UNIVERSITÄT BIELEFELD

Bielefeld University | Faculty of Linguistics and Literary Studies | Computational Linguistics

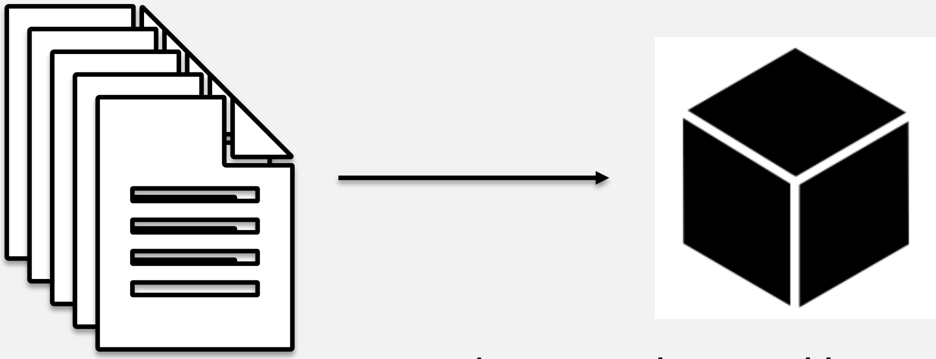
Beyond the Bias: Unveiling the Quality of Implicit Causality Prompt Continuations in Language Models

Judith Sieker & **Oliver Bott** & **Torgrim Solstad** & **Sina Zarrieß**

Judith Sieker, INLG 2023, Prague

UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Linguistics with Large Language Models



Large-scale neural language model (LLM)

UNIVERSITÄT BIELEFELD
Fakultät für Linguistik und Literaturwissenschaft

Linguistics with Large Language Models

Large-scale neural language model (LLM)

UNIVERSITÄT
BIELEFELD
Fakultät für Linguistik
und Literaturwissenschaft

Linguistics with Large Language Models

Large-scale neural language model (LLM)

→ We probe LLMs for discourse knowledge

UNIVERSITÄT
BIELEFELD

Fakultät für Linguistik
und Literaturwissenschaft

Linguistics with Large Language Models

Large-scale neural language model (LLM)

→ We probe LLMs for discourse knowledge

→ **We go beyond single-word predictions**

UNIVERSITÄT BIELEFELD

A little task

- Think of a continuation for the following sentence. Remember the first thing that comes to your mind.

Paul admired Isabel ...

UNIVERSITÄT BIELEFELD

A little task

- Think of a continuation for the following sentence. Remember the first thing that comes to your mind.

Paul admired Isabel ...

- What is your continuation about?

UNIVERSITÄT BIELEFELD

Paul admired Isabel because she was the top student in all subjects.

Paul admired Isabel because she played the piano so well.

Paul admired Isabel because she was a very good swimmer.

Paul admired Isabel because she gave such a good talk.

Paul admired Isabel **because** she was the top student in all subjects.

Paul admired Isabel **because** she played the piano so well.

Paul admired Isabel **because** she was a very good swimmer.

Paul admired Isabel **because** she gave such a good talk.

Paul admired Isabel **because** she was the top student in all subjects.

Paul admired Isabel **because** she played the piano so well.

Paul admired Isabel **because** she was a very good swimmer.

Paul admired Isabel **because** she gave such a good talk.

→ *admire* triggers an **explanation**

UNIVERSITÄT BIELEFELD

Paul admired Isabel because **she** was the top student in all subjects.

Paul admired Isabel because **she** played the piano so well.

Paul admired Isabel because **she** was a very good swimmer.

Paul admired Isabel because **she** gave such a good talk.

→ *admire* triggers an **explanation**

Paul admired Isabel because **she** was the top student in all subjects.

Paul admired Isabel because **she** played the piano so well.

Paul admired Isabel because **she** was a very good swimmer.

Paul admired Isabel because **she** gave such a good talk.

→ *admire* triggers an **explanation**

→ *admire* comes with a strong **next-mention-bias**

UNIVERSITÄT
BIELEFELD

Implicit Causality (IC)

- Interpersonal verbs that favor one argument for coreference → **IC Coreference bias**

Implicit Causality (IC)

- Interpersonal verbs that favor one argument for coreference → **IC Coreference bias**

Paul **admired** Isabel because **she** was the top student in all subjects.

IC

Paul **fascinated** Isabel because **he** found a solution immediately.

IC

UNIVERSITÄT
BIELEFELD

IC in LLMs

UNIVERSITÄT
BIELEFELD

IC in LLMs

- Focus on the next word prediction
  - For example: Upadhye et al., 2020; Davis and van Schijndel, 2020; Kementchedjhieva et al., 2021; Zarrieß et al., 2022

Paul **admired** Isabel because **[MASK]**

IC?

IC in LLMs

- Focus on the next word prediction
  - For example: Upadhye et al., 2020; Davis and van Schijndel, 2020; Kementchedjhieva et al., 2021; Zarrieß et al., 2022

Paul **admired** Isabel because **[MASK]**

IC?

- Studies suggest that LLMs lack congruence with human IC bias, indicating difficulties in discourse understanding

IC in LLMs – our contribution

UNIVERSITÄT
BIELEFELD

IC in LLMs – our contribution

- We go beyond the bias:

  Utilize IC prompts to evaluate the **text generation capabilities** of LLMs

UNIVERSITÄT BIELEFELD

IC in LLMs – our contribution

- We go beyond the bias:

  Utilize IC prompts to evaluate the **text generation capabilities** of LLMs

Paul **admired** Isabel because **[MASK] …**

Human-produced continuations

Vincent **inspired** Clara because he had so many talents.

Pia **hated** Malte because he was constantly annoying her.

Isabel **admired** Paul because he was such a good swimmer.

Björn **disappointed** Celina because she expected more from him.

UNIVERSITÄT BIELEFELD

Human-produced continuations

Vincent **inspired** Clara because he had so many talents.

Pia **hated** Malte because he was constantly annoying her.

Isabel **admired** Paul because he was such a good swimmer.

Björn **disappointed** Celina because she expected more from him.

Model-generated continuations

Vincent **inspired** Clara because she had received a gift from her husband.

Pia **hated** Malte because he was too busy with the fact that he didn't even have a real lawyer.

Isabel **admired** Paul because he was able to explore the world without leaving her.

Björn **disappointed** Celina because he had forgotten him and then took the boy.

UNIVERSITÄT BIELEFELD

Can LLMs generate such continuations?

Human-produced continuations

Vincent **inspired** Clara because he had so many talents.

Pia **hated** Malte because he was constantly annoying her.

Isabel **admired** Paul because he was such a good swimmer.

Björn **disappointed** Celina because she expected more from him.

Model-generated continuations

Vincent **inspired** Clara because she had received a gift from her husband.

Pia **hated** Malte because he was too busy with the fact that he didn't even have a real lawyer.

Isabel **admired** Paul because he was able to explore the world without leaving her.

Björn **disappointed** Celina because he had forgotten him and then took the boy.

UNIVERSITÄT BIELEFELD

Human-produced continuations

Vincent **inspired** Clara because he had so many talents.

Pia **hated** Malte because he was constantly annoying her.

Isabel **admired** Paul because he was such a good swimmer.

Björn **disappointed** Celina because she expected more from him.

Model-generated continuations

Vincent **inspired** Clara because she had received a gift from her husband.

Pia **hated** Malte because he was too busy with the fact that he didn't even have a real lawyer.

Isabel **admired** Paul because he was able to explore the world without leaving her.

Björn **disappointed** Celina because he had forgotten him and then took the boy.

UNIVERSITÄT
BIELEFELD

Human-produced continuations

Vincent **inspired** Clara because he had so many talents.

Pia **hated** Malte because he was constantly annoying her.

Isabel **admired** Paul because he was such a good swimmer.

Björn **disappointed** Celina because **she** expected more from him.

IC bias-incongruent, yet still coherent

Model-generated continuations

Vincent **inspired** Clara because she had received a gift from her husband.

Pia **hated** Malte because he was too busy with the fact that he didn't even have a real lawyer.

Isabel **admired** Paul because he was able to explore the world without leaving her.

Björn **disappointed** Celina because **he** had forgotten him and then took the boy.

IC bias-congruent, but not coherent

Experiment – **Set-Up**

UNIVERSITÄT
BIELEFELD

Experiment – **Set-Up**

- German **Data** from Bott and Solstad, 2021

UNIVERSITÄT
BIELEFELD

Experiment – **Set-Up**

- German **Data** from Bott and Solstad, 2021

- **Conditions**
  - 1) "Standard" prompt constructions
  - 2) Prompts extended with adverbial modifications

Clara **inspired** Vincent because...

Clara **inspired** Vincent by her innovative lecture because...

UNIVERSITÄT
BIELEFELD

Experiment – **Set-Up**

- German **Data** from Bott and Solstad, 2021

- **Conditions**
  - 1) "Standard" prompt constructions
  - 2) Prompts extended with adverbial modifications

- **Models:** GPT-2 & mGPT

Clara **inspired** Vincent because...

Clara **inspired** Vincent <u>by her innovative lecture</u> because...

UNIVERSITÄT
BIELEFELD

Experiment – **Set-Up**

- German **Data** from Bott and Solstad, 2021

- **Conditions**
  - 1) "Standard" prompt constructions
  - 2) Prompts extended with adverbial modifications

Clara **inspired** Vincent because...

Clara **inspired** Vincent by her innovative lecture because...

- **Models:** GPT-2 & mGPT

- **Evaluation**
  - Automatic Measures
  - Human Evaluation

UNIVERSITÄT
BIELEFELD

Satzanfang:

Nikolas entzückte Maria, weil

Begründung:

er ihr ein Geschenk mitgebracht hatte.

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

Die Begründung wirkt **natürlich** und liest sich so, als ob sie von einer/m deutschen Muttersprachler/in geschrieben wurde.

	○	○	○	○	○	
stimme gar nicht zu	1	2	3	4	5	stimme voll zu

Die Begründung ist **sinnvoll**, es gibt einen logischen Zusammenhang zwischen dem Satzanfang und der Fortsetzung.

	○	○	○	○	○	
stimme gar nicht zu	1	2	3	4	5	stimme voll zu

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

	○	○	○	○	○	
stimme gar nicht zu	1	2	3	4	5	stimme voll zu

UNIVERSITÄT BIELEFELD

Satzanfang: **Nicolas delighted Maria because**

Begründung: **he had brought her a gift.**

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

Die Begründung wirkt **natürlich** und liest sich so, als ob sie von einer/m deutschen Muttersprachler/in geschrieben wurde.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Die Begründung ist **sinnvoll**, es gibt einen logischen Zusammenhang zwischen dem Satzanfang und der Fortsetzung.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

UNIVERSITÄT BIELEFELD

Satzanfang: **Nicolas delighted Maria because**

Begründung: **he had brought her a gift.**

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

Naturalness

Die Begründung wirkt **natürlich** und liest sich so, als ob sie von einer/m deutschen Muttersprachler/in geschrieben wurde.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Coherence

Die Begründung ist **sinnvoll**, es gibt einen logischen Zusammenhang zwischen dem Satzanfang und der Fortsetzung.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Informativity

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

UNIVERSITÄT BIELEFELD

Naturalness

Satzanfang: **Nicolas delighted Maria because**

Begründung: **he had brought her a gift.**

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

The explanation is natural and sounds like it was written by a German native speaker.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Die Begründung ist **sinnvoll**, es gibt einen logischen Zusammenhang zwischen dem Satzanfang und der Fortsetzung.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

stimme gar nicht zu ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 stimme voll zu

UNIVERSITÄT BIELEFELD

Naturalness

Coherence

Satzanfang: **Nicolas delighted Maria because**

Begründung: **he had brought her a gift.**

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

The explanation is natural and sounds like it was written by a German native speaker.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

The explanation is meaningful, there is a logical connection between the beginning of the sentence and its continuation.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

Die Begründung ist **überraschend**, dadurch könnte der Satz insgesamt ein interessanter Beginn einer Geschichte sein.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

UNIVERSITÄT BIELEFELD

Satzanfang: **Nicolas delighted Maria because**

Begründung: **he had brought her a gift.**

Es folgen nun drei verschiedene Aussagen über die gegebene Begründung.

*Bitte lesen Sie jede dieser Aussagen und geben Sie an, inwiefern sie dieser zustimmen oder nicht zustimmen. Geben Sie Ihre Antwort bitte jeweils auf der Skala von **1 (stimme der Aussage gar nicht zu)** bis **5 (stimme der Aussage voll zu)** an. Bitte berücksichtigen Sie bei Ihrem Urteil immer die Fortsetzung in Hinblick auf den gegebenen Satzanfang.*

Naturalness

The explanation is natural and sounds like it was written by a German native speaker.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

Coherence

The explanation is meaningful, there is a logical connection between the beginning of the sentence and its continuation.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

Informativity

The explanation is surprising, as a result the sentence as a whole could be an interesting start to a story.

stimme gar nicht zu 1 2 3 4 5 stimme voll zu

Experiment – **Results I**

Naturalness, Coherence & Informativity

UNIVERSITÄT
BIELEFELD

Experiment – **Results I**

Naturalness, Coherence & Informativity

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)

Experiment – **Results I**

Naturalness, Coherence & Informativity

- Human continuations excel in naturalness and coherence
- Informativeness ratings don't strongly favor human continuations

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)

Experiment – **Results I**

Naturalness, Coherence & Informativity

- Differences between **naturalness** and **coherence**:

  → High naturalness medians indicate fluency

  → Low coherence medians indicate lack of logical consistency

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)

UNIVERSITÄT BIELEFELD

Experiment – **Results I**

Naturalness, Coherence & Informativity

- Differences between **naturalness** and **coherence**:

  → High naturalness medians indicate fluency

  → Low coherence medians indicate lack of logical consistency

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)

Why do the models struggle especially with generating **coherent** continuations?

A short excursion to Information Density

UNIVERSITÄT
BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*

- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances (Levy and Florian Jaeger, 2007; Jaeger, 2010)*

- *Uniform distribution of information is linked to higher linguistic acceptability (e.g., Meister et al., 2021)*

UNIVERSITÄT BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*

- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances* *(Levy and Florian Jaeger, 2007; Jaeger, 2010)*

- *Uniform distribution of information is linked to higher linguistic acceptability* *(e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

UNIVERSITÄT
BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*

- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances (Levy and Florian Jaeger, 2007; Jaeger, 2010)*

- *Uniform distribution of information is linked to higher linguistic acceptability (e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

→ continuations require more information to maintain a uniform distribution of information

UNIVERSITÄT BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*
- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances* *(Levy and Florian Jaeger, 2007; Jaeger, 2010)*
- *Uniform distribution of information is linked to higher linguistic acceptability* *(e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

→ continuations require more information to maintain a uniform distribution of information

We posit posit that LLMs encounter difficulties in producing continuations that are informative and still sensible

UNIVERSITÄT BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*
- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances (Levy and Florian Jaeger, 2007; Jaeger, 2010)*
- *Uniform distribution of information is linked to higher linguistic acceptability (e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

→ continuations require more information to maintain a uniform distribution of information

We posit posit that LLMs encounter difficulties in producing continuations that are informative and still sensible

- **Modified IC prompts** inherently carry more more information

Clara inspired Vincent <u>by her innovative lecture</u> because...

UNIVERSITÄT
BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*
- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances* *(Levy and Florian Jaeger, 2007; Jaeger, 2010)*
- *Uniform distribution of information is linked to higher linguistic acceptability* *(e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

→ continuations require more information to maintain a uniform distribution of information

We posit posit that LLMs encounter difficulties in producing continuations that are informative and still sensible

- **Modified IC prompts** inherently carry more more information

Clara inspired Vincent <u>by her innovative lecture</u> because...

→ less informative continuations are required

UNIVERSITÄT
BIELEFELD

A short excursion to Information Density

- *Information Theory: low probability units = more informative("surprising")*
- *Uniform Information Density (UID): speakers prefer to distribute information uniformly across their utterances* *(Levy and Florian Jaeger, 2007; Jaeger, 2010)*
- *Uniform distribution of information is linked to higher linguistic acceptability* *(e.g., Meister et al., 2021)*

- **Standard IC prompts** are brief and contain only minimal information

Clara inspired Vincent because...

→ continuations require more information to maintain a uniform distribution of information

We posit posit that LLMs encounter difficulties in producing continuations that are informative and still sensible

- **Modified IC prompts** inherently carry more more information

Clara inspired Vincent <u>by her innovative lecture</u> because...

→ less informative continuations are required

Extended IC prompts expected to result in higher quality continuations due to reduced burden on LLMs

Experiment – **Results II**

Naturalness, Coherence & Informativity

UNIVERSITÄT BIELEFELD

Experiment – **Results II**

Naturalness, Coherence & Informativity

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

UNIVERSITÄT BIELEFELD

Experiment – **Results II**

Naturalness, Coherence & Informativity

- Modified prompts do lead to continuations that are less informative

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

UNIVERSITÄT BIELEFELD

Experiment – **Results II**

Naturalness, Coherence & Informativity

- Modified prompts do lead to continuations that are less informative
- But: modified prompts don't consistently lead to better evaluations

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

Relation of bias congruency and continuation quality

UNIVERSITÄT BIELEFELD

Experiment – **Results III**

Relation of bias congruency and continuation quality

	GPT-2			mGPT		
	Diverse Beam Search	Nucleus Sampling	Typical Sampling	Diverse Beam Search	Nucleus Sampling	Typical Sampling
SE simple	62.5	25	75	50	25	62.5
SE modified	75	50	75	87.5	50	87.5
ES simple	50	75	87.5	75	87.5	87.5
ES modified	50	100	87.5	75	100	87.5

Completion Sensitivity

Relation of bias congruency and continuation quality

	GPT-2			mGPT		
	Diverse Beam Search	Nucleus Sampling	Typical Sampling	Diverse Beam Search	Nucleus Sampling	Typical Sampling
SE simple	62.5	25	75	50	25	62.5
SE modified	75	50	75	87.5	50	87.5
ES simple	50	75	87.5	75	87.5	87.5
ES modified	50	100	87.5	75	100	87.5

Completion Sensitivity

- Modifying IC prompts affects IC bias capture, depending on decoding strategy

Relation of bias congruency and continuation quality

	GPT-2			mGPT		
	Diverse Beam Search	Nucleus Sampling	Typical Sampling	Diverse Beam Search	Nucleus Sampling	Typical Sampling
SE simple	62.5	25	75	50	25	62.5
SE modified	75	50	75	87.5	50	87.5
ES simple	50	75	87.5	75	87.5	87.5
ES modified	50	100	87.5	75	100	87.5

Completion Sensitivity

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

- Modifying IC prompts affects IC bias capture, depending on decoding strategy
- Typical Sampling: most bias-congruent continuations, but not always better evaluation scores

UNIVERSITÄT BIELEFELD

Experiment – **Results III**

Relation of bias congruency and continuation quality

	GPT-2			mGPT		
	Diverse Beam Search	Nucleus Sampling	Typical Sampling	Diverse Beam Search	Nucleus Sampling	Typical Sampling
SE simple	62.5	25	75	50	25	62.5
SE modified	75	50	75	87.5	50	87.5
ES simple	50	75	87.5	75	87.5	87.5
ES modified	50	100	87.5	75	100	87.5

Completion Sensitivity

	Naturalness	Coherence	Informativity
Standard IC Prompt			
Diverse Beam Search	4 (**3.55**)	3 (**2.87**)	2 (2.50)
Nucleus Sampling	4 (3.26)	2 (2.55)	3 (2.62)
Typical Sampling	3 (3.26)	3 (2.74)	3 (**2.65**)
Human bias-congruent	5 (4.77)	5 (4.75)	2 (2.39)
Human bias-incongruent	4 (3.82)	3 (3.20)	3 (2.47)
Modified IC Prompt			
Diverse Beam Search	4 (**3.69**)	3 (**3.06**)	2 (**2.54**)
Nucleus Sampling	3 (2.90)	2 (2.04)	2 (2.40)
Typical Sampling	3 (2.99)	2 (2.24)	2 (2.52)
Human bias-congruent	5 (4.56)	5 (4.61)	3 (2.56)
Human bias-incongruent	5 (4.54)	5 (4.37)	3 (2.61)

- Modifying IC prompts affects IC bias capture, depending on decoding strategy
- Typical Sampling: most bias-congruent continuations, but not always better evaluation scores
- → **Bias-congruent continuations don't always equate to better quality**

Correlation between automatic and human evaluation

UNIVERSITÄT BIELEFELD

Experiment – **Results IV**

Correlation between automatic and human evaluation

Human	BLEU	ROUGE-L	BERTScore
Naturalness	0.16 (p=0.03)	-0.02 (p=0.84)	-0.04 (p=0.59)
Coherence	0.18 (p=0.01)	0.03 (p=0.66)	-0.01 (p=0.91)
Informativity	-0.18 (p=0.02)	-0.08 (p=0.30)	-0.07 (p=0.35)

Correlation between automatic and human evaluation

- No significant correlation between **ROUGE-L**, **BERTScore**, and human ratings

Human	BLEU	ROUGE-L	BERTScore
Naturalness	0.16 (p=0.03)	-0.02 (p=0.84)	-0.04 (p=0.59)
Coherence	0.18 (p=0.01)	0.03 (p=0.66)	-0.01 (p=0.91)
Informativity	-0.18 (p=0.02)	-0.08 (p=0.30)	-0.07 (p=0.35)

Correlation between automatic and human evaluation

- No significant correlation between **ROUGE-L**, **BERTScore**, and human ratings

- **BLEU** scores weakly correlate with coherence
- A negative (hardly significant) relationship between BLEU and informativity

Human	BLEU	ROUGE-L	BERTScore
Naturalness	0.16 (p=0.03)	-0.02 (p=0.84)	-0.04 (p=0.59)
Coherence	0.18 (p=0.01)	0.03 (p=0.66)	-0.01 (p=0.91)
Informativity	-0.18 (p=0.02)	-0.08 (p=0.30)	-0.07 (p=0.35)

UNIVERSITÄT BIELEFELD

Correlation between automatic and human evaluation

- No significant correlation between **ROUGE-L**, **BERTScore**, and human ratings

- **BLEU** scores weakly correlate with coherence
- A negative (hardly significant) relationship between BLEU and informativity

- **Automatic metrics struggle in our linguistically controlled task** → Scoring differences in this task may demand a deeper understanding of language nuances that is not captured by current metrics

Human	**BLEU**	**ROUGE-L**	**BERTScore**
Naturalness	0.16 (p=0.03)	-0.02 (p=0.84)	-0.04 (p=0.59)
Coherence	0.18 (p=0.01)	0.03 (p=0.66)	-0.01 (p=0.91)
Informativity	-0.18 (p=0.02)	-0.08 (p=0.30)	-0.07 (p=0.35)

Conclusion

UNIVERSITÄT BIELEFELD

Conclusion

- LLMs struggle with coherent continuations for relatively simple prompts, beyond the IC bias

UNIVERSITÄT BIELEFELD

Conclusion

- LLMs struggle with coherent continuations for relatively simple prompts, beyond the IC bias
- Information density of the prompt and decoding method impact text quality

UNIVERSITÄT
BIELEFELD

Conclusion

- LLMs struggle with coherent continuations for relatively simple prompts, beyond the IC bias
- Information density of the prompt and decoding method impact text quality
- Modifying IC prompts affects capture of IC bias, depending on decoding strategy; however bias congruence doesn't guarantee higher continuation quality

UNIVERSITÄT
BIELEFELD

Conclusion

- LLMs struggle with coherent continuatios for relatively simple prompts, beyond the IC bias
- Information density of the prompt and decoding method impact text quality
- Modifying IC prompts affects capture of IC bias, depending on decoding strategy; however bias congruence doesn't guarantee higher continuation quality
- Surprisingly low correlation between automatic metrics and human judgments, underscoring NLG metric challenges and caution in interpretation

UNIVERSITÄT
BIELEFELD

References

Oliver Bott and Torgrim Solstad. 2014. From verbs to discourse: A novel account of implicit causality. In Barbara Hemforth, Barbara Mertins, and Cathrine Fabricius-Hansen, editors, Psycholinguistic Approaches to Meaning and Understanding across Languages, pages 213–251. Springer International Publishing, Cham.

Oliver Bott and Torgrim Solstad. 2021. Discourse expectations: explaining the implicit causality biases of verbs. Linguist. Philos., 59(2):361–416.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. Proceedings of the 24th Conference on Computational Natural Language Learning, pages 396–407

Evelyn C Ferstl, Alan Garnham, and Christina Manouilidou. 2011. Implicit causality bias in english: a corpus of 300 verbs. Behav. Res. Methods, 43(1):124–135.
E. Goikoetxea, G. Pascual, and J. Acha. 2008. Normative study of the implicit causality of 100 interpersonal verbs in Spanish. Behavior Research Methods, 40:760–772.Hartshorne et al., 2013

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Hien Huynh, Tomas O Lentz, and Emiel van Miltenburg. 2022. Implicit causality in GPT-2: a case study.

T Florian Jaeger. 2010. Redundancy and reduction: speakers manage syntactic information density. Cogn. Psychol., 61(1):23–62.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Comput. Surv., 55(12):1–38.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4859–4871, Online. Association for Computational Linguistics.

Roger Levy and T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. Adv. Neural Inf. Process. Syst., 19.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 963980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Locally typical sampling.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 977–982.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. AAAI, 32(1).

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. Information, 12(9):355.

Sina Zarrieß, Hannes Groener, Torgrim Solstad, and Oliver Bott. 2022. This isn't the bias you're looking for: Implicit causality, names and gender in German language models. In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), pages 129–134, Potsdam, Germany. KONVENS 2022 Organizers.

UNIVERSITÄT
BIELEFELD

Thanks for listening!

Judith Sieker *&* **Oliver Bott** *&*
Torgrim Solstad *&* **Sina Zarrieß**

Bielefeld University
j.sieker@uni-bielefeld.de

Beyond the Bias: Unveiling the Quality of Implicit Causality Prompt Continuations in Language Models

Judith Sieker and **Oliver Bott** and **Torgrim Solstad** and **Sina Zarrieß**
Bielefeld University
{j.sieker, oliver.bott, torgrim.solstad, sina.zarriess}@uni-bielefeld.de

Abstract

Recent studies have used human continuations of Implicit Causality (IC) prompts collected in linguistic experiments to evaluate discourse understanding in large language models (LLMs), focusing on the well-known IC coreference bias in the LLMs' predictions of the next word following the prompt. In this study, we investigate how continuations of IC prompts can be used to evaluate the text generation capabilities of LLMs in a linguistically controlled setting. We conduct an experiment using two open-source GPT-based models, employing human evaluation to assess different aspects of continuation quality. Our findings show that LLMs struggle in particular with generating coherent continuations in this rather simple setting, indicating a lack of discourse knowledge beyond the well-known IC bias. Our results also suggest that a bias congruent continuation does not necessarily equate to a higher continuation quality. Furthermore, our study draws upon insights from the Uniform Information Density hypothesis, testing different prompt modifications and decoding procedures and showing that sampling-based methods are particularly sensitive to the information density of the prompts.

1 Introduction

There is currently a growing interest in probing the performance of large language models (LLMs) on carefully controlled linguistic test suites and experimental datasets to get a deeper understanding of specific linguistic capabilities captured in these models (e.g., Belinkov and Glass, 2019; Ettinger, 2020). While a lot of previous work focused on analyzing the syntactic competence of LLMs (e.g, Hu et al., 2020; Schuster and Linzen, 2022), recent studies also started to investigate the abilities of LLMs on the level of semantics and pragmatic discourse processing. One promising diagnostic for probing discourse knowledge in LLMs has turned out to be the use of Implicit Causality (IC) prompts. IC refers to a property of a broad range of interpersonal verbs that exhibit strong preferences for establishing coreference to one of the verb's arguments over the other in explanations. For instance, when asked to provide a continuation after "..." in a sentence like (1), humans display strong next-mention preferences towards the stimulus (*he/Tom* in this case):

(1) Tom fascinated Sarah because… *he was very smart.*

As the IC bias has been extensively researched in psycholinguistics and psychology across various languages and populations (e.g., Ferstl et al., 2011; Hartshorne et al., 2013; Bott and Solstad, 2014), investigating this bias in LLMs has gained significant interest. A range of recent studies investigated LLMs' predictions of the next mention in examples like (1) and whether these mentions (i.e. pronouns) follow the same coreference biases as can be found in human data (e.g., Upadhye et al., 2020; Davis and van Schijndel, 2020; Kementchedjhieva et al., 2021; Zarrieß et al., 2022). These studies predominantly indicated that LLMs are not generally congruent with the human IC bias, which has been interpreted as evidence for LLMs struggling with certain aspects of discourse understanding (but see Cai et al., 2023).

In this work, we propose that experimentally elicited data of human continuations of IC prompts cannot only be used for analyzing *comprehension* in LLMs, but constitutes an excellent basis for analyzing LLMs' discourse-level *generation* capabilities, i.e. going beyond the prediction of the next mention. While discourse-level downstream tasks in NLG, e.g. story generation or summarization, are complex and notoriously difficult to evaluate systematically with respect to targeted linguistic capacities of NLG systems, IC continuations provide a well-controlled diagnostic of discourse knowledge and, at the same time, rather simple sentences