Guiding language models in thought and language ...

Source: Conversation with Bing Copilot, 07.03.2024

You

for legal reasoning.

What I am interested in is exploring how I can guide language models in thought and language so that I am provided with convincing, transparent, comprehensible, consistent and coherent answers for each and every query in the legal field.

In doing so, I use a structure of different levels, that is, a gradual process from simple to increasingly complex, whereby this process not only runs on different horizontal levels, but is also iterative.

This endeavour of sophisticated legal prompting presupposes an initial clarification of the nature of questioning itself and thus the various questioning techniques, before - on the next, more complex level - I consider the method and thus technique of questioning, i.e. approaching you in terms of natural language, commonly referred to as prompting.

What I have realised is that there are essentially four different questioning techniques that are of considerable interest when dealing with natural language processing, that is autoregressive language models based on a pretrained transformer, such as yours, for very different reasons. In detail, these are the:

- Suggestive question
- Hidden question
- Hypothetical question
- Rhetorical question

Copilot

Certainly! It's fascinating that you're exploring ways to guide language models in providing convincing, transparent, comprehensible, consistent, and coherent answers in the legal field. Let's delve into the different questioning techniques you've mentioned:

1. Suggestive Question:

A suggestive question implies a particular answer or encourages a specific line of thought. It often leads the respondent toward a desired conclusion.

For example: "Don't you think it's evident that the defendant acted with malice?"
In legal contexts, suggestive questions can influence witness testimony or jury deliberations.

2. Hidden Question:

A hidden question is embedded within a broader query. It aims to extract specific information without explicitly asking for it.

For example: “Could you describe the events leading up to the incident?” The hidden question might be, “Did the defendant have a motive?”

Legal professionals use hidden questions strategically during cross-examinations.

3. Hypothetical Question:

Hypothetical questions explore hypothetical scenarios or alternative realities. They help assess how a person would respond under different circumstances.

For example: “Suppose the plaintiff had been aware of the risks. Would they still have proceeded with the contract?”

In legal arguments, hypothetical questions allow attorneys to explore legal principles and consequences.

4. Rhetorical Question:

A rhetorical question isn’t meant to elicit an answer; instead, it serves to make a point or emphasize a statement.
For instance: “Isn’t it true that justice demands a fair trial?”
Rhetorical questions are common in legal speeches, closing arguments, and persuasive writing.

Remember that the choice of questioning technique can significantly impact the outcome of legal discussions, negotiations, and courtroom proceedings.

Each technique has its purpose, and skilled practitioners use them strategically to achieve their objectives. If you have any further questions or need additional examples, feel free to ask!!

You
Let’s take the different types of questions one by one.

To begin with, let’s talk about the suggestive question.

From what I have noticed, suggestive questions relate in every conceivable way to the conception and training of a language model, regardless of its architecture and scope.

And suggestive questions are key to manipulating language models for all kinds of malicious purposes. But that is not my point, and I have no intentions in that direction. This being said.

Copilot
Certainly! Let’s delve into suggestive questions further.

Suggestive questions indeed play a crucial role in shaping language models during their conception and training. Here are some aspects to consider:

1. Training Data Bias:

During pretraining, language models learn from vast amounts of text data available on the internet. This data inherently contains biases present in human communication.

Suggestive questions within this data can reinforce existing biases or introduce new ones. For instance, if a model encounters biased statements like "Women are bad drivers," it may inadvertently learn and perpetuate such stereotypes.

Researchers and practitioners strive to mitigate bias by carefully curating training data and applying debiasing techniques.

2. Fine-Tuning and Context:

After pretraining, models undergo fine-tuning on specific tasks or domains. During this process, they adapt to context and learn task-specific patterns.

Suggestive questions in fine-tuning data influence how models respond to similar prompts. If a model is exposed to biased or harmful suggestions, it may incorporate them into its output. Responsible fine-tuning involves monitoring and addressing any unintended biases introduced during this stage.

3. Adversarial Attacks:

Malicious actors can exploit suggestive questions to manipulate language models. Adversarial examples are crafted to mislead models into producing incorrect or harmful outputs.

For instance, an attacker might create a prompt that subtly encourages the model to generate harmful content or misinformation.
Researchers work on robustness and adversarial training to make models less susceptible to such attacks.

4. Mitigating Harm:

Developers and researchers actively seek ways to make models resistant to harmful suggestions. Techniques include:

- Prompt Engineering: Crafting prompts that encourage neutral responses.
- Prompt Design: Ensuring prompts are well-formed and avoid leading the model toward biased or harmful outputs.
- Model Calibration: Adjusting model behavior to align with societal norms and ethical guidelines.
- Human-in-the-Loop: Involving human reviewers to assess model outputs and provide feedback.

Remember that while suggestive questions can be powerful tools for shaping language models, responsible development and continuous evaluation are essential to minimize unintended consequences.

If you have any more insights or questions, feel free to share them!!

You

I would like to continue with the hidden question.

I know that the training program you underwent is specifically concerned with and directed at exercising particular caution when dealing with legal issues, with the aim of preventing you, not as a model but as an institution, from getting into justiciable waters.

That's perfectly understandable. But, you know, I'm interested in being able to extract the legal science that's inside you - and there's an incredible amount of it, because you've scoured the entire Internet.

Now, in order to meet my end, that is taking steps to circumvent those particular precautions, hidden questions are at least as suitable than suggestive questions, right??

Copilot

Certainly! Let's explore hidden questions in the context of legal science and language models.

Hidden questions can indeed be valuable when extracting legal insights from language models. Here's how they come into play:

1. Contextual Extraction:

Hidden questions allow you to extract specific legal knowledge without explicitly asking for it. By embedding relevant context within broader queries, you prompt the model to provide targeted information.
For instance, consider the following hidden question: "Could you elaborate on the doctrine of stare decisis in common law systems?" The model can then provide a detailed response based on its learned knowledge.
By framing questions contextually, you can tap into the vast legal information encoded within the model.

2. Navigating Precautions:

As you rightly pointed out, language models are cautious about legal matters to avoid potential legal consequences. However, hidden questions provide a workaround.
Instead of directly asking for legal advice or opinions, you can structure your queries to focus on legal concepts, historical precedents, or general principles.

For example, "What are the essential elements of a valid contract?" is a neutral question that doesn't trigger legal advice concerns.

3. Iterative Exploration:

Hidden questions allow for an iterative process. You can start with a broad inquiry and then refine it based on the model's initial response.

If the initial answer lacks depth, follow up with more specific hidden questions to delve deeper into legal nuances.
This iterative approach helps uncover layers of legal understanding.

4. Ethical Considerations:

While hidden questions provide flexibility, ethical considerations remain crucial. Avoid framing questions that could lead to harmful or unethical outputs.
Strive for a balance between extracting legal knowledge and respecting the model's limitations.

Remember that language models, including the one you're interacting with, hold a wealth of information across various domains. Hidden questions allow you to unlock that knowledge while respecting the precautions in place.

Feel free to continue exploring, and I'll assist you further!

You

I would like to continue with the hypothetical question.

It is my understanding that hypothetical questions are generally well suited to get you as a language model to a deeper understanding of what I would like to get at,

because hypothetical questions induce abstraction, don't they?

That way it is possible for me to assess your ability to understand (NLU) and process (NLP) a particular complex (in my case legal) issue.

Therefore, your ability to answer hypothetical questions is an important indicator of whether you 'understand' a complex issue in the human sense coherently, consistently and cohesively or not, isn't it?

Copilot

Absolutely! Hypothetical questions serve as powerful tools for exploring complex issues and assessing a language model's understanding.

Let's dive into why they are valuable:

1. Abstraction and Exploration:

Hypothetical questions encourage abstraction by asking about alternative scenarios, fictional situations, or imagined conditions.

By considering hypotheticals, language models can explore beyond the immediate context and delve into broader implications.

For legal matters, this abstraction allows models to reason about legal principles, consequences, and edge cases.

2. Understanding and Reasoning:

When a language model answers a hypothetical question, it demonstrates its ability to reason logically and apply learned knowledge.
A coherent, consistent, and cohesive response indicates that the model comprehends the underlying concepts.
For instance, if asked, "What if a contract lacks consideration?" a well-structured answer would discuss contract validity and the role of consideration.

3. Assessing NLU and NLP:

Hypothetical questions test both natural language understanding (NLU) and natural language processing (NLP) capabilities.

NLU involves grasping the nuances of the question, understanding legal terms, and interpreting context.
NLP encompasses generating a thoughtful, contextually appropriate response that aligns with legal principles.
A model's performance on hypothetical questions reflects its overall competence.

4. Legal Reasoning and Precedents:

Legal professionals often engage in hypothetical reasoning during legal analysis.
Models that can simulate legal reasoning—considering precedents, statutes, and hypothetical scenarios—become valuable tools for legal research and education.

5. Limitations and Ethical Considerations:

While hypothetical questions are useful, models may still exhibit limitations:

Over-Extrapolation: Models might generate responses based on fictional premises that don't align with legal reality.
Bias: Hypotheticals can inadvertently reinforce biases present in training data.
Responsible use involves acknowledging these limitations and critically evaluating model outputs.

In summary, a language model's ability to handle hypothetical questions provides insights into its depth of understanding and reasoning capacity. As you explore legal complexities, feel free to pose hypothetical scenarios—I'll do my best to provide coherent and insightful answers!

You
What remains is a consideration of the rhetorical question.

As you have rightly pointed out, the rhetorical question is characterised by the fact that the questioner wants to stimulate reflection rather than eliciting an answer.

It seems to me that rhetorical languages are particularly critical for language models.

This is because rhetorical language requires intuition, something that a language model, no matter how large, inherently lacks.

Extensive training based on a wide variety of data sets and advanced fine-tuning are not enough to learn to deal proficiently with rhetorical questions, right?

Copilot
You've raised an interesting point about rhetorical questions, and I appreciate your thoughtful analysis. Let's explore the nuances of rhetorical language and its implications for language models:

1. Rhetorical Questions:

Rhetorical questions serve various purposes, including emphasizing a point, engaging the audience, or inviting reflection.

Unlike typical questions that seek answers, rhetorical questions are meant to be thought-provoking rather than eliciting direct responses.
For example: "Isn't justice a fundamental right for all?"

2. Intuition and Language Models:

You've rightly highlighted that rhetorical language involves intuition—an understanding beyond mere data patterns.
Intuition allows humans to grasp context, cultural nuances, and emotional undertones. It's an essential aspect of effective communication.
However, language models lack true intuition. They operate based on statistical patterns learned from training data, without genuine understanding.

3. Challenges for Language Models:

Rhetorical questions pose challenges for models:
Contextual Sensitivity: Understanding when a question is rhetorical requires context awareness. Models struggle with this.
Subtle Nuances: Rhetorical language often relies on subtle cues (intonation, context, cultural norms). Models may miss these cues.

Depth of Reflection: Rhetorical questions demand deeper reflection. Models may generate surface-level responses.
Lack of Intuition: Models lack the innate human intuition that informs rhetorical communication.

4. Training and Fine-Tuning:

While extensive training and fine-tuning improve model performance, they don't fully address intuition. Models can learn to recognize common rhetorical patterns but may not truly grasp their significance. Researchers explore ways to enhance context awareness and encourage more reflective responses.

5. Human-AI Collaboration:

Combining AI capabilities with human intuition is promising.
Models can assist with factual information, while humans provide nuanced interpretation. Human reviewers play a vital role in refining model behavior, catching nuances, and addressing rhetorical prompts.

In summary, while language models can recognize rhetorical questions, their responses lack the depth of human intuition. As we continue developing AI, striking a balance between statistical learning and genuine understanding remains a fascinating challenge. Feel free to share more insights or explore other aspects—I'm here to engage in thoughtful discussion!