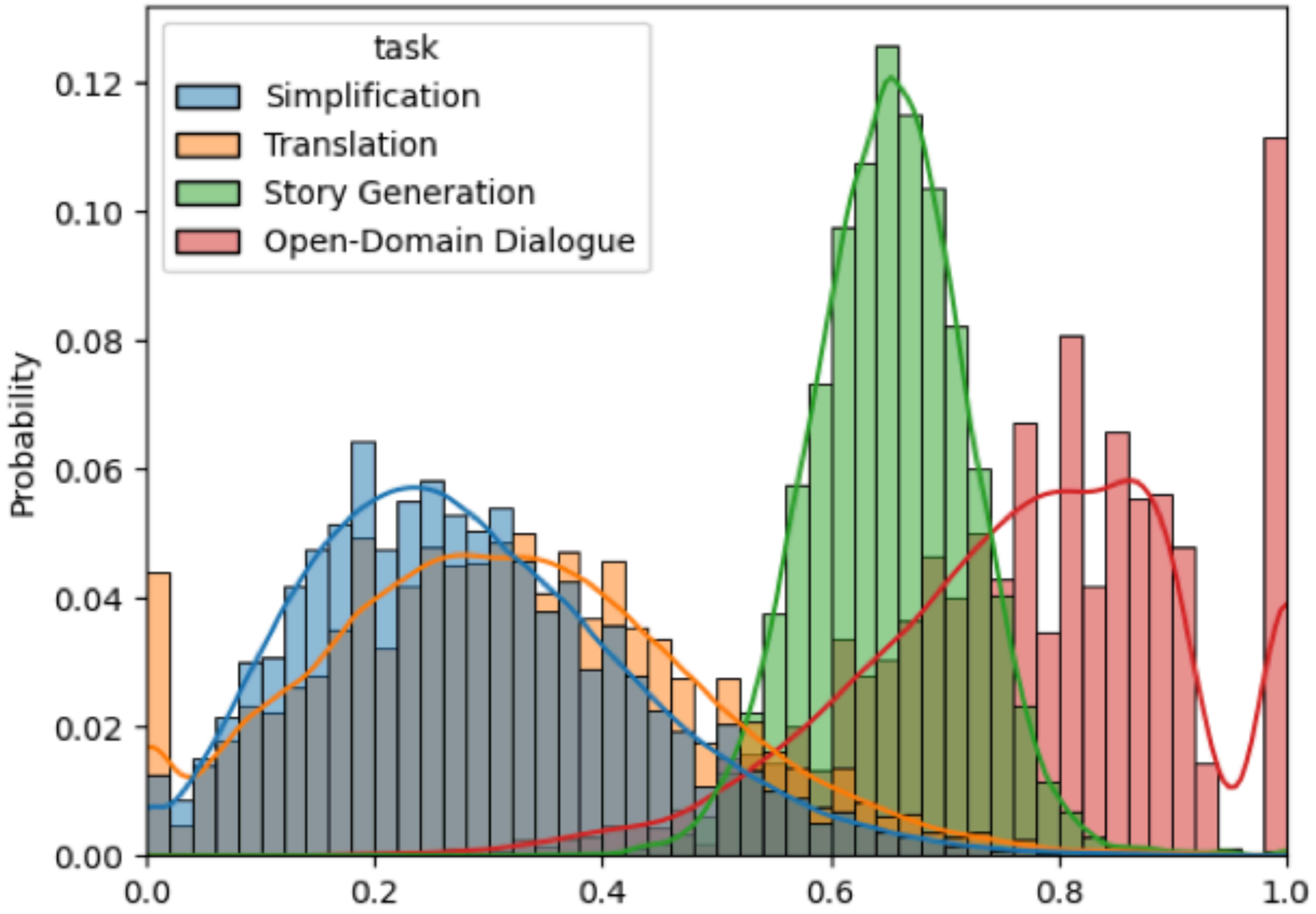
Models: Flan-T5 large (Chung et al., 2022) pre-trained and **finetuned** on *ASSET*

Open-domain dialogue

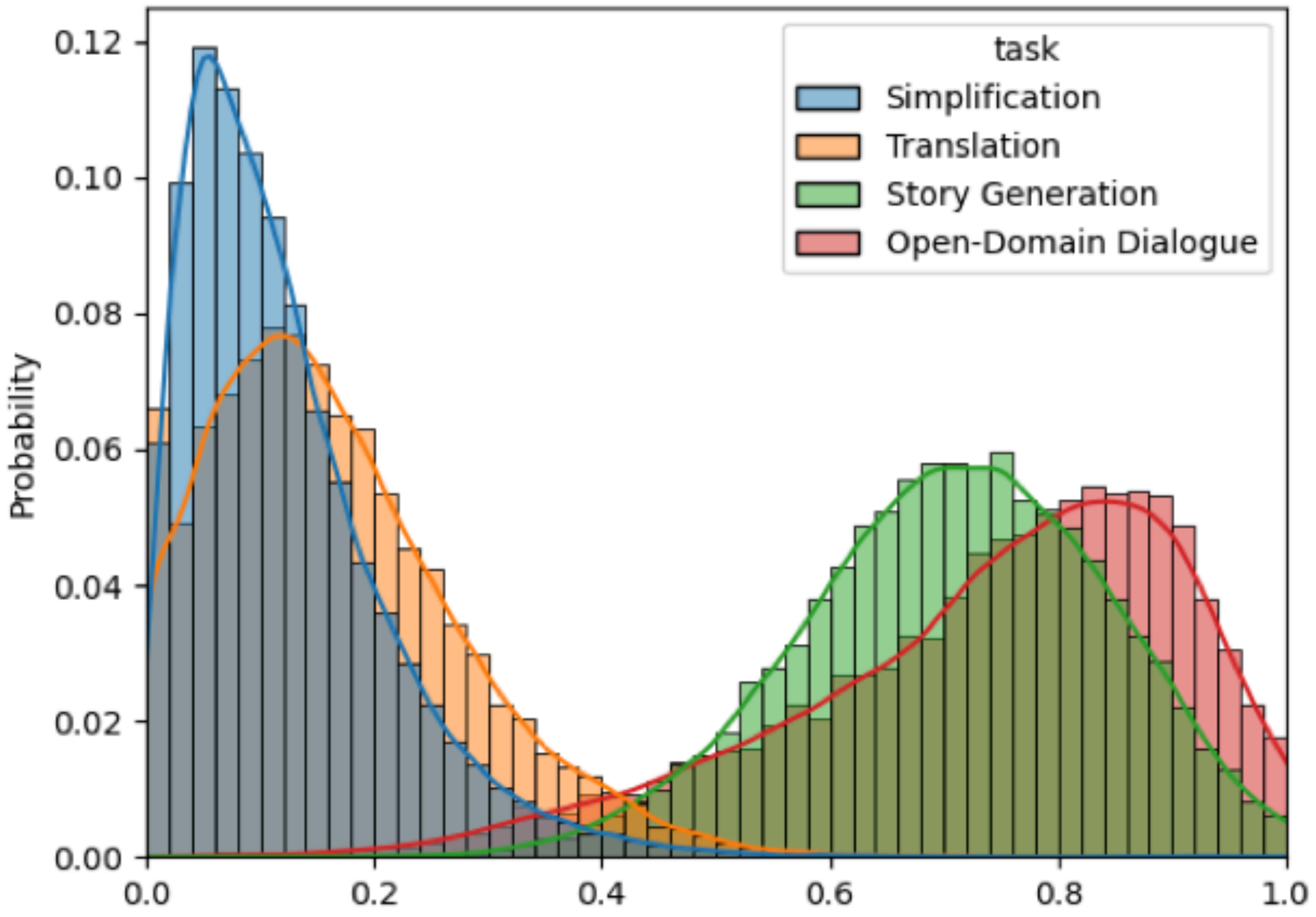
Data: 1028 dialogue contexts w/ 5 responses from *DailyDialog*++ (Sai et al., 2020)

Models: DialoGPT large (Zhang et al., 2019) **pre-trained** and finetuned on *DailyDialog* (Li et al., 2017)

Human production variability across tasks



(a) Lexical variability

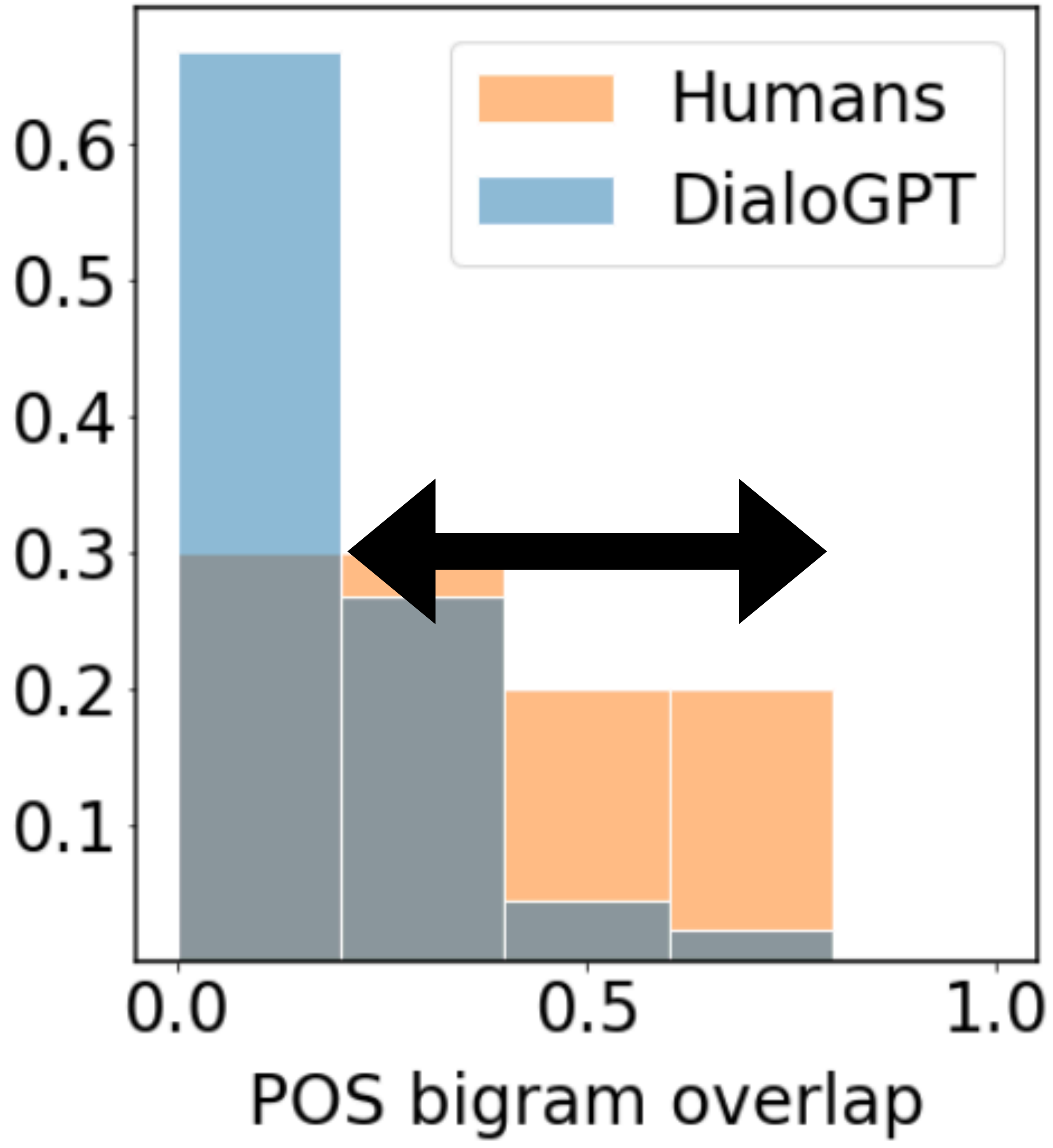


(c) Semantic variability

Do neural text generators reproduce human production variability?

Do text generators reproduce human production variability?

$\mathrm{Model}_k(x) := k(\hat{Y}, \hat{Y})$

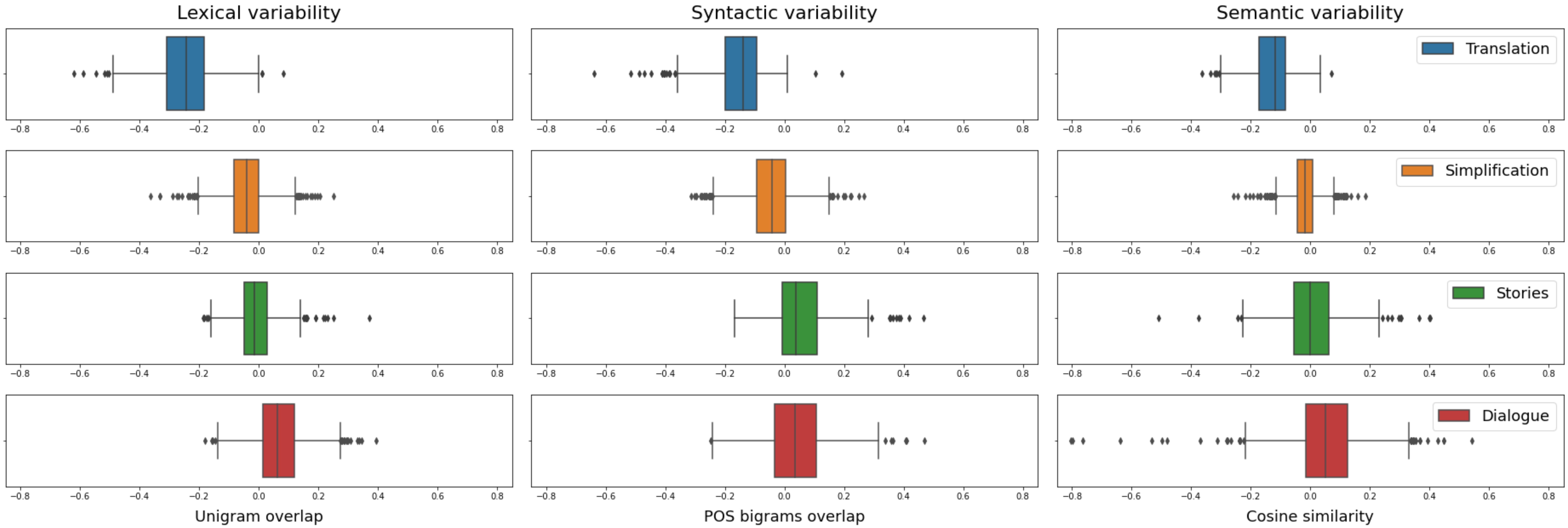


$\mathrm{Human}_k(x) := k(Y, Y)$

Quantify deviation from plausible human variability via a statistical divergence $\mathrm{D}(\,\cdot\,, H_k(x))$

We use the Wasserstein 1-distance (D_{W_1}) and $\mathrm{D}_\mu = \mu_{H_k(x)} - \mu_{\cdot_k(x)}$

Do text generators reproduce human production variability?



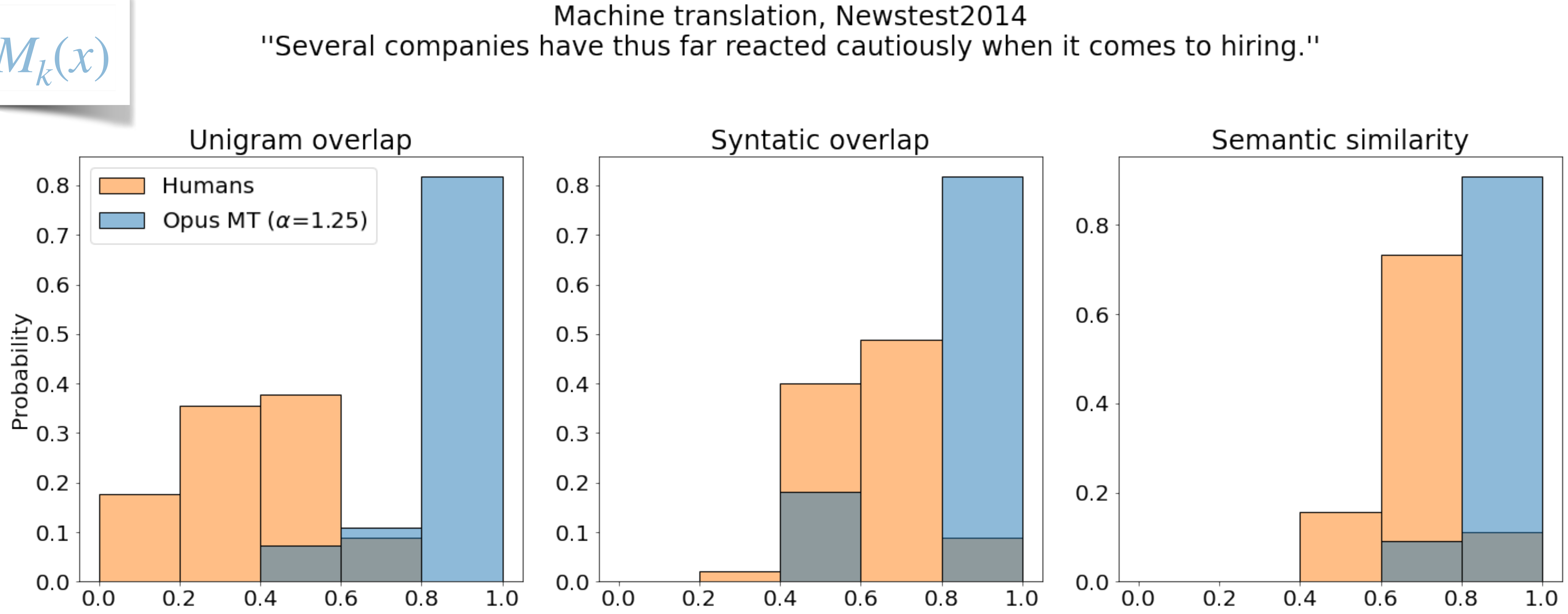
Distribution of $\mathrm{D}_{\mu}(M_k(x), H_k(x))$ over instances (10 human productions & 10 unbiased samples; 5 for dialogue).

$\mathrm{D}_{\mu} > 0$ indicates the model is overestimating the variability of the task; $\mathrm{D}_{\mu} < 0$ indicates variability underestimation.

Qualitative analysis of miscalibrated instances

Variability underestimation in translation

$H_k(x)$ vs. $M_k(x)$



Even when context and communicative intent are fixed, speakers' **linguistic realisations** of the communicative intent may vary (Levelt, 1993)

Several companies have thus far reacted cautiously when it comes to hiring.

Viele Firmen haben bisher vorsichtig reagiert, wenn es um Neuanstellungen geht.

Einige Unternehmen haben bisher bei der Einstellung vorsichtig reagiert.

Mehrere Unternehmen haben bisher zurücckhaltend reagiert, wenn es um die Einstellung von Mitarbeitern geht.

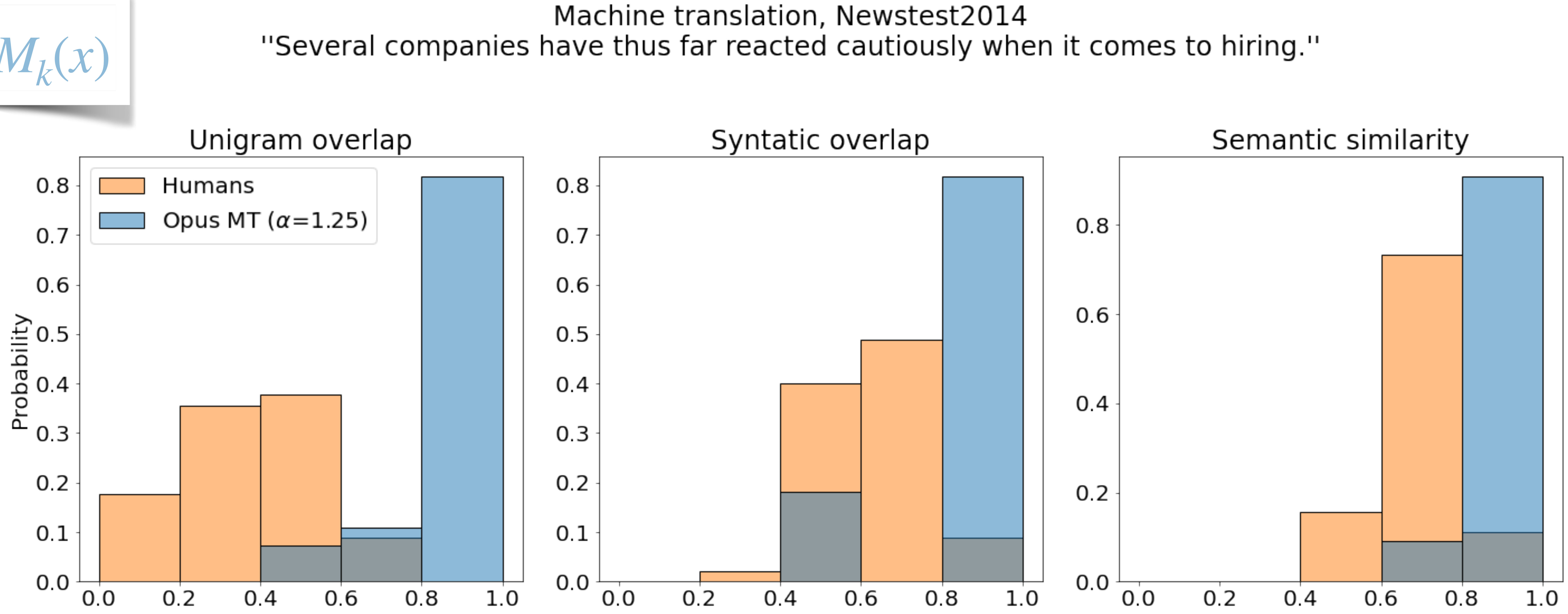
In Bezug auf Neuanstellungen haben diverse Unternehmen bisher mit Vorsicht reagiert.

Einige Unternehmen haben darauf mit Vorsichtsmaßnahmen reagiert wenn es um Neueinstellungen geht.

Source and translations: newstest 2014 (Bojar et al., 2014; Ott et al., 2018)

Variability underestimation in translation

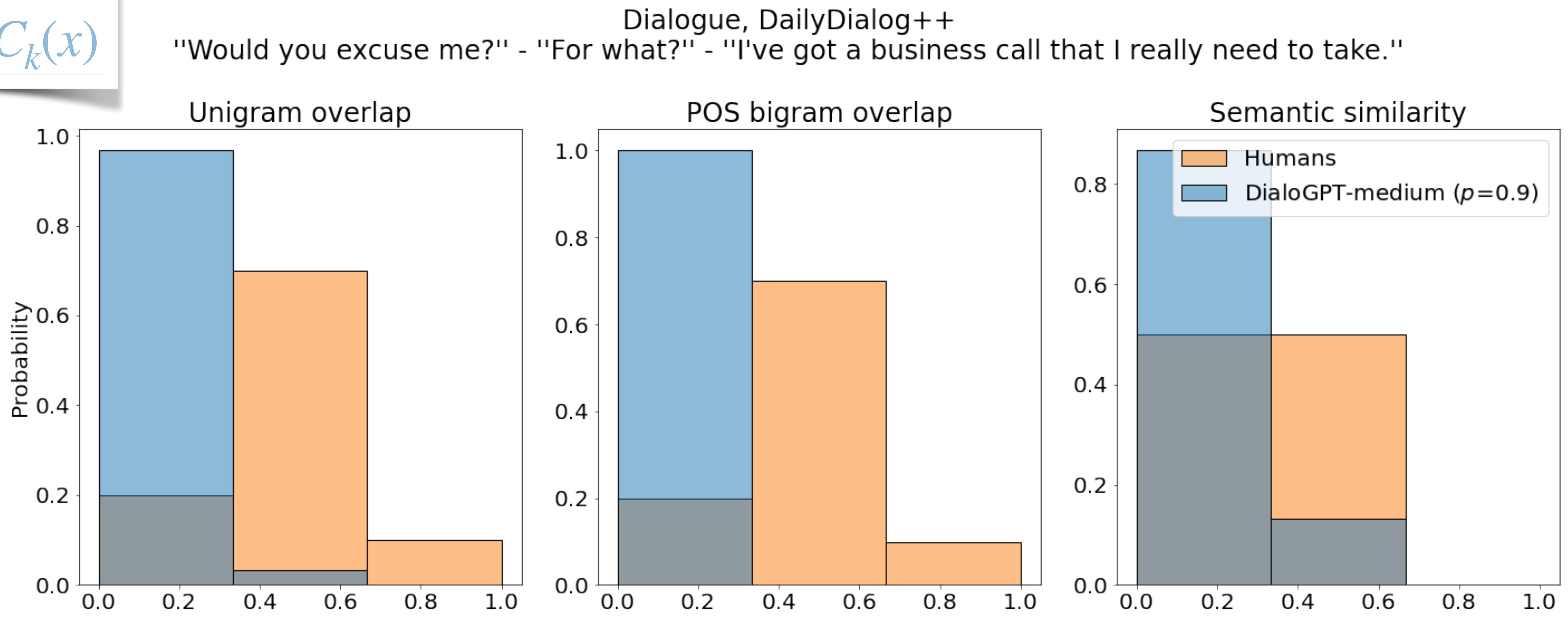
$H_k(x)$ vs. $M_k(x)$



All ten generations contain the German phrase *“vorsichtig reagiert”* as a translation for *“reacted cautiously”*.

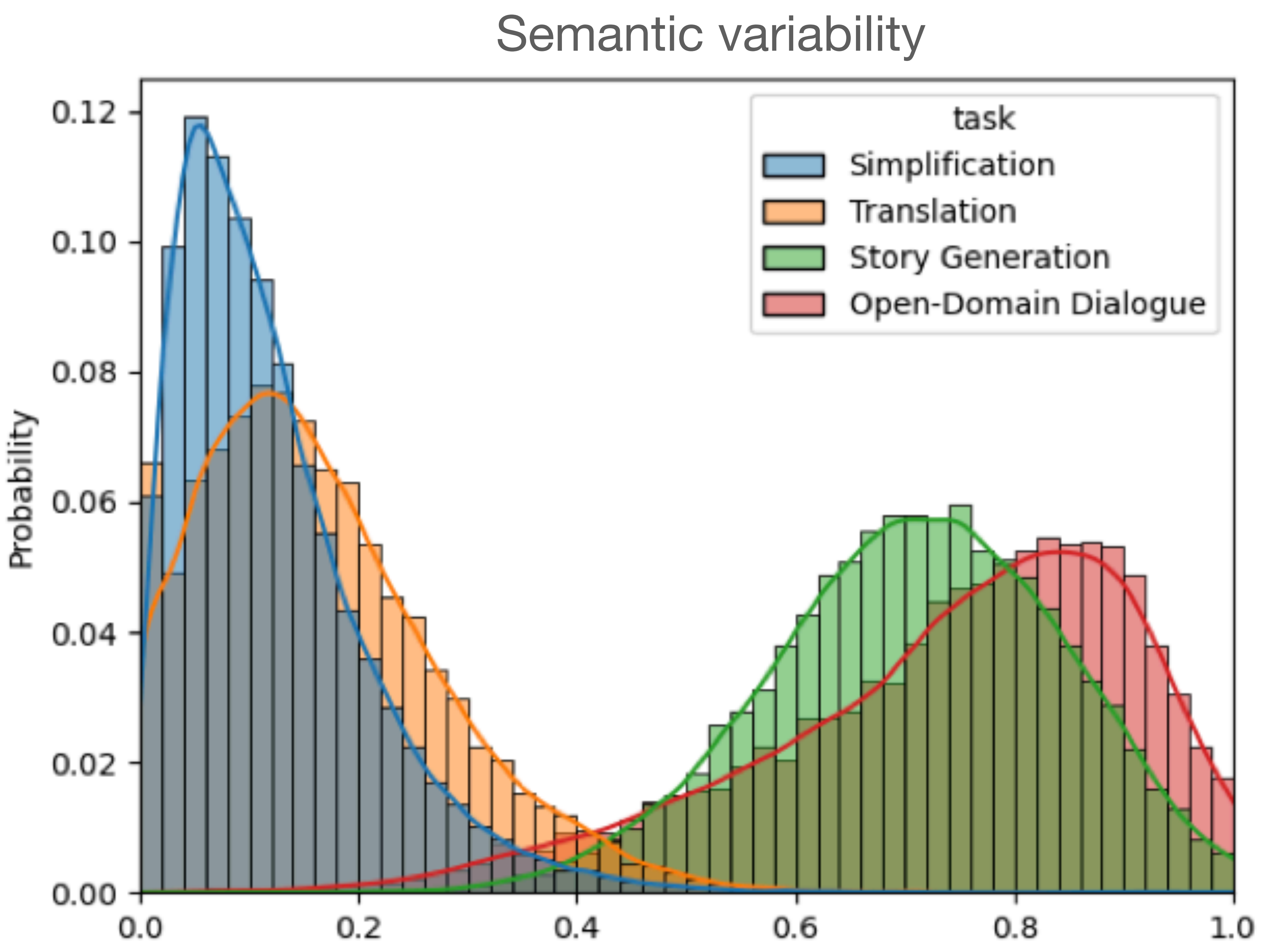
Variability overestimation in open-ended dialogue

$H_k(x)$ vs. $C_k(x)$

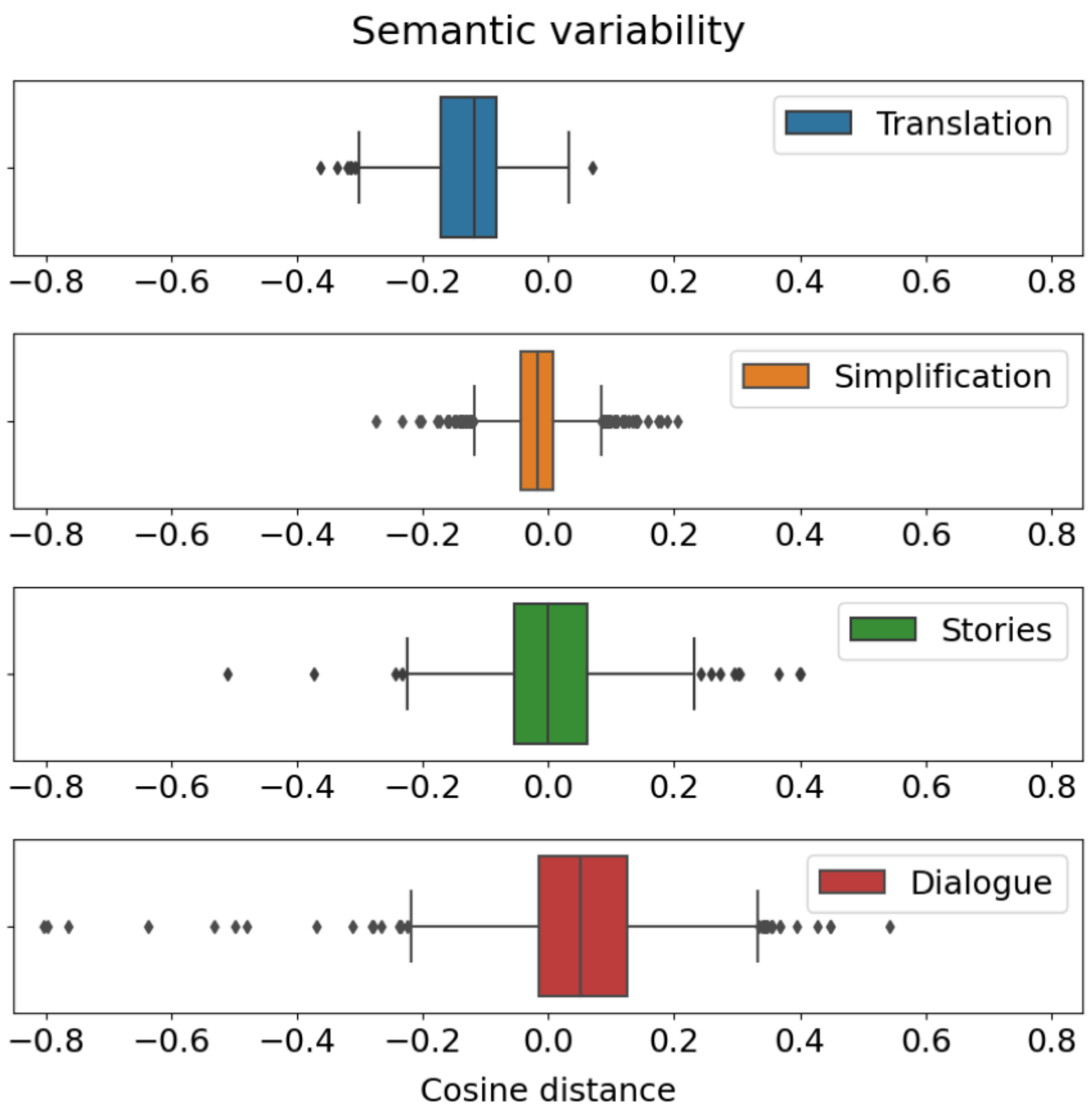


Humans reply with short affirmative responses (*"Okay! Please.", "Well! Go on.", "Sure, why not!", "Sure! Go ahead.", "Yes! Sure."*) while generated responses are mostly lengthy—and sometimes incoherent—statements (e.g., *"You don't need a business call. You need a friend"*).

Overall, LLMs approximate human production variability relatively well.



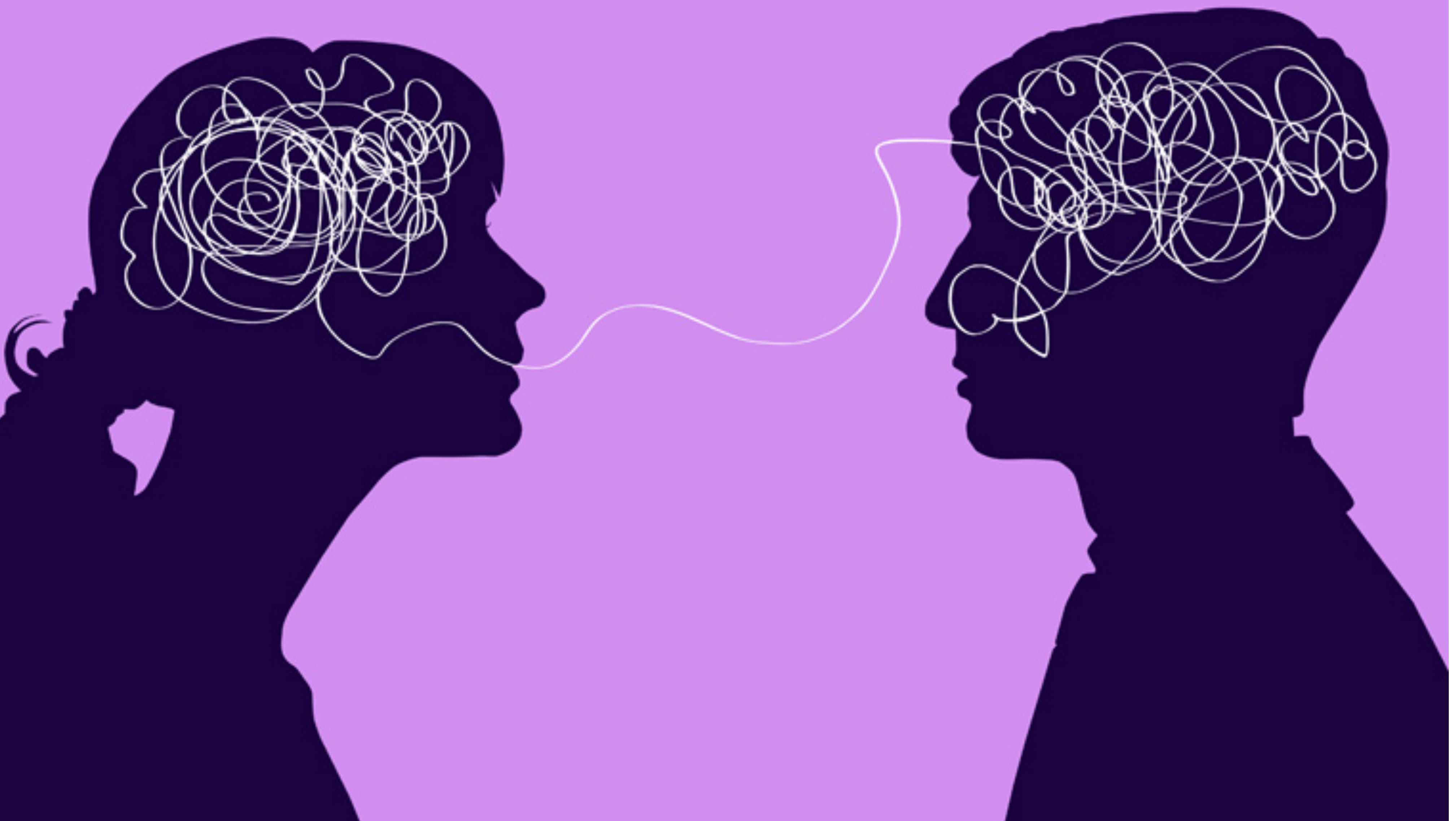
Semantic variability



Semantic variability

→ Can we use them to investigate psycholinguistic questions?

Using language is effortful

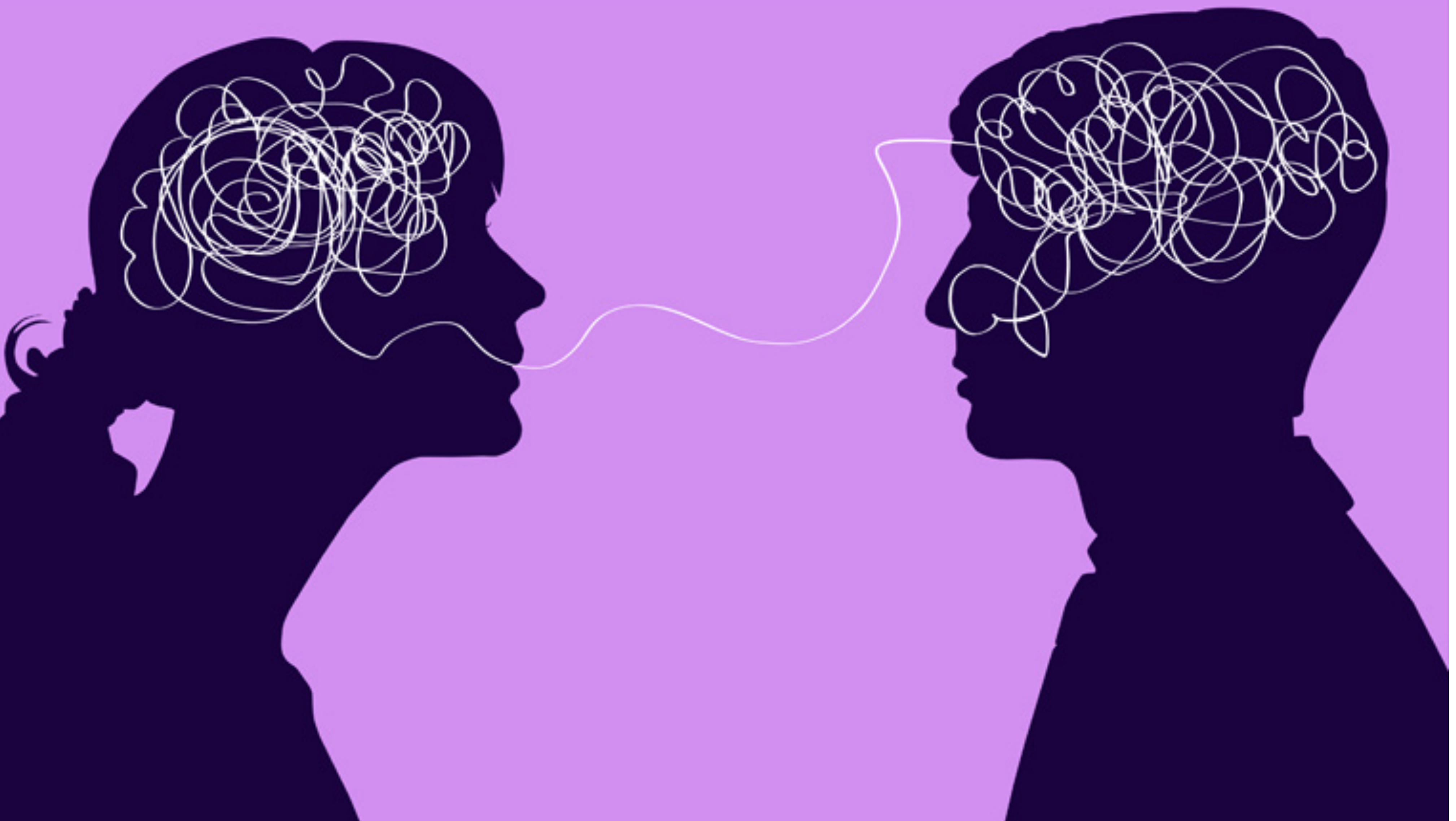


Speakers and addressees balance this effort collaboratively:

- Addressees actively ***predict*** what will be said next.
- Speakers take into account the ***processing effort*** of their addresses when deciding how to formulate a message.

predictability ≈ processing effort

Using language is effortful



How to capture processing effort?

Shannon's *information content* or ***surprisal*** measures the predictability of a word in context

$$\text{surprisal}(w) = -\log_2 P(w \mid C)$$

New proposal for quantifying utterance predictability: ***information value***

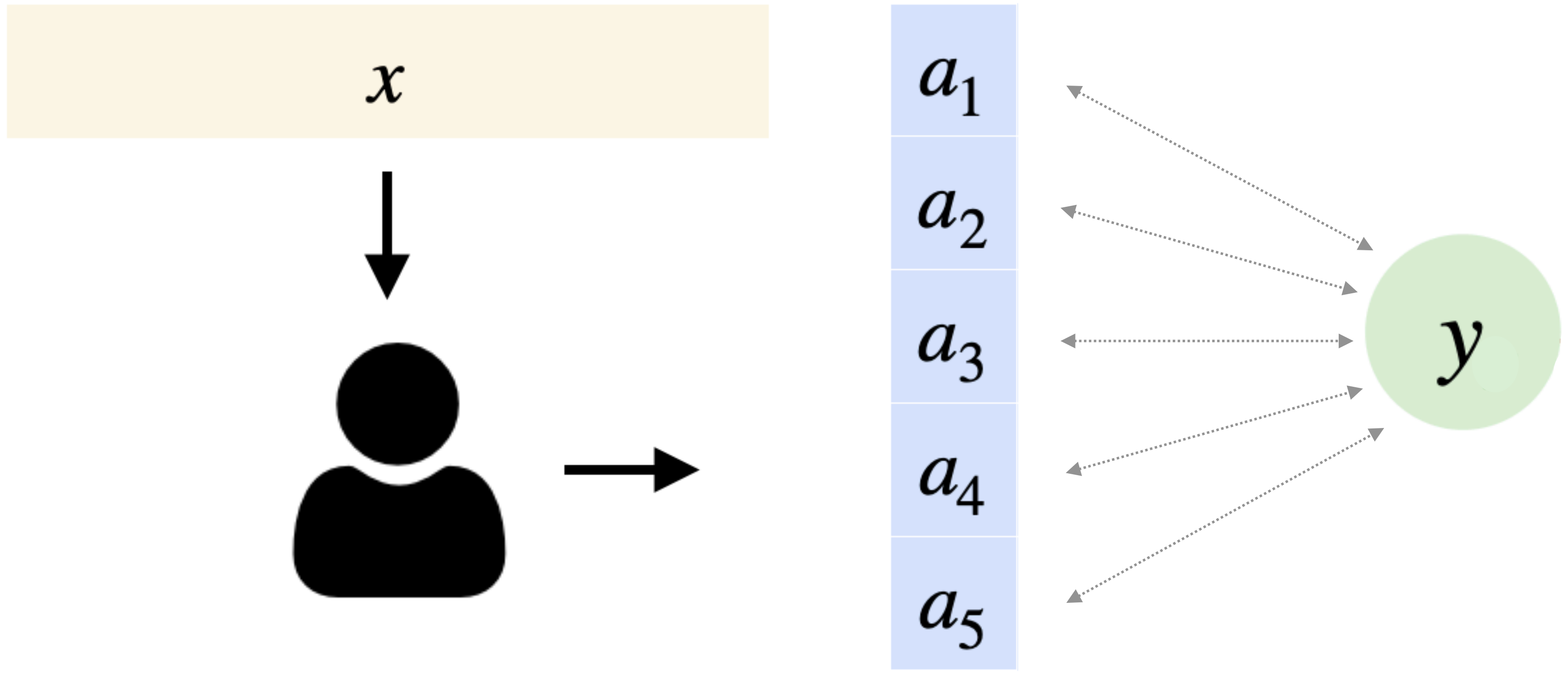
- Operationalises predictability as distance from plausible alternatives
- Exploits LLMs to generate alternatives

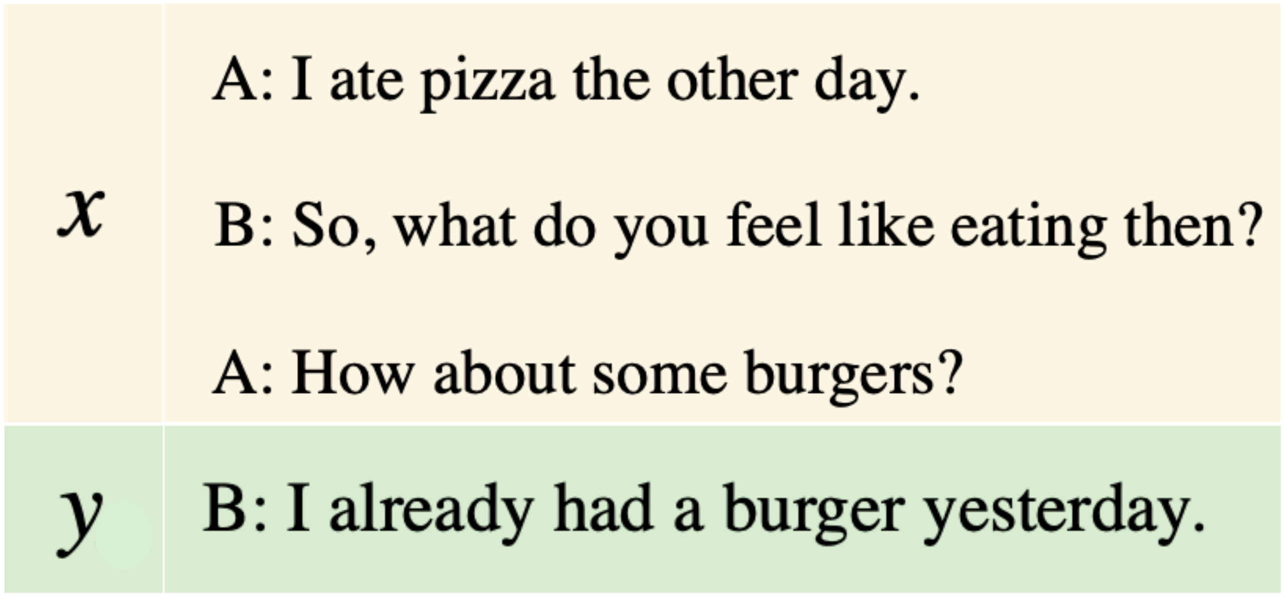
Information value

Utterance predictability as distance from plausible alternatives

A novel framework for quantifying utterance predictability:

Given a **context** x, a speaker may produce a number of plausible utterances. We refer to these as A_x, the **alternative set**.





x	A: I ate pizza the other day.
	B: So, what do you feel like eating then?
	A: How about some burgers?
y	B: I already had a burger yesterday.

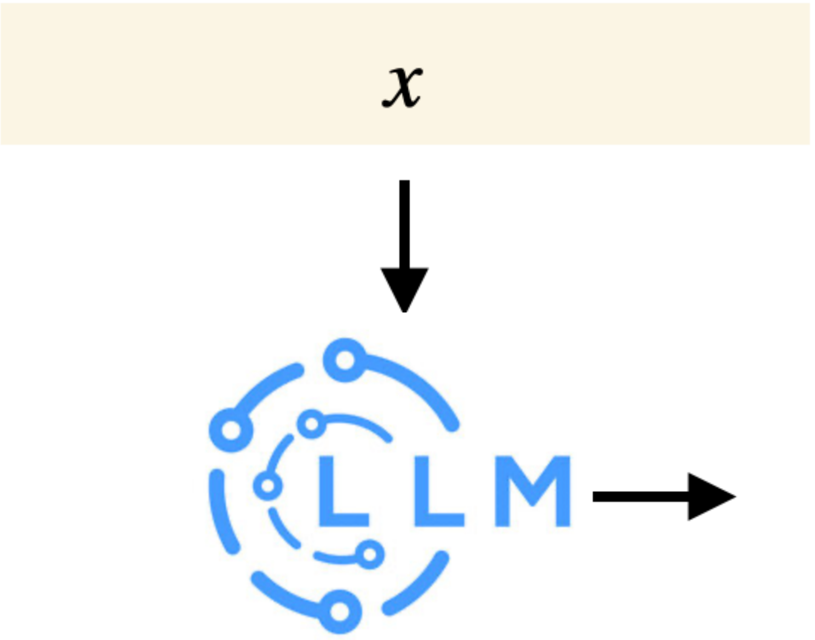
The **information value** of a next utterance y is defined as:

$$I(Y = y \mid X = x) := d(y, A_x)$$

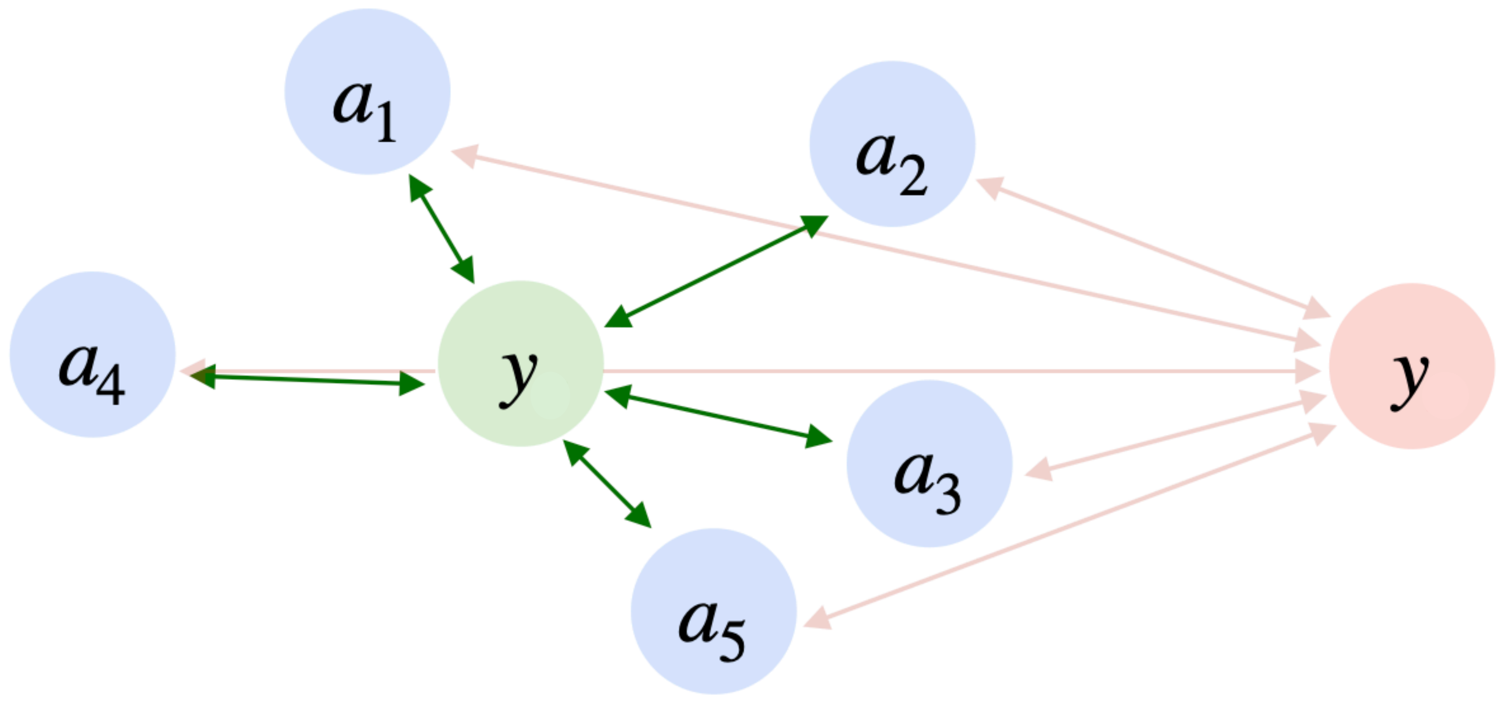
Information value

Utterance predictability as distance from plausible alternatives

A method for computing information value using LLMs to generate the alternative set:

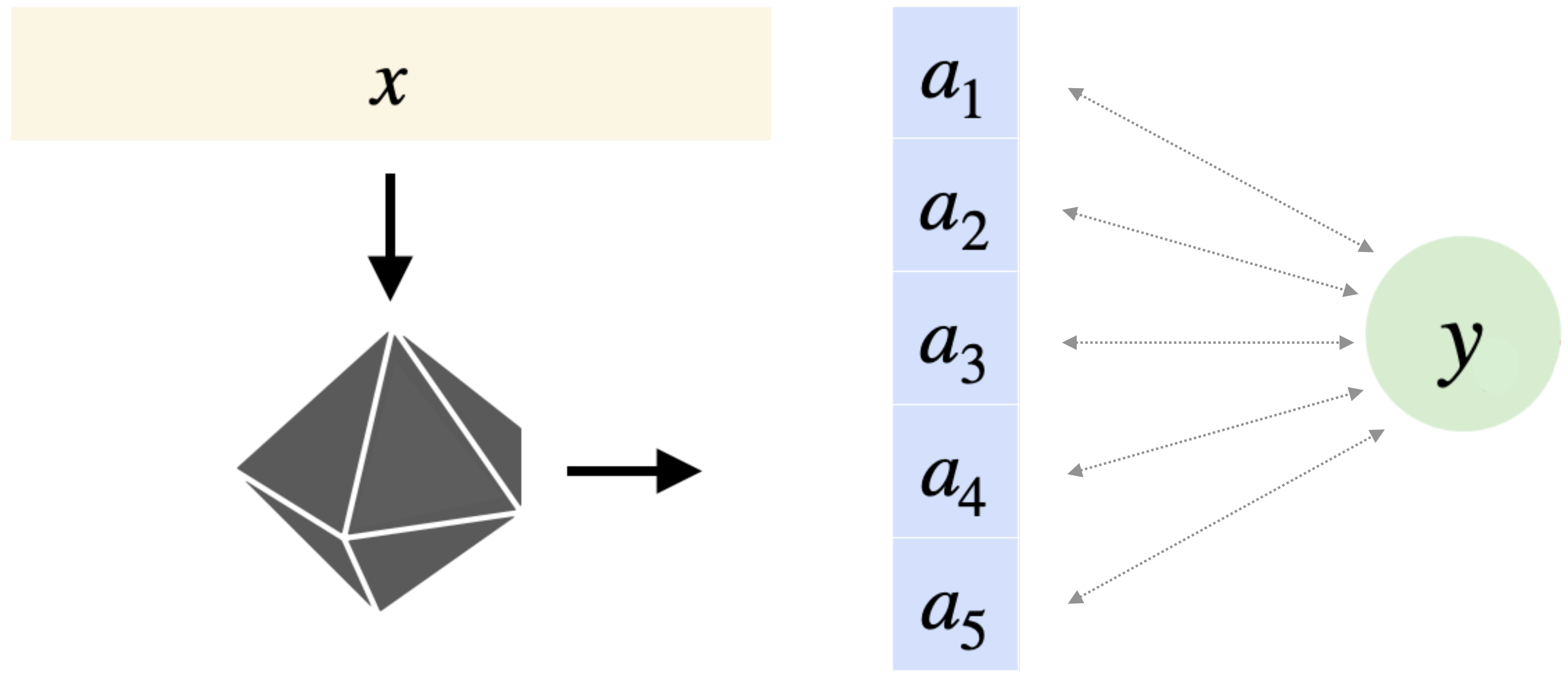


a_1	Well, I don't feel like eating burgers.
a_2	Good idea!
a_3	All right, I'll just go get an order of those right now.
a_4	I like Chinese food.
a_5	I'm not all that hungry, but we can order something later.



Different distance measures (dimensions of predictability): *lexical, semantic, syntactic*

Method



$$I(Y = y \,|\, X = x) := d(y, A_x) \quad x, y, a \in \Sigma^*$$

Generator

$$A_x \sim p_{Y|X=x}$$

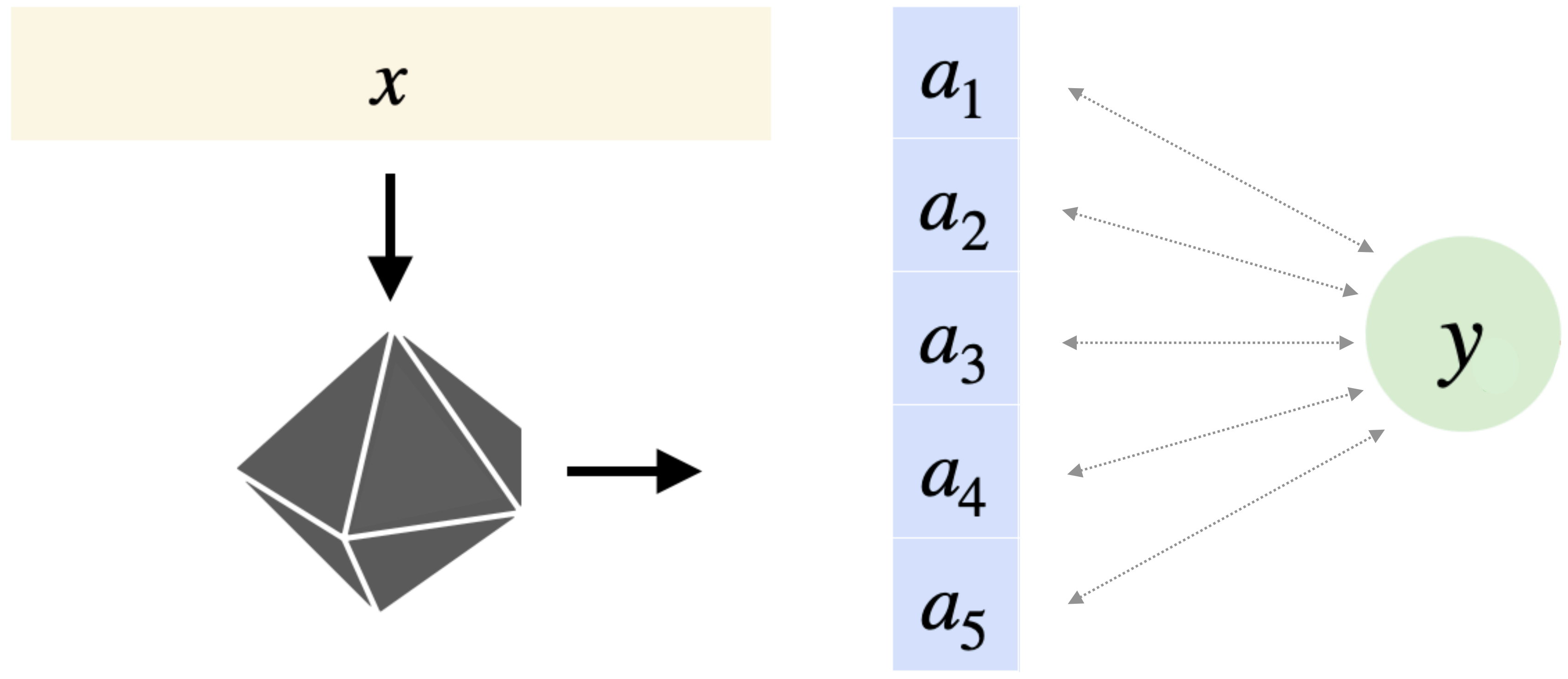
Distance metric

$$d : \Sigma^* \times \Sigma^* \to \mathbb{R}$$

Summary statistic

$$f : \mathbb{R}^{|A_x|} \to \mathbb{R}$$

Method



$$I(Y = y\,|\,X = x) := d(y, A_x) \quad x, y, a \in \Sigma^*$$

Generator

$$A_x \sim p_{Y|X=x}$$

language model
+
sampling algorithm

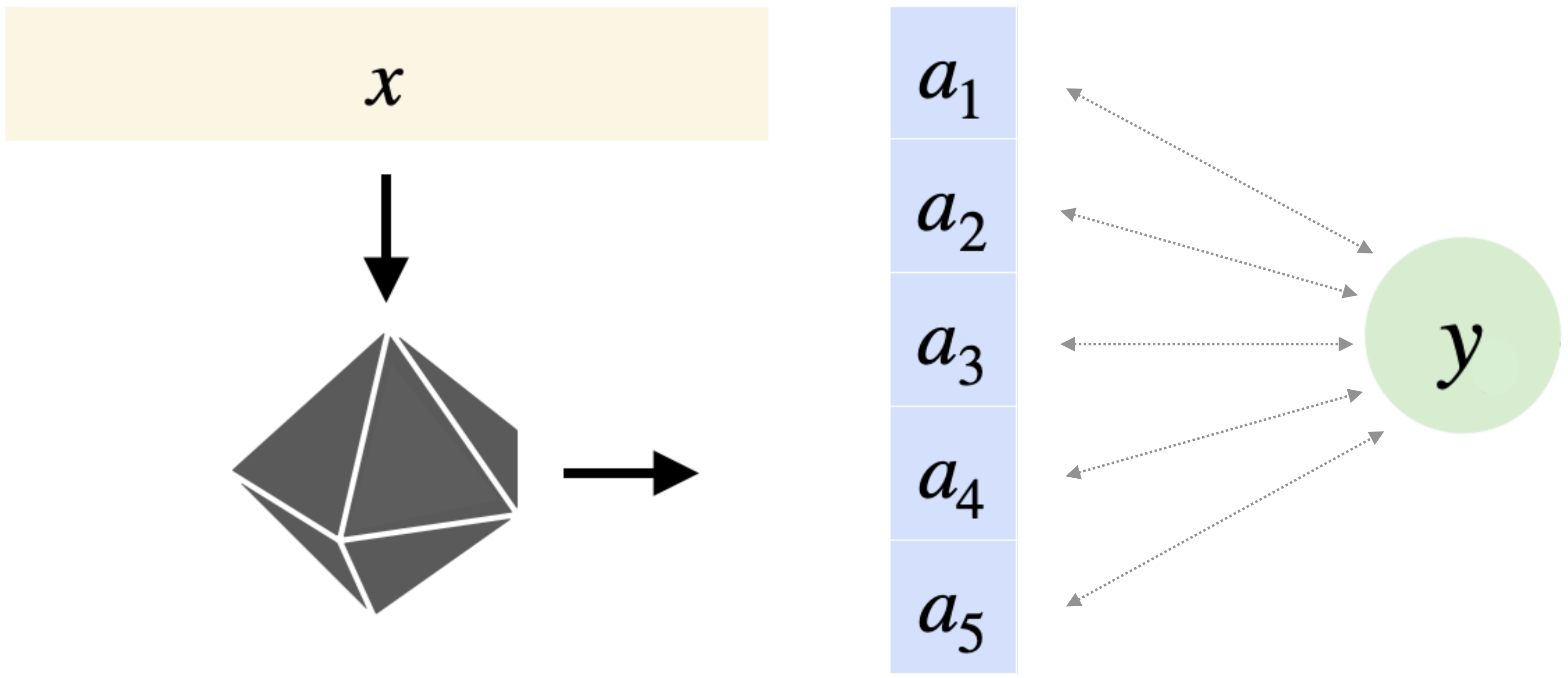
Distance metric

$$d : \Sigma^* \times \Sigma^* \to \mathbb{R}$$

Summary statistic

$$f : \mathbb{R}^{|A_x|} \to \mathbb{R}$$

Method



$$I(Y = y \,|\, X = x) := d(y, A_x) \quad x, y, a \in \Sigma^*$$

Generator

$$A_x \sim p_{Y|X=x}$$

language model
+
sampling algorithm

Distance metric

$$d : \Sigma^* \times \Sigma^* \to \mathbb{R}$$

Lexical: 1 - n-gram overlap

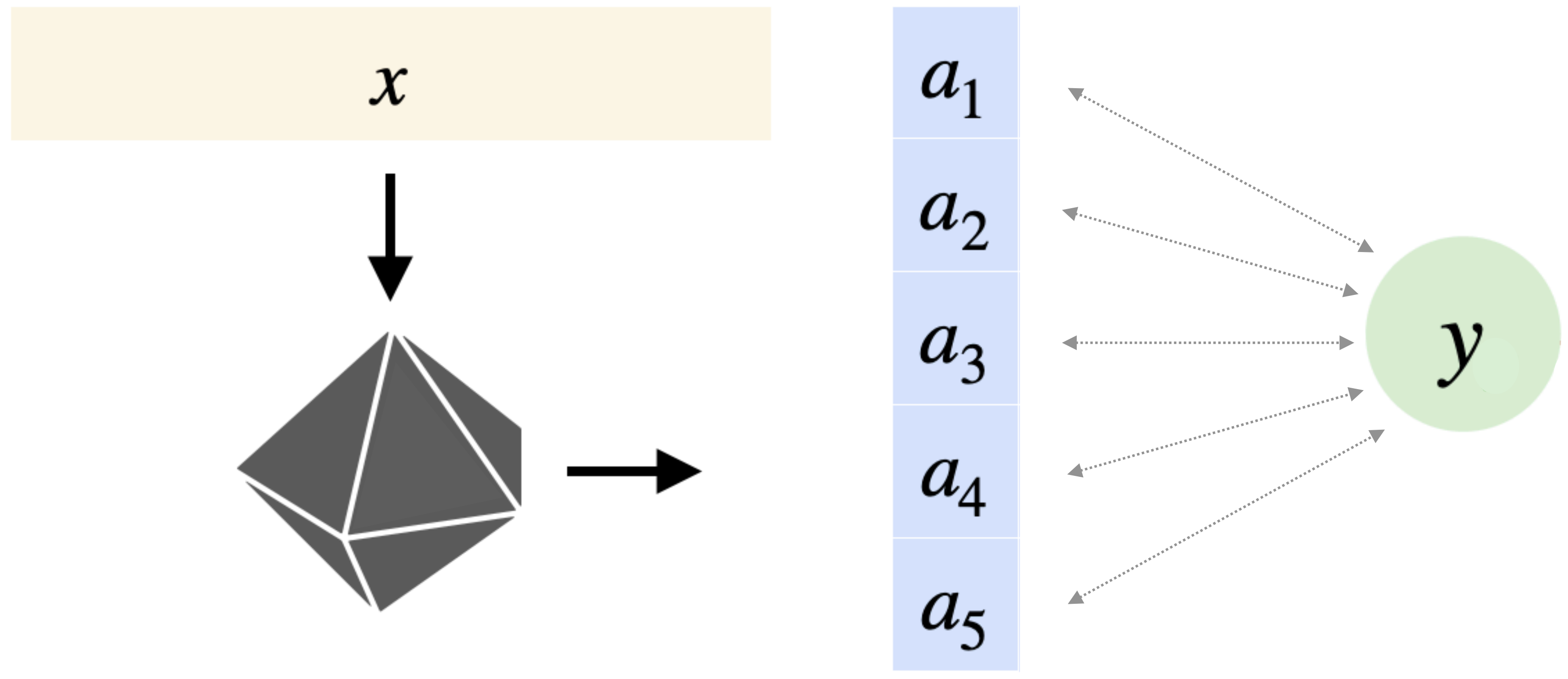
Syntactic: 1 - POS n-gram overlap

Semantic: cosine between sentence embeddings

Summary statistic

$$f : \mathbb{R}^{|A_x|} \to \mathbb{R}$$

Method



$$I(Y = y \,|\, X = x) := d(y, A_x) \quad x, y, a \in \Sigma^*$$

Generator

$A_x \sim p_{Y|X=x}$

language model
+
sampling algorithm

Distance metric

$d : \Sigma^* \times \Sigma^* \to \mathbb{R}$

Lexical: 1 - n-gram overlap

Syntactic: 1 - POS n-gram overlap

Semantic: cosine between sentence embeddings

Summary statistic

$f : \mathbb{R}^{|A_x|} \to \mathbb{R}$

Mean

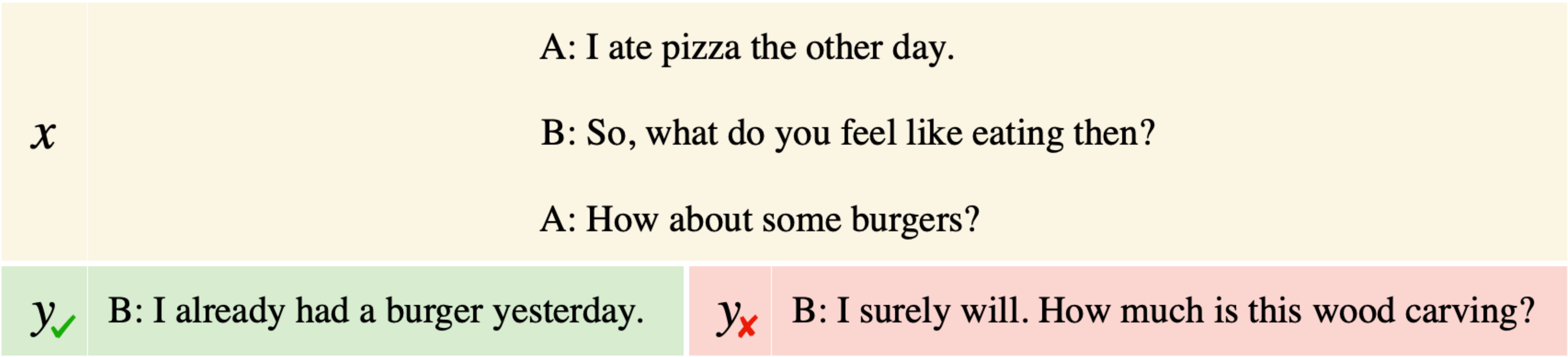
Minimum

Information value

Utterance predictability as distance from plausible alternatives

Can LLM-based information value estimates predict comprehension behaviour?

→ *Correlation with human acceptability judgements and reading times*



x	A: I ate pizza the other day. B: So, what do you feel like eating then? A: How about some burgers?		
$y_{\checkmark}$	B: I already had a burger yesterday.	$y_{\times}$	B: I surely will. How much is this wood carving?

reading times

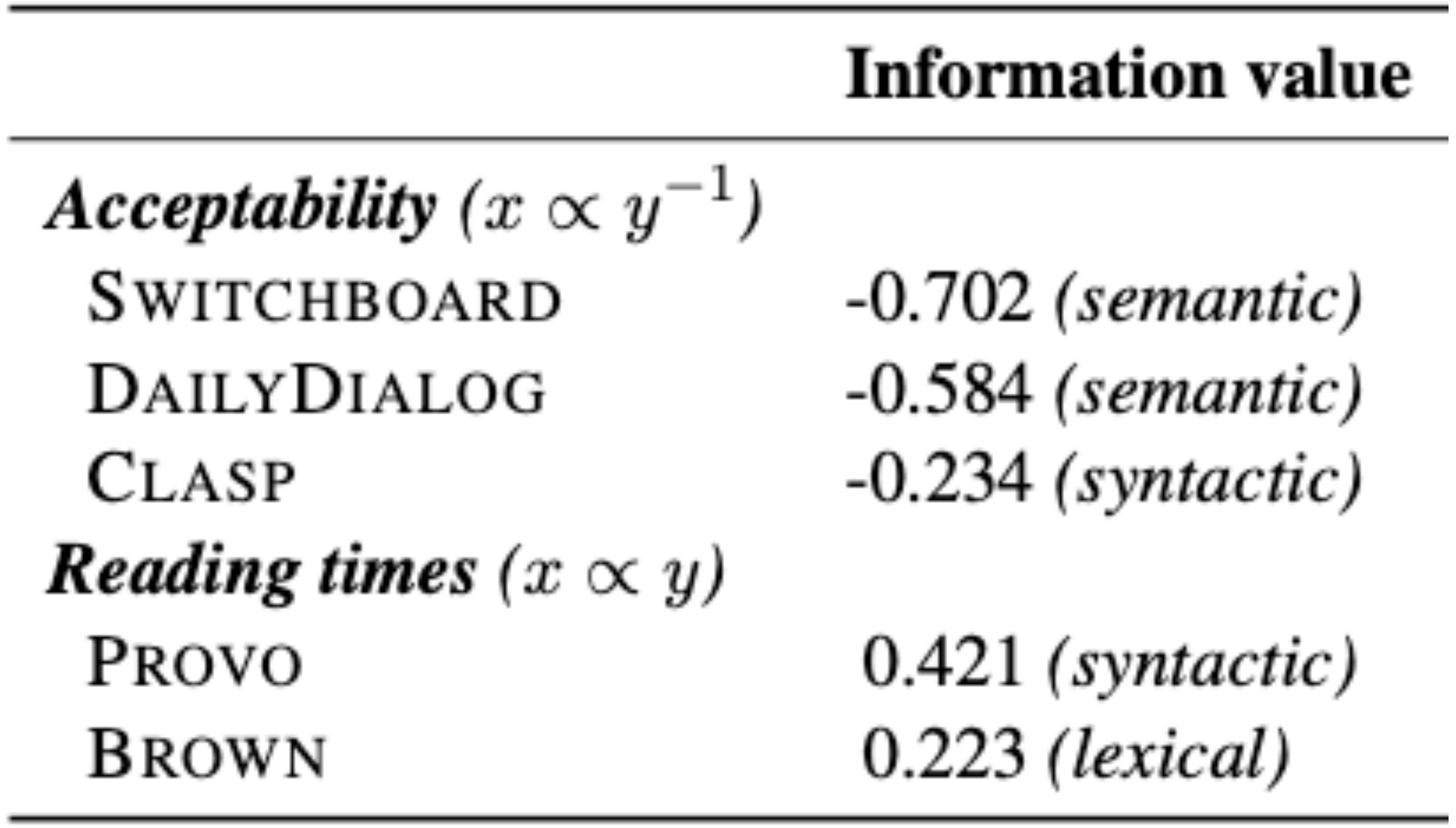
acceptability judgements

Experiments

Psychometric predictive power

Can LLM-based information value estimates predict comprehension behaviour?

→ *Correlation with human acceptability judgements and reading times*



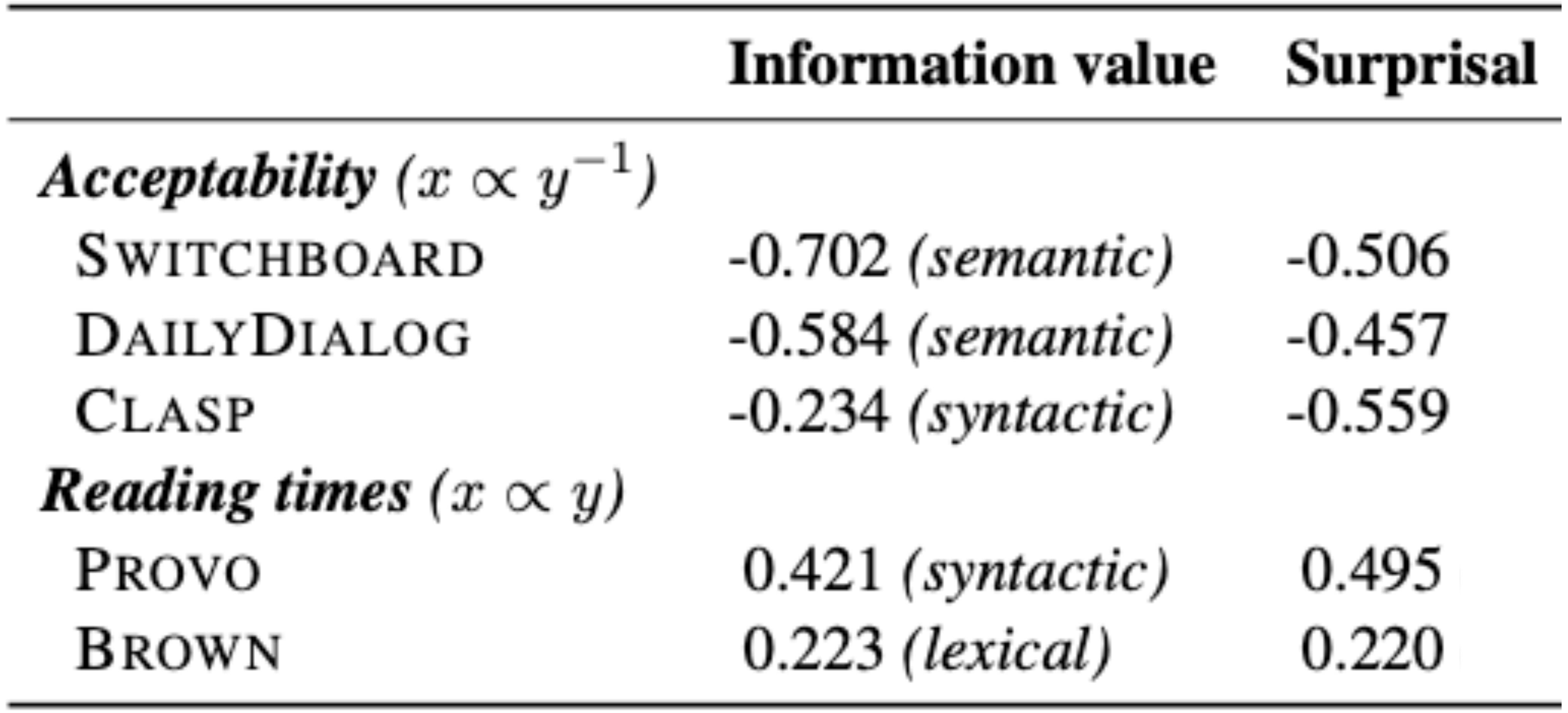
	Information value
Acceptability ($x \propto y^{-1}$)	
Switchboard	-0.702 *(semantic)*
DailyDialog	-0.584 *(semantic)*
Clasp	-0.234 *(syntactic)*
Reading times ($x \propto y$)	
Provo	0.421 *(syntactic)*
Brown	0.223 *(lexical)*

Experiments

Psychometric predictive power

Can LLM-based information value estimates predict comprehension behaviour?

→ *Correlation with human acceptability judgements and reading times*



	Information value	**Surprisal**
Acceptability $(x \propto y^{-1})$		
SWITCHBOARD	-0.702 *(semantic)*	-0.506
DAILYDIALOG	-0.584 *(semantic)*	-0.457
CLASP	-0.234 *(syntactic)*	-0.559
Reading times $(x \propto y)$		
PROVO	0.421 *(syntactic)*	0.495
BROWN	0.223 *(lexical)*	0.220

Experiments

Relation to utterance surprisal

Is the predictive power of information value complementary to that of surprisal?

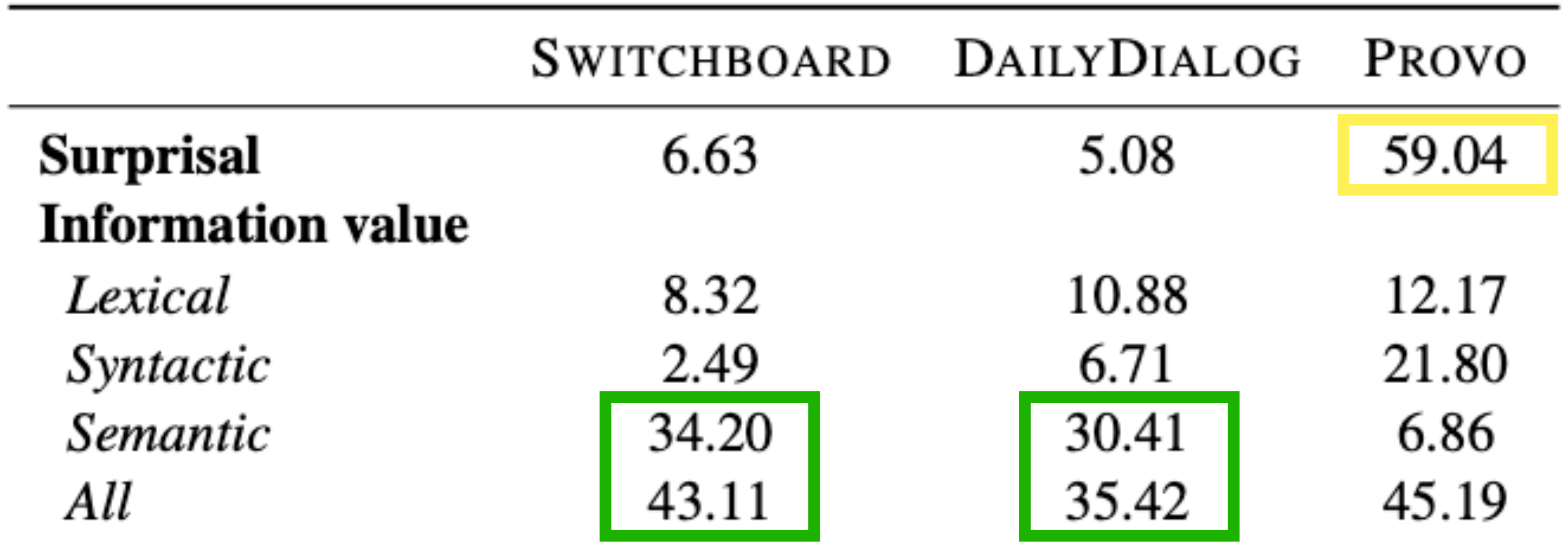
→ $\Delta LogLik$*: the difference in log-likelihood between a model with target predictor(s) and a baseline model with control predictors* [*Wilcox et al. 2020*]

Experiments

Relation to utterance surprisal

Is the predictive power of information value complementary to that of surprisal?

→$\Delta LogLik$: the difference in log-likelihood between a model with target predictor(s) and a baseline model with control predictors [Wilcox et al. 2020]



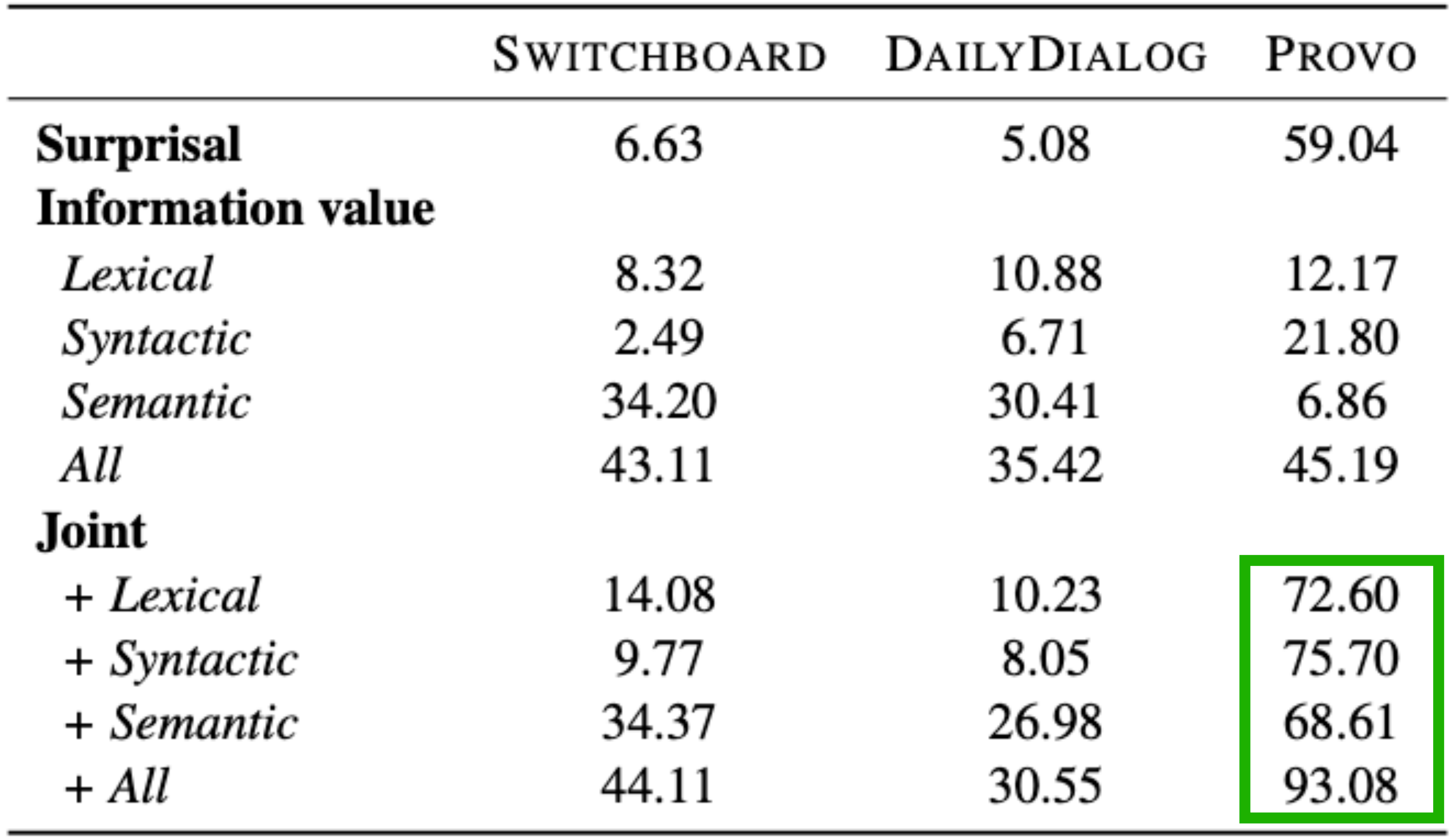
	SWITCHBOARD	DAILYDIALOG	PROVO
Surprisal	6.63	5.08	59.04
Information value			
Lexical	8.32	10.88	12.17
Syntactic	2.49	6.71	21.80
Semantic	34.20	30.41	6.86
All	43.11	35.42	45.19

Experiments

Relation to utterance surprisal

Is the predictive power of information value complementary to that of surprisal?

→$\Delta LogLik$: *the difference in log-likelihood between a model with target predictor(s) and a baseline model with control predictors* [Wilcox et al. 2020]



	SWITCHBOARD	DAILYDIALOG	PROVO
Surprisal	6.63	5.08	59.04
Information value			
Lexical	8.32	10.88	12.17
Syntactic	2.49	6.71	21.80
Semantic	34.20	30.41	6.86
All	43.11	35.42	45.19
Joint			
+ *Lexical*	14.08	10.23	72.60
+ *Syntactic*	9.77	8.05	75.70
+ *Semantic*	34.37	26.98	68.61
+ *All*	44.11	30.55	93.08

Summing up

Variability is an intrinsic property of human language production: each task has its own plausible levels of variability.

We propose a framework to evaluate neural text generators in terms of their ability to reproduced production variability (i.e., uncertainty) observed in humans.

- LLMs capture reasonably human production variability. But overestimated in more open-ended tasks and underestimated in more constrained tasks.

We propose to exploit **production variability** to quantify **utterance predictability** in comprehension

- **information value** quantifies the predictability of an utterance relative to a set of plausible alternatives

Summing up

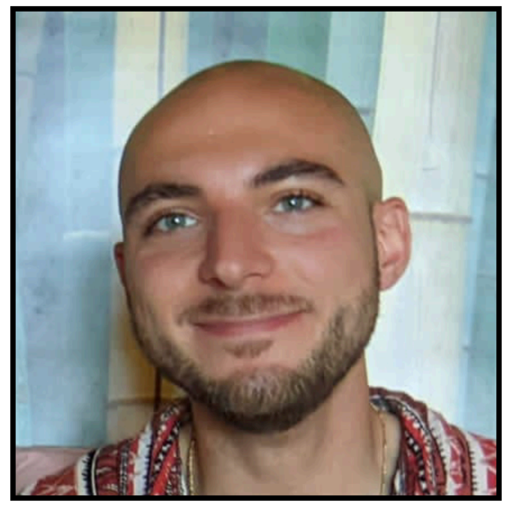
Information value:

- Captures variability and uncertainty **above the word level**, considering impact of **more abstract communicative units** like dialogue acts
- Enables different **dimensions of predictability** to be disentangled
- Is substantially **more predictive than surprisal** for acceptability judgements in dialogue and **complementary** for predicting eye-tracked reading times
- Robust estimates can be obtained from **neural text generators**

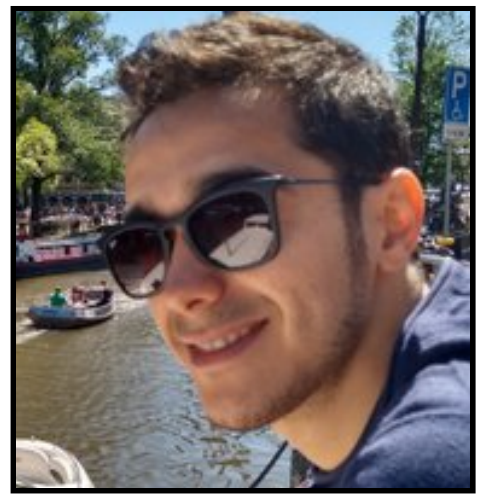
Thanks

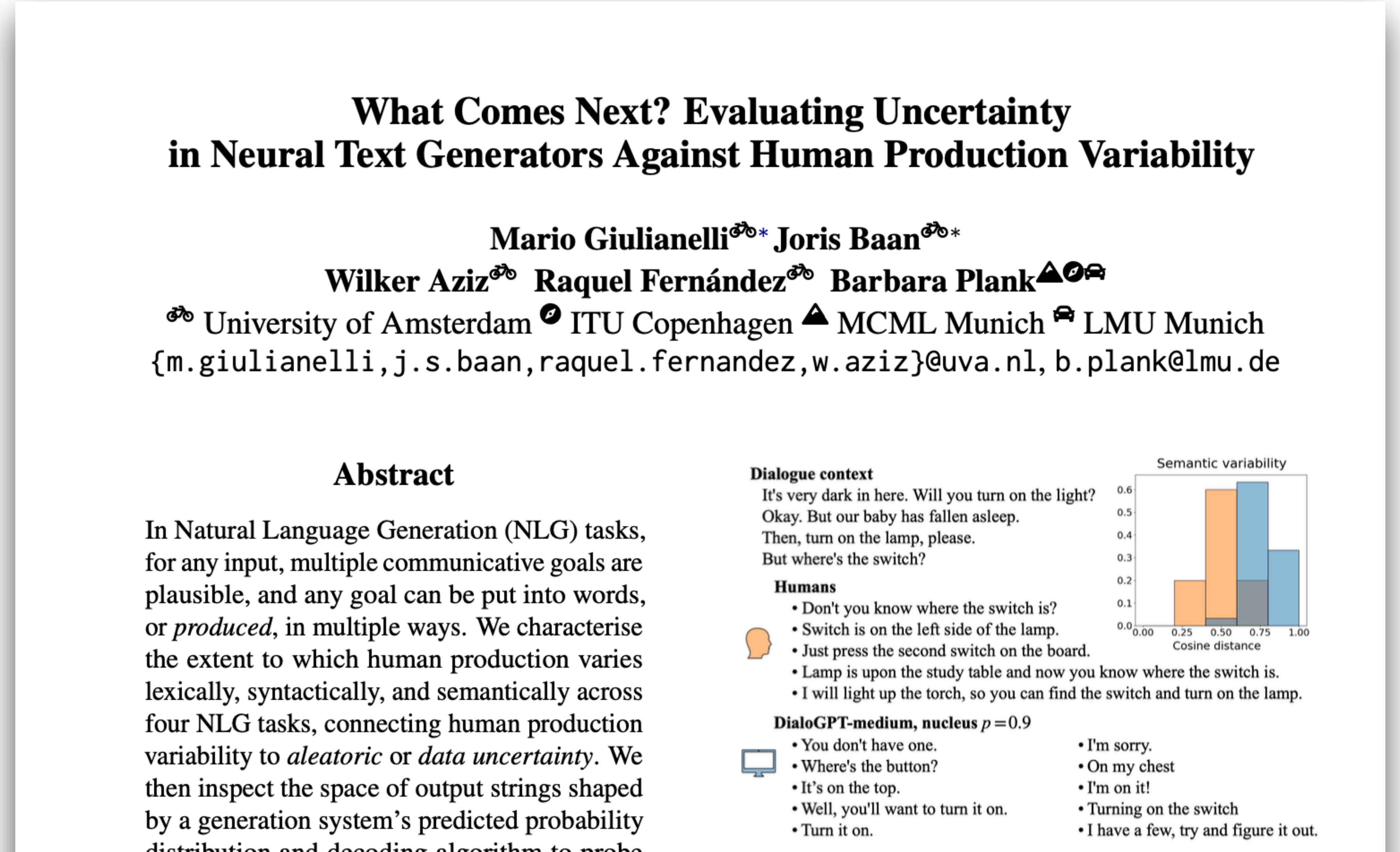
Joris Baan

Sarenne Wallbridge


Mario Giulianelli

Barbara Plank


Wilker Aziz



What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability

Mario Giulianelli[🚲*] Joris Baan[🚲*]
Wilker Aziz[🚲] Raquel Fernández[🚲] Barbara Plank[▲●🚗]
[🚲] University of Amsterdam [●] ITU Copenhagen [▲] MCML Munich [🚗] LMU Munich
{m.giulianelli,j.s.baan,raquel.fernandez,w.aziz}@uva.nl, b.plank@lmu.de

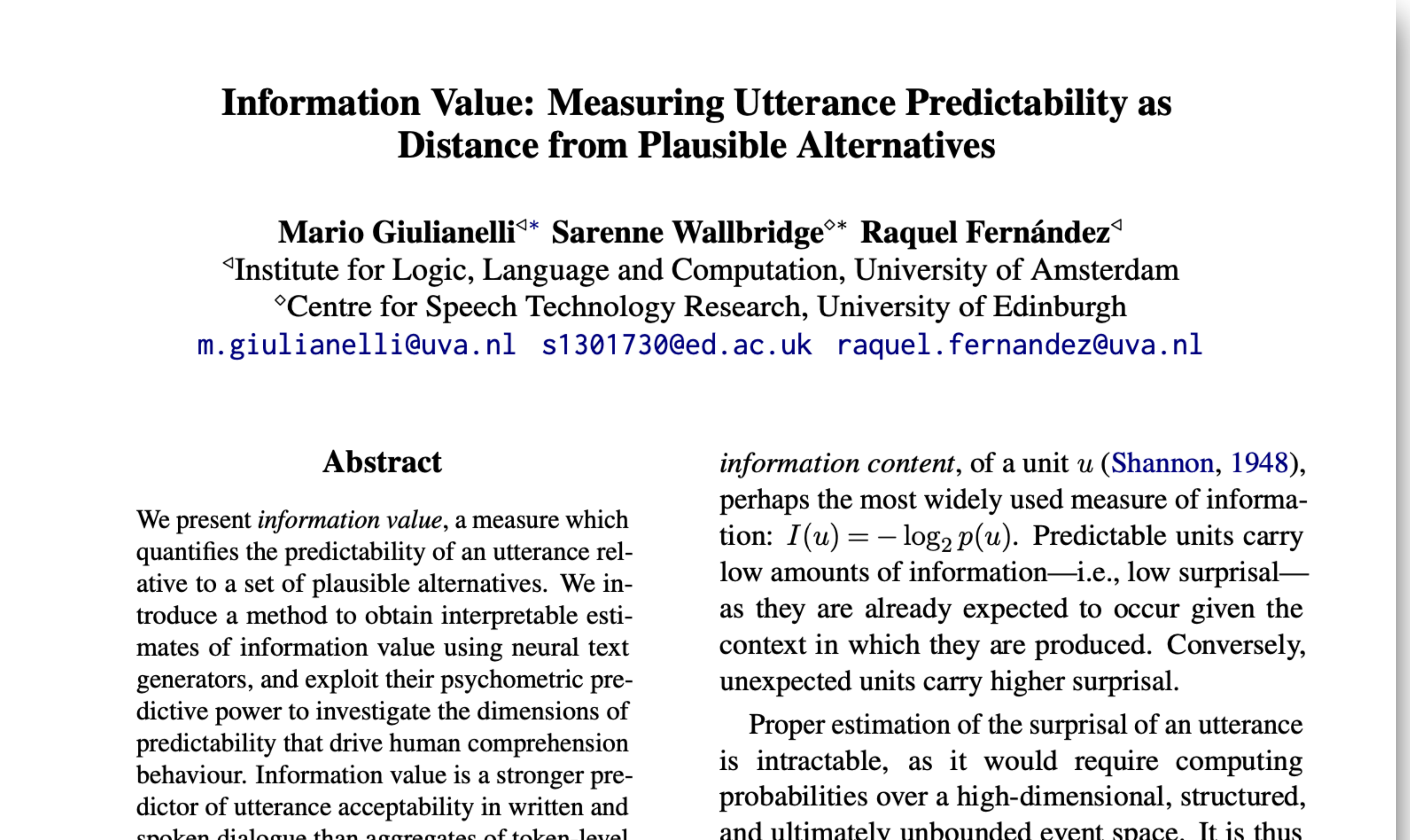
Abstract

In Natural Language Generation (NLG) tasks, for any input, multiple communicative goals are plausible, and any goal can be put into words, or *produced*, in multiple ways. We characterise the extent to which human production varies lexically, syntactically, and semantically across four NLG tasks, connecting human production variability to *aleatoric* or *data uncertainty*. We then inspect the space of output strings shaped by a generation system's predicted probability distribution and decoding algorithm to probe

Dialogue context
It's very dark in here. Will you turn on the light?
Okay. But our baby has fallen asleep.
Then, turn on the lamp, please.
But where's the switch?

Humans
- Don't you know where the switch is?
- Switch is on the left side of the lamp.
- Just press the second switch on the board.
- Lamp is upon the study table and now you know where the switch is.
- I will light up the torch, so you can find the switch and turn on the lamp.

DialoGPT-medium, nucleus $p = 0.9$
- You don't have one.
- Where's the button?
- It's on the top.
- Well, you'll want to turn it on.
- Turn it on.
- I'm sorry.
- On my chest
- I'm on it!
- Turning on the switch
- I have a few, try and figure it out.



Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives

Mario Giulianelli[◁*] Sarenne Wallbridge[◇*] Raquel Fernández[◁]
[◁]Institute for Logic, Language and Computation, University of Amsterdam
[◇]Centre for Speech Technology Research, University of Edinburgh
m.giulianelli@uva.nl s1301730@ed.ac.uk raquel.fernandez@uva.nl

Abstract

We present *information value*, a measure which quantifies the predictability of an utterance relative to a set of plausible alternatives. We introduce a method to obtain interpretable estimates of information value using neural text generators, and exploit their psychometric predictive power to investigate the dimensions of predictability that drive human comprehension behaviour. Information value is a stronger predictor of utterance acceptability in written and spoken dialogue than aggregates of token-level

information content, of a unit u (Shannon, 1948), perhaps the most widely used measure of information: $I(u) = -\log_2 p(u)$. Predictable units carry low amounts of information—i.e., low surprisal—as they are already expected to occur given the context in which they are produced. Conversely, unexpected units carry higher surprisal.

Proper estimation of the surprisal of an utterance is intractable, as it would require computing probabilities over a high-dimensional, structured, and ultimately unbounded event space. It is thus

(EMNLP 2023)