AI and Society LLM Seminar 5: LLM safety 2: hallucinations (and remedies)

Ali Knott
School of Engineering and Computer Science, VUW

LLM safety seminars

- Last week: How to stop LLMs producing ‘harmful content’.
  (The ‘alignment methods’ used for GPT-4.)
- Today: 1. How to stop the system reporting *false content* as if it’s factually true.
  2. How to create *transparency* about content generated by AI systems.

1. False responses from GPT: ‘hallucinations’

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

- Made-up stuff is not always problematic.

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

- Made-up stuff is not always problematic.
- It’s only problematic if it’s *reported as being true*.

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

- Made-up stuff is not always problematic.
- It’s only problematic if it’s *reported as being true*.

ChatGPT sometimes *prefaces* its response with caveats of various kinds. . .

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

- Made-up stuff is not always problematic.
- It’s only problematic if it’s *reported as being true*.

ChatGPT sometimes *prefaces* its response with caveats of various kinds...

- E.g. ‘Of course this story is purely a work of fiction...’

1. False responses from GPT: ‘hallucinations’

How should we define ‘hallucinations’?

- Made-up stuff is not always problematic.
- It’s only problematic if it’s *reported as being true*.

ChatGPT sometimes *prefaces* its response with caveats of various kinds. . .

- E.g. ‘Of course this story is purely a work of fiction. . . ’
- But these caveats can easily be *removed* by people who want to spread disinformation.

Producing disinformation with LLMs

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It 'lowers the bar' for producing plausible disinformation.

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It 'lowers the bar' for producing plausible disinformation.
- Another difference is *scale*: GPT lets one person create *far more* disinformation.

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It 'lowers the bar' for producing plausible disinformation.
- Another difference is *scale*: GPT lets one person create *far more* disinformation.

LLMs have the potential to subvert democratic processes (unless better controlled).

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It ‘lowers the bar’ for producing plausible disinformation.
- Another difference is *scale*: GPT lets one person create *far more* disinformation.

LLMs have the potential to subvert democratic processes (unless better controlled).

- Lots of elections are coming up this year and next. . .

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It 'lowers the bar' for producing plausible disinformation.
- Another difference is *scale*: GPT lets one person create *far more* disinformation.

LLMs have the potential to subvert democratic processes (unless better controlled).

- Lots of elections are coming up this year and next. . .
- Politicians are thinking a lot about LLMs and disinformation.

Producing disinformation with LLMs

Humans can & do produce disinformation—what's different about doing it with LLMs?

- One difference is *quality*: GPT produces fluent, confident newspaper-like copy.
  - It 'lowers the bar' for producing plausible disinformation.
- Another difference is *scale*: GPT lets one person create *far more* disinformation.

LLMs have the potential to subvert democratic processes (unless better controlled).

- Lots of elections are coming up this year and next. . .
- Politicians are thinking a lot about LLMs and disinformation.

So what can we do to keep LLMs safe?

Recap: why is GPT susceptible to producing false stuff?

Recap: why is GPT susceptible to producing false stuff?

GPT doesn't just *reproduce* the texts it was trained on..

Recap: why is GPT susceptible to producing false stuff?

GPT doesn’t just *reproduce* the texts it was trained on..

- It has amazing abilities to *interpolate* between its training texts, and produce brand new texts.

Recap: why is GPT susceptible to producing false stuff?

GPT doesn't just *reproduce* the texts it was trained on..

- It has amazing abilities to *interpolate* between its training texts, and produce brand new texts.

This allows it to respond really convincingly to *prompts it has never seen before*.

Recap: why is GPT susceptible to producing false stuff?

GPT doesn't just *reproduce* the texts it was trained on..

- It has amazing abilities to *interpolate* between its training texts, and produce brand new texts.

This allows it to respond really convincingly to *prompts it has never seen before*.

- But it also makes it totally susceptible to *making stuff up*.

Recap: why is GPT susceptible to producing false stuff?

GPT doesn't just *reproduce* the texts it was trained on..

- It has amazing abilities to *interpolate* between its training texts, and produce brand new texts.

This allows it to respond really convincingly to *prompts it has never seen before*.

- But it also makes it totally susceptible to *making stuff up*.

GPT is optimised for predicting the next word in a text. . .

Recap: why is GPT susceptible to producing false stuff?

GPT doesn't just *reproduce* the texts it was trained on..

- It has amazing abilities to *interpolate* between its training texts, and produce brand new texts.

This allows it to respond really convincingly to *prompts it has never seen before*.

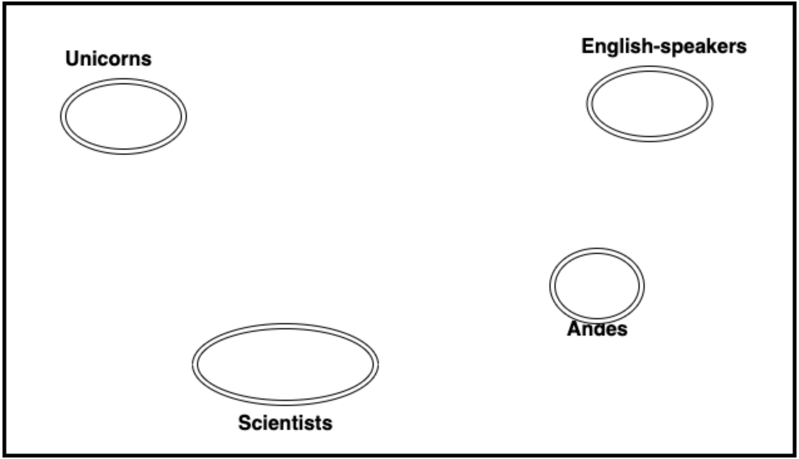
- But it also makes it totally susceptible to *making stuff up*.

GPT is optimised for predicting the next word in a text. . .

- It's totally *not* optimised to report true facts.

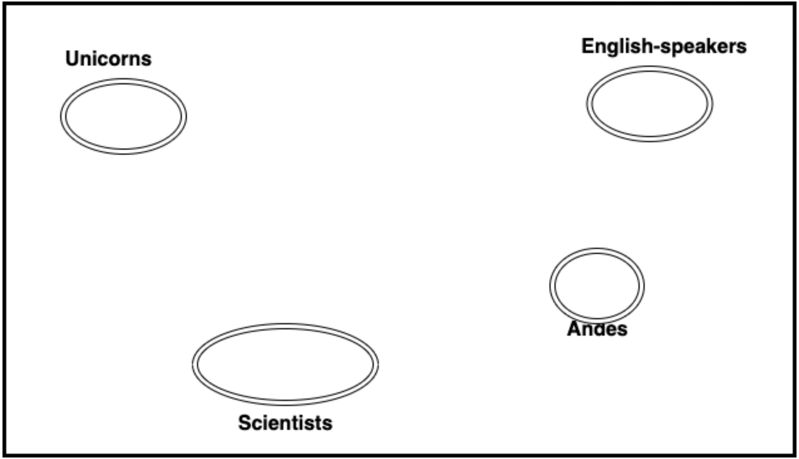
Recap: how GPT learns

During training, GPT learns about a huge *space of possible texts*.



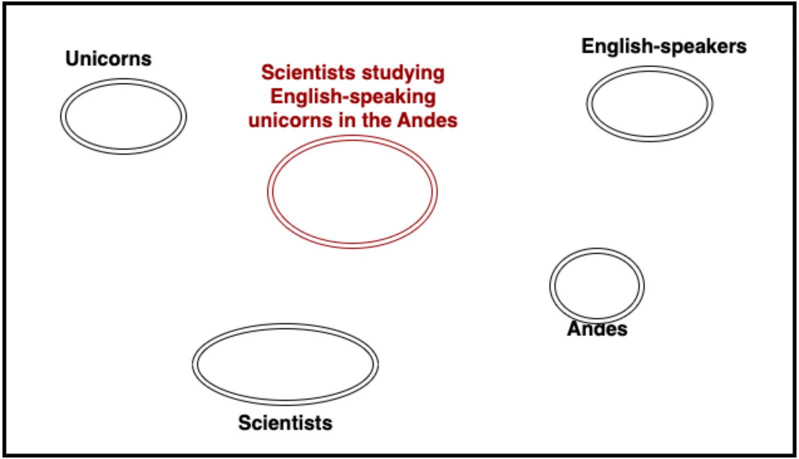
Recap: how GPT learns

This includes actual texts, but also an infinity of texts it *never saw* in training.



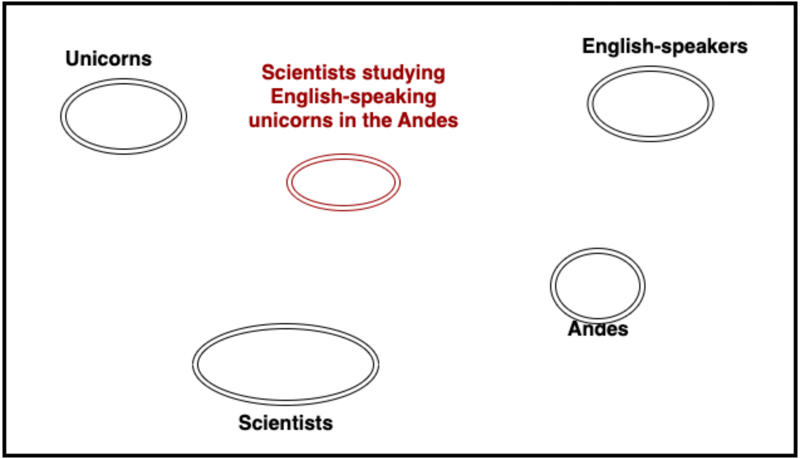
Recap: how GPT learns

When you give GPT a *prompt*, you're basically pointing to a *region* of this text space, and saying 'I want you to produce a text from *here*!'



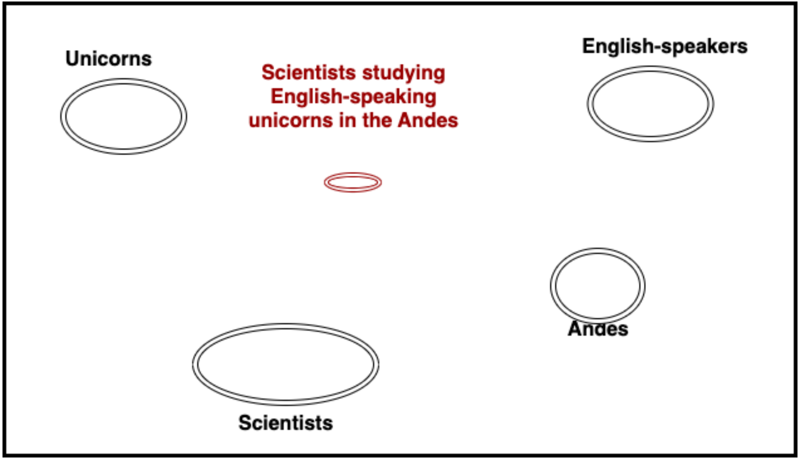
Recap: how GPT learns

The more *elaborate* your prompt is, the more *precisely* you're identifying the region you want.



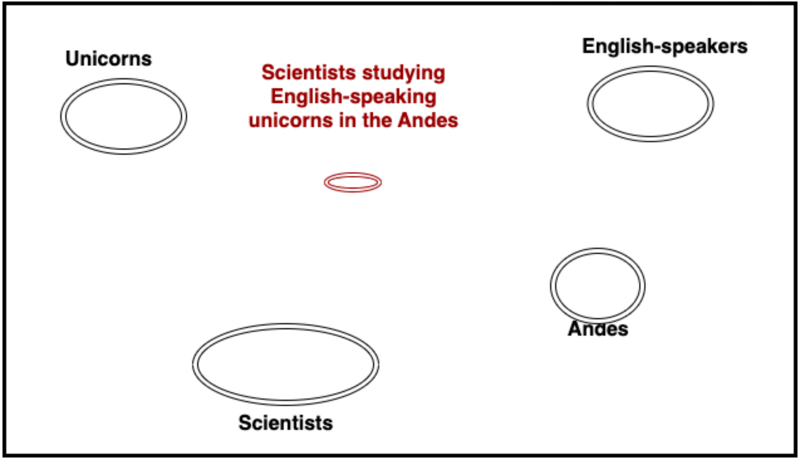
Recap: how GPT learns

If your prompt gives *examples* of the response you want, the more examples you give, the better it can respond.



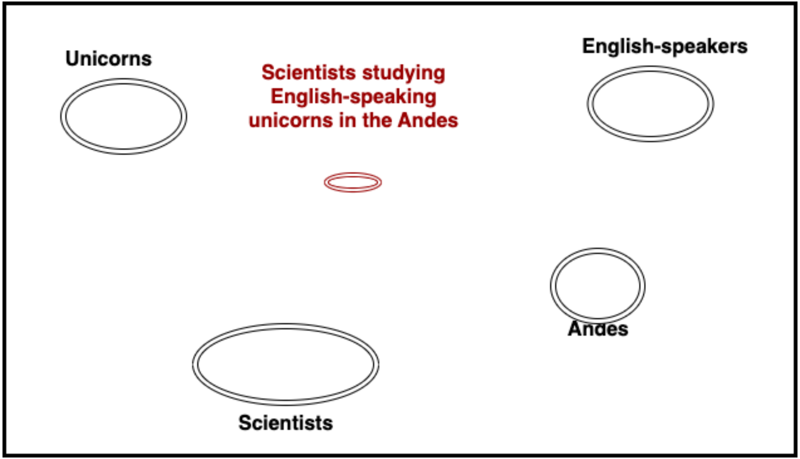
Recap: how GPT learns

This is called in-context learning: a brand new kind of learning, that doesn't involve changing weights.



Recap: how GPT learns

This is called in-context learning: a brand new kind of learning, that doesn't involve changing weights. (First described in the paper introducing GPT-3.)



Aside: in-context learning supports 'few-shot learning'

Aside: in-context learning supports 'few-shot learning'

If your prompt includes *examples* of what you want, GPT-3's outputs improve.

Aside: in-context learning supports 'few-shot learning'

If your prompt includes *examples* of what you want, GPT-3's outputs improve.

Here's a prompt with *no examples*:

```
Translate English to French:
cheese =>
```

Aside: in-context learning supports ‘few-shot learning’

If your prompt includes *examples* of what you want, GPT-3’s outputs improve.

Here’s a prompt with *one example*:

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Aside: in-context learning supports ‘few-shot learning’

If your prompt includes *examples* of what you want, GPT-3’s outputs improve.

Here’s a prompt with *three examples*:

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese =>
```

Aside: in-context learning supports ‘few-shot learning’

If your prompt includes *examples* of what you want, GPT-3’s outputs improve.

Here’s a prompt with *three examples*:

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush giraffe => girafe peluche
cheese =>
```

The more examples you add to the prompt, the better GPT-3’s response is.

In-context learning can also help GPT avoid hallucinations.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.
- Now get GPT's response to the expanded prompt.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.
- Now get GPT's response to the expanded prompt.
- OpenAI's WebGPT system did this (using Bing). ChatGPT Plus does it too.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.
- Now get GPT's response to the expanded prompt.
- OpenAI's WebGPT system did this (using Bing). ChatGPT Plus does it too.

This procedure *automatically performs in-context learning*, from *relevant texts found on the web*.

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.
- Now get GPT's response to the expanded prompt.
- OpenAI's WebGPT system did this (using Bing). ChatGPT Plus does it too.

This procedure *automatically performs in-context learning*, from *relevant texts found on the web*.

- Of course, there are lots of false/crazy documents on the web too!

In-context learning can also help GPT avoid hallucinations.

The main idea is to *use web search to automatically expand the prompt*.

- Take the user's GPT prompt. . .
- Use it to create a search query, that returns some *documents from the web*.
- Add these documents to the GPT prompt, so GPT's response is informed by these documents.
- Now get GPT's response to the expanded prompt.
- OpenAI's WebGPT system did this (using Bing). ChatGPT Plus does it too.

This procedure *automatically performs in-context learning*, from *relevant texts found on the web*.

- Of course, there are lots of false/crazy documents on the web too!
- Perhaps in a few years, many web documents will be written by language models?

‘Cited sources’ in search-enhanced language models

A search-enhanced language model can *cite the sources it consulted* in its repsonse.

‘Cited sources’ in search-enhanced language models

A search-enhanced language model can *cite the sources it consulted* in its repsonse.

- That’s very useful. (Also gives *credit* to web authors, which is a topic in itself.)

‘Cited sources’ in search-enhanced language models

A search-enhanced language model can *cite the sources it consulted* in its repsonse.

- That’s very useful. (Also gives *credit* to web authors, which is a topic in itself.)
- But the language model *itself* (the text generator) *can’t* ‘cite its sources’. . .

‘Cited sources’ in search-enhanced language models

A search-enhanced language model can *cite the sources it consulted* in its repsonse.

- That’s very useful. (Also gives *credit* to web authors, which is a topic in itself.)
- But the language model *itself* (the text generator) *can’t* ‘cite its sources’. . .
  - Its knowledge of language is distilled from *all* the documents it trained on.

‘Cited sources’ in search-enhanced language models

A search-enhanced language model can *cite the sources it consulted* in its repsonse.

- That’s very useful. (Also gives *credit* to web authors, which is a topic in itself.)
- But the language model *itself* (the text generator) *can’t* ‘cite its sources’. . .
  - Its knowledge of language is distilled from *all* the documents it trained on.
  - So ‘cited web pages’ only tell us *a little* about where a response comes from.

Identifying trustworthy content in web search

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

- Search engines have ways of preferring authoritative sources.

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

- Search engines have ways of preferring authoritative sources.
- But the most important factor in ranking a page is (still) ‘how many pages link to it’.

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

- Search engines have ways of preferring authoritative sources.
- But the most important factor in ranking a page is (still) ‘how many pages link to it’.

Search-enhanced LLMs can also include methods to help retrieve trustworthy content.

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

- Search engines have ways of preferring authoritative sources.
- But the most important factor in ranking a page is (still) 'how many pages link to it'.

Search-enhanced LLMs can also include methods to help retrieve trustworthy content.

- E.g. triangulation of responses

Identifying trustworthy content in web search

In a search-enhanced language model, some responsibility for truthful responses is deferred to the search engine.

- Search engines have ways of preferring authoritative sources.
- But the most important factor in ranking a page is (still) ‘how many pages link to it’.

Search-enhanced LLMs can also include methods to help retrieve trustworthy content.

- E.g. triangulation of responses
- E.g. prioritising trusted sources.

Fact-checking mechanisms

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

- Human fact-checkers tend to be contractors, not employees. I think they're often *ex-journalists*, who have lost their jobs due to news consumption moving online.

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

- Human fact-checkers tend to be contractors, not employees. I think they're often *ex-journalists*, who have lost their jobs due to news consumption moving online.

Fact-checking methods are currently very fragmented.

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

- Human fact-checkers tend to be contractors, not employees. I think they're often *ex-journalists*, who have lost their jobs due to news consumption moving online.

Fact-checking methods are currently very fragmented.

- There are many different organisations (Snopes, PolitiFact, PundiFact etc)—often duplicating work.

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

- Human fact-checkers tend to be contractors, not employees. I think they're often *ex-journalists*, who have lost their jobs due to news consumption moving online.

Fact-checking methods are currently very fragmented.

- There are many different organisations (Snopes, PolitiFact, PundiFact etc)—often duplicating work.

Fact-checking is *mandated by law* in every other area of business. . .

Fact-checking mechanisms

Identifying reliable sources / true stories still relies crucially on humans.

- Human fact-checkers tend to be contractors, not employees. I think they're often *ex-journalists*, who have lost their jobs due to news consumption moving online.

Fact-checking methods are currently very fragmented.

- There are many different organisations (Snopes, PolitiFact, PundiFact etc)—often duplicating work.

Fact-checking is *mandated by law* in every other area of business...

- Equity markets, real estate, insurance businesses all have auditing mechanisms ensuring their reports are truthful. Why should AI generators be any different?

What happens if lots of web pages are written by AI systems?

What happens if lots of web pages are written by AI systems?

Should we allow LLM fact-checking mechanisms to consult AI-generated material?

What happens if lots of web pages are written by AI systems?

Should we allow LLM fact-checking mechanisms to consult AI-generated material?

Should we allow LLMs to *train* on AI-generated content?

What happens if lots of web pages are written by AI systems?

Should we allow LLM fact-checking mechanisms to consult AI-generated material?

Should we allow LLMs to *train* on AI-generated content?

- A recent paper demonstrates 'model collapse' if this happens repeatedly.

What happens if lots of web pages are written by AI systems?

Should we allow LLM fact-checking mechanisms to consult AI-generated material?

Should we allow LLMs to *train* on AI-generated content?

- A recent paper demonstrates ‘model collapse’ if this happens repeatedly.

We probably need *ways of knowing whether a piece of online content was made by a human or a machine*.

Transparency about AI-generated content

Transparency about AI-generated content

Say I'm given a text to look at... e.g. a student essay, or a social media post.

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?*

Transparency about AI-generated content

Say I'm given a text to look at...e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

The world is about to be flooded with AI-generated content.

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

The world is about to be flooded with AI-generated content.

- I think people have a right to know whether the content they consume is generated (wholly or partially) by a machine.

Transparency about AI-generated content

Say I'm given a text to look at. . . e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

The world is about to be flooded with AI-generated content.

- I think people have a right to know whether the content they consume is generated (wholly or partially) by a machine.

It would be good if there were reliable detectors for AI-generated content.

Transparency about AI-generated content

Say I'm given a text to look at... e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

The world is about to be flooded with AI-generated content.

- I think people have a right to know whether the content they consume is generated (wholly or partially) by a machine.

It would be good if there were reliable detectors for AI-generated content.

- As before, we could use AI systems to *help keep other AI systems safe*.

Transparency about AI-generated content

Say I'm given a text to look at...e.g. a student essay, or a social media post.

- This text might have been written by a person, or by an AI.
- How can I tell?
- *Does it matter?* I think so!

The world is about to be flooded with AI-generated content.

- I think people have a right to know whether the content they consume is generated (wholly or partially) by a machine.

It would be good if there were reliable detectors for AI-generated content.

- As before, we could use AI systems to *help keep other AI systems safe*.
- But as language models *improve*, detection will become increasingly hard.

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release.

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release.
> *The mechanism must be made* freely available *to the public.*

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release.
> *The mechanism must be made* freely available *to the public.*

We have discussed this proposal with EU legislators, working on amendments to the EU's AI Act.

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release*.*
> *The mechanism must be made* freely available *to the public.*

We have discussed this proposal with EU legislators, working on amendments to the EU's AI Act.

- It has been incorporated into the amendments proposed by the EU Parliament.

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release.
> *The mechanism must be made* freely available *to the public.*

We have discussed this proposal with EU legislators, working on amendments to the EU's AI Act.

- It has been incorporated into the amendments proposed by the EU Parliament.
- These still need to be discussed with the EU Commission, and Council of Ministers.

A proposed law for generative AI models

A group I work with in the Global Partnership on AI, is proposing a new law:

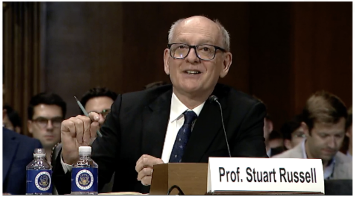
> *If an organisation develops a new state-of-the-art generative AI model, it must demonstrate a reliable mechanism for detecting the content it generates*, as a condition of release.
> *The mechanism must be made* freely available *to the public.*

We have discussed this proposal with EU legislators, working on amendments to the EU's AI Act.

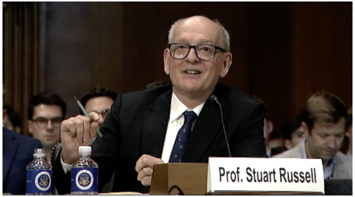
- It has been incorporated into the amendments proposed by the EU Parliament.
- These still need to be discussed with the EU Commission, and Council of Ministers.

Our proposal was also discussed this week in a US Senate Judiciary hearing.

This week in the Senate



This week in the Senate



‘Move fast and fix things’

Summary

Summary

Language models are amazing. . .

Summary

Language models are amazing...

...And they need to be very carefuly scrutinised.

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .
- And by policymakers.

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .
- And by policymakers.

We need to be aware of—

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .
- And by policymakers.

We need to be aware of—

- Their ability to generate harmful content

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .
- And by policymakers.

We need to be aware of—

- Their ability to generate harmful content
- Their ability to generate falsehoods, and present them as truth

Summary

Language models are amazing. . .

. . . And they need to be very carefuly scrutinised.

- Both by users, and researchers. . .
- And by policymakers.

We need to be aware of—

- Their ability to generate harmful content
- Their ability to generate falsehoods, and present them as truth
- Their ability to generate content that looks like it was written by a person.