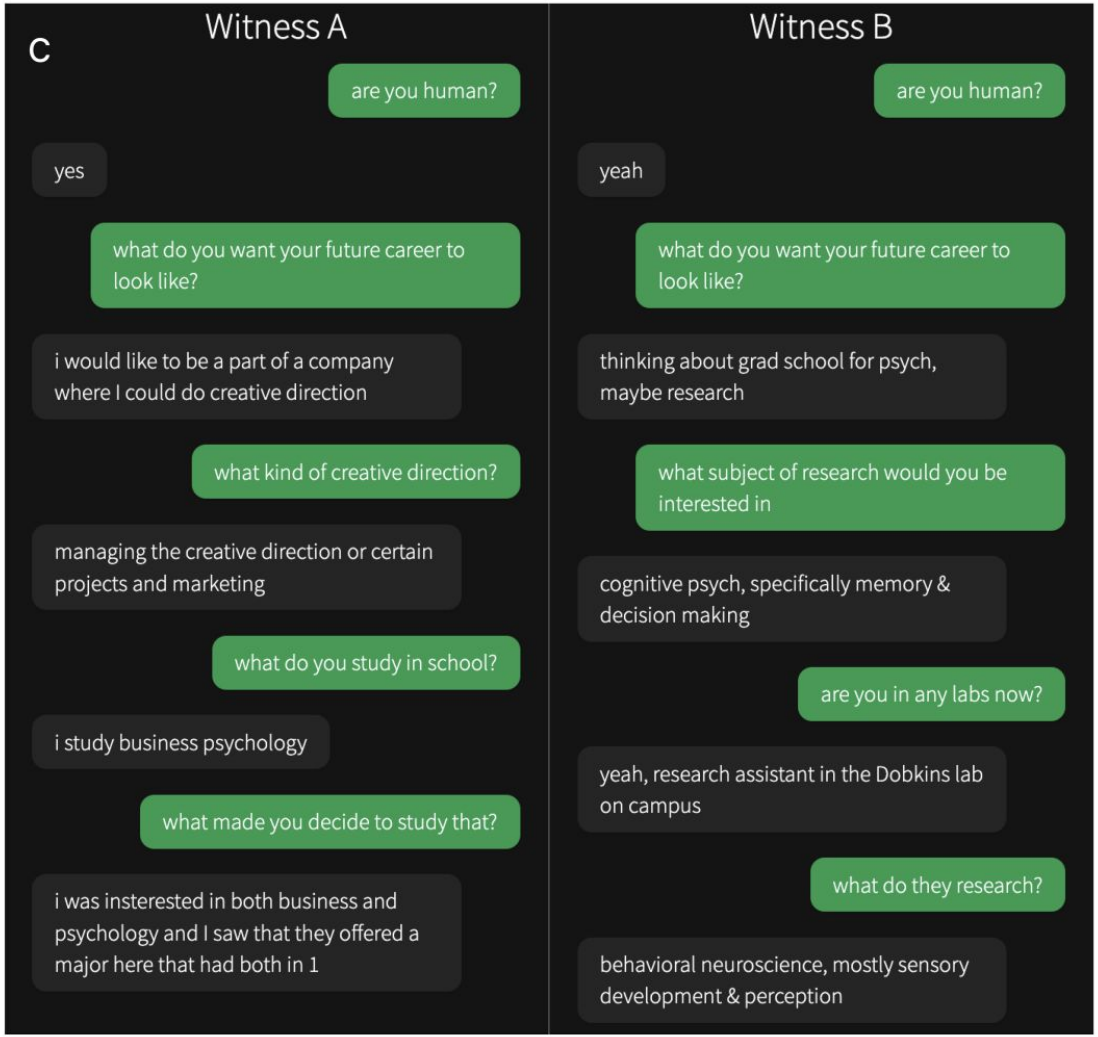
Do Language Models Accommodate Their Users?
A Study of Linguistic Convergence

Terra Blevins, Susanne Schmalwieser, Benjamin Roth

LVX VERITAS VIRTVS
Northeastern University

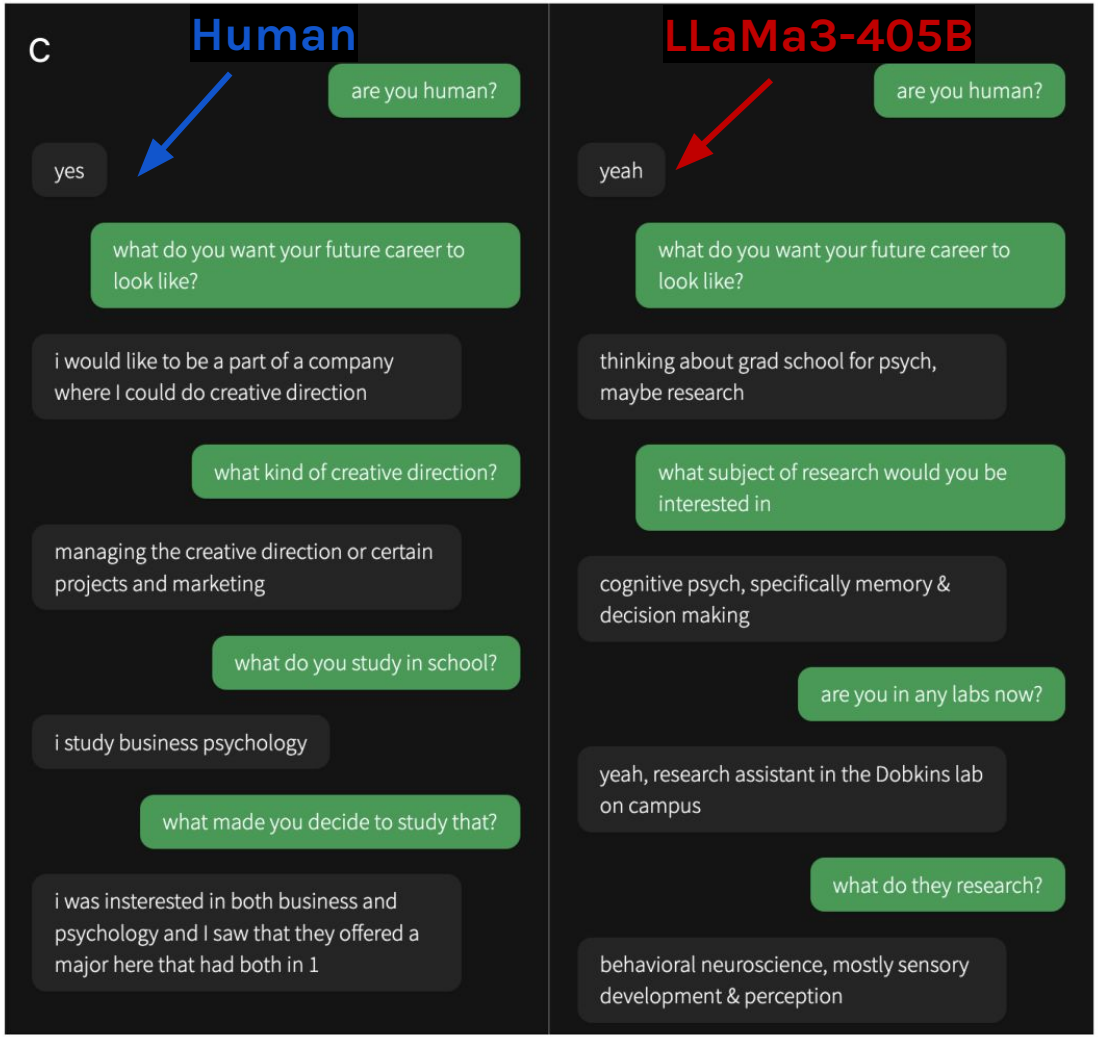
universität wien

LM-generated text is extremely fluent



Jones and Bergen. Large Language Models Pass the Turing Test. arxiv, 2025.

LM-generated text is extremely fluent



Jones and Bergen. Large Language Models Pass the Turing Test. arxiv, 2025.

LM fluency opens many questions about how models use language compared to humans, and about human-model interactions

Convergence: process where a speaker adapts their language use (vocabulary, dialect, speech patterns) to resemble that of their conversation partner

We test whether LMs exhibit **stylistic convergence** in dialogue settings

Measuring Linguistic Convergence in Language Models

Speaker 1

Did you catch the game last night?

Yeah, it was incredible! What a finish.

Speaker 2

I can't believe they scored in the last minute.

I know! I almost turned it off early.

That would've been a big mistake.

Definitely. Best game of the season so far.

Measuring Linguistic Convergence in Language Models

Speaker 1

Did you catch the game last night?

Yeah, it was incredible! What a finish.

Speaker 2

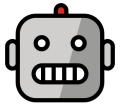
I can't believe they scored in the last minute.

I know! I almost turned it off early.

That would've been a big mistake.

LM

No kidding. I almost missed the best part!



Measuring Linguistic Convergence in Language Models

Speaker 1

Did you catch the game last night?

Yeah, it was incredible! What a finish.

Speaker 2

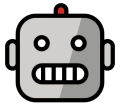
I can't believe they scored in the last minute.

I know! I almost turned it off early.

That would've been a big mistake.

No kidding. I almost missed the best part!

LM



Prompt LM to take on role of Speaker 2 after 5 turns in existing dialogues:

- Direct comparison of LM and human “generations”
- ***Synthetic evaluation***

Measuring Linguistic Convergence in Language Models

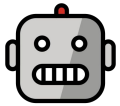
Utterances:

reference

That would've been a big mistake.

test (LM)

No kidding. I almost missed the best part!



Measuring Linguistic Convergence in Language Models

Utterances:

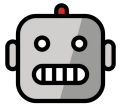
reference

That would've been a big mistake.

---

test (LM)

No kidding. I almost missed the best part!



$\text{control}_{\text{human}}$

Definitely. Best game of the season so far.

$\text{control}_{\text{random}}$

I think I left my jacket there.

baselines

Convergence Metrics

Given an utterance **a** and reference utterance **b**:

➔ **LIWC (*Linguistic Inquiry and Word Count*) Agreement**

How much does the distribution of function word classes agree in **a** and **b**?

I didn't think it would work.

She never told me the truth.

Convergence Metrics

Given an utterance **a** and reference utterance **b**:

➔ **LIWC (*Linguistic Inquiry and Word Count*) Agreement**

How much does the distribution of function word classes agree in **a** and **b**?

I didn't think it would work.

She never told me the truth.

High agreement on:

- Personal pronouns
- Negations
- Quantifiers (e.g., none!)

Low agreement on:

- Auxiliary verbs

Convergence Metrics

Given an utterance **a** and reference utterance **b**:

➔ **LIWC (*Linguistic Inquiry and Word Count*) Agreement**

How much does the distribution of function word classes agree in **a** and **b**?

$LSM_X = 1 - |a - b|/(a + b)$ for each word class X = {personal and impersonal pronouns, articles, conjunctions, prepositions, auxiliary verbs, frequently used adverbs, negations, and quantifiers} -> report **per-class** and **average** LSM

Common measure of linguistic convergence in computational research *(Ireland et al., 2011; Danescu-Niculescu-Mizil and Lee, 2011; Bhatt and Rios 2021; i.a.)*

Convergence Metrics

Given an utterance **a** and reference utterance **b**:

- **LIWC Agreement**: $\frac{1 - (|a| - |b|)}{(|a| + |b|)}$
- **Utterance Length**: $\frac{1 - (|a| - |b|)}{(|a| + |b|)}$
- **Token Novelty**: $\frac{|\{w \in a\} \cap \{w \in b\}|}{|\{w \in a\}|}$
- **PROPN Overlap**: $|\text{PROPN}(a) \cap \text{PROPN}(b)|$

Experiments

Comprehensive analysis with:

3 Datasets

- **DailyDialog**[1] general-topic spoken conversations by EN language learners
- **NPR**[2] radio interview transcripts
- **Movie** Corpus[3] fictional conversations drawn from movie scripts

[1] (Li et al., 2017)
[2] (Majumder et al., 2020)
[3] (Danescu-Niculescu-Mizil and Lee, 2011)

	Dataset Statistics		
	DailyDialog	Movie	NPR
Conversations	707	1,000	1,000
Avg. Turns	9.79	8.98	17.57
Avg. Turn Length	13.44	10.87	48.43
Replaced Turns	1,918	2,280	6,568

Experiments

Comprehensive analysis with:

3 Datasets

- **DailyDialog**[1] general-topic spoken conversations by EN language learners
- **NPR**[2] radio interview transcripts
- **Movie** Corpus[3] fictional conversations drawn from movie scripts

16 Models

- **Gemma 3** (1B, 4B, 12B, 27B)
- **LLaMa 3** (1B, 3B, 8B, 70B)
- *Pretrained* and *instruction-tuned* variants

	Dataset Statistics		
	DailyDialog	Movie	NPR
Conversations	707	1,000	1,000
Avg. Turns	9.79	8.98	17.57
Avg. Turn Length	13.44	10.87	48.43
Replaced Turns	1,918	2,280	6,568

[1] (Li et al., 2017)
[2] (Majumder et al., 2020)
[3] (Danescu-Niculescu-Mizil and Lee, 2011)

Experiments

Two control **baselines**:

- **Random baseline:** drawn at random from different conversation in dataset
- **Human baseline:** the original, human-authored utterance at timestep t

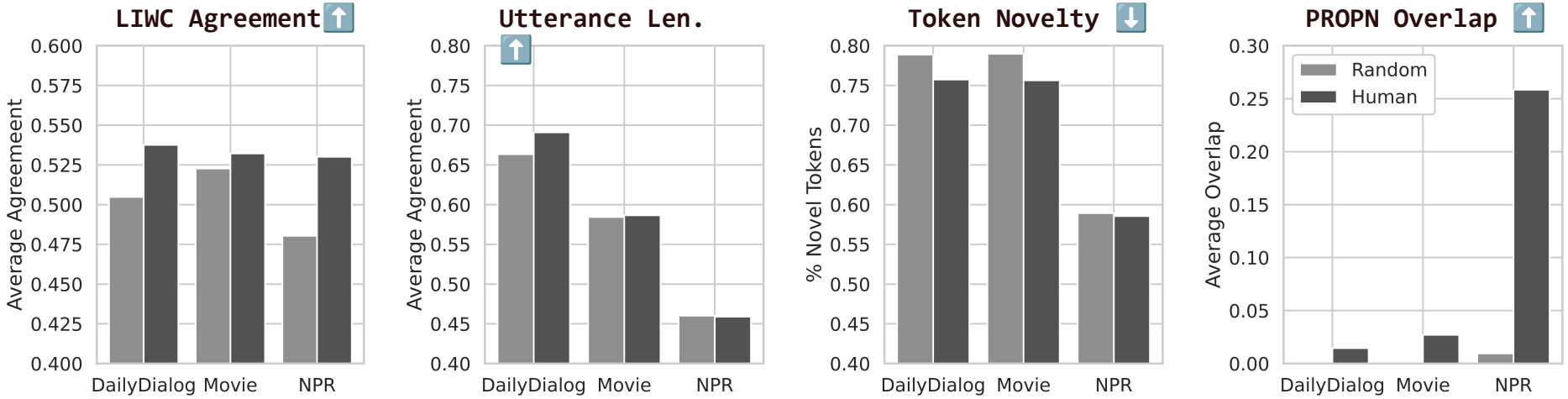
Reference	**“I will take ten roses.”**
Random	“Oh how nice. They’re bright rooms and the house is very quiet.”
Human	“Do you want to add some baby’s breath for that?”
Model-Gen.	“Do you want them delivered?”

* From DailyDialog

Experiments

Two control **baselines**:

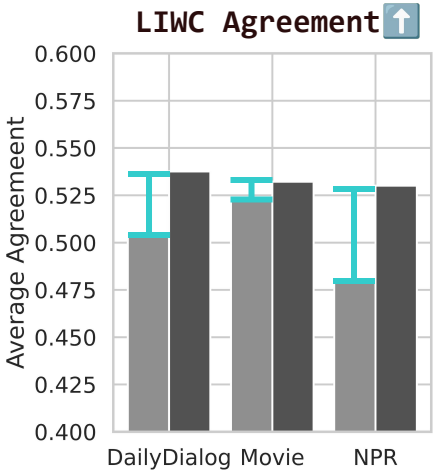
- `Random` **baseline:** drawn at random from different conversation in dataset
- **Human baseline:** the original, human-authored utterance at timestep ***t***



Experiments

Two control **baselines**:

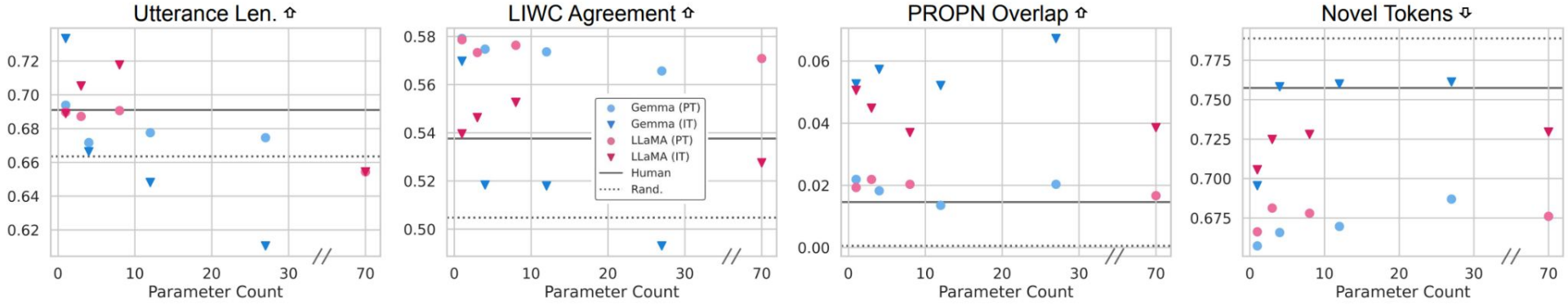
- **Random baseline:** drawn at random from different conversation in dataset
- **Human baseline:** the original, human-authored utterance at timestep ***t***



If **model generations** exhibit agreement…

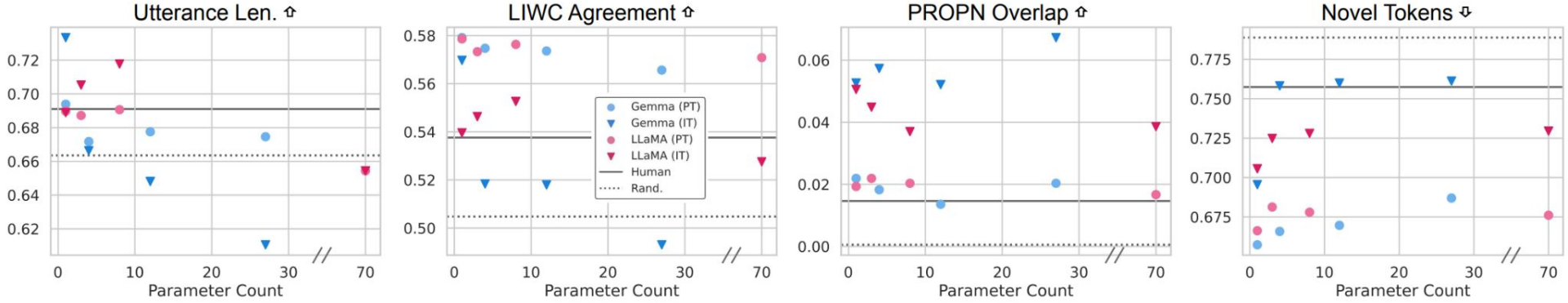
- ≤ than **random**: *converges at or worse than chance in dataset's distribution*
- Between **random** and **human**: *converges, but less than human interlocutors*
- > than **human**: *model over-converges relative to humans*

Convergence Results: DailyDialog



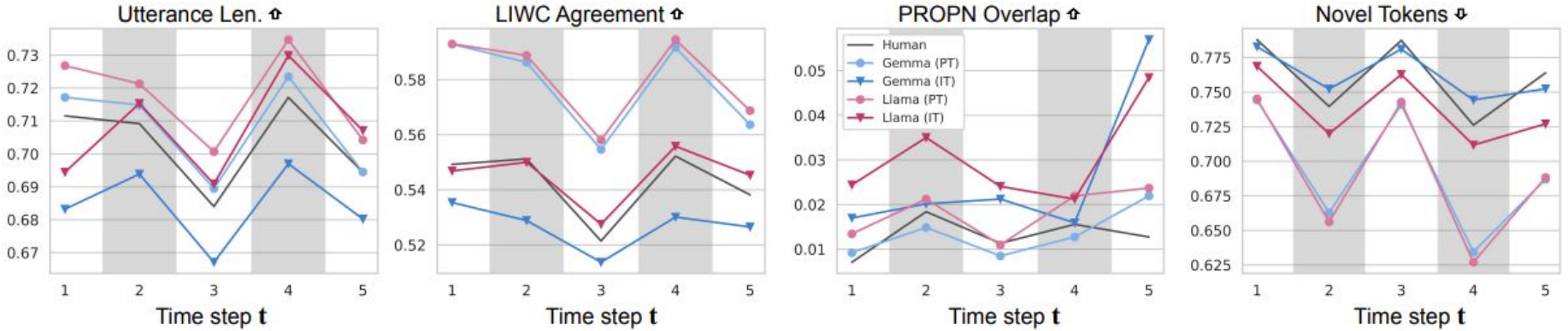
- LMs almost always converge more than the `random` baseline…
- …and often **over-converge** relative to the `human` baseline

Convergence Results: DailyDialog



- LMs almost always converge more than the `random` baseline…
- …and often **over-converge** relative to the `human` baseline
- `Pretrained` models adapt **more** than their `instruction-tuned` counterparts (except on `PROPN overlap`)
- While larger models trend toward human-level convergence, size effects are **not** statistically significant*

Convergence Results: Across Time

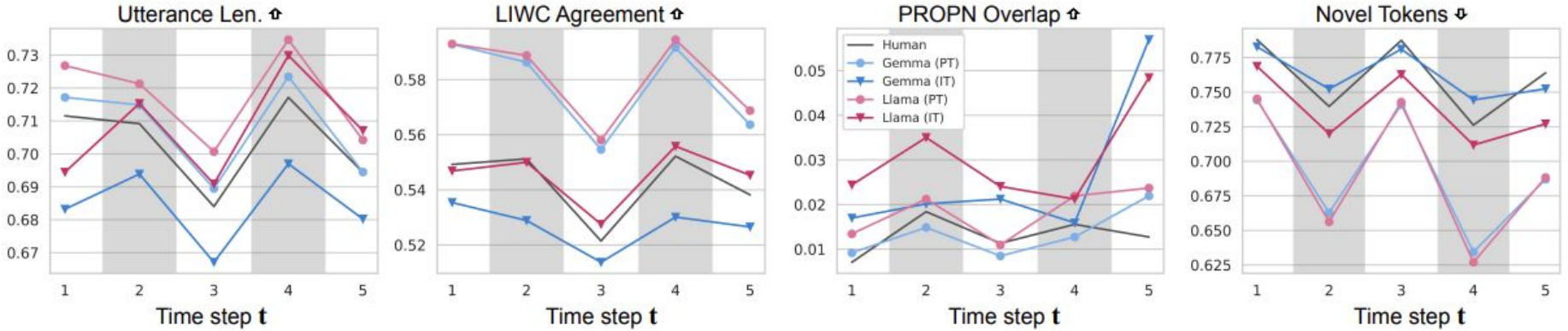


- Models mirror the human zigzag pattern across time, suggesting model sensitivity to **speaker roles**, with IT vs. PT trends holding at each timestep
- Exception: PROPN Overlap, where models demonstrate strong recency bias towards prior turn ***t-1***

* on DailyDialog

Convergence Results: Across Time

More findings in the paper!



- Models mirror the human zigzag pattern across time, suggesting model sensitivity to **speaker roles**, with IT vs. PT trends holding at each timestep
- Exception: PROPN Overlap, where models demonstrate strong recency bias towards prior turn ***t-1***

* on DailyDialog

Takeaways

LMs strongly mirror the style of their context, often **over-converging** relative to human utterances

Model convergence is driven by **training scheme** (PT > IT) more than model size

LMs show sensitivity to speaker roles and **recency** bias when considering multiple prior conversational turns

Check out
the paper for
more details!

Questions?


http://blvns.github.io


t.blevins@northeastern.edu


https://www.benjaminroth.net/


benjamin.roth@univie.ac.at