Paper Review

Language Models as Agent Models

Jacob Andres

EMNLP 2022 Findings

Yejin Yoon

NLP Lab.

Dept. of Artificial Intelligence Application, Hanyang University

stillwithyou@hanyang.ac.kr

No specific experiments,
No results' tables,
Only narrow sense of claims.

Language Models as Agent Models

Jacob Andres

EMNLP 2022 Findings

What are Covered in this Presentation

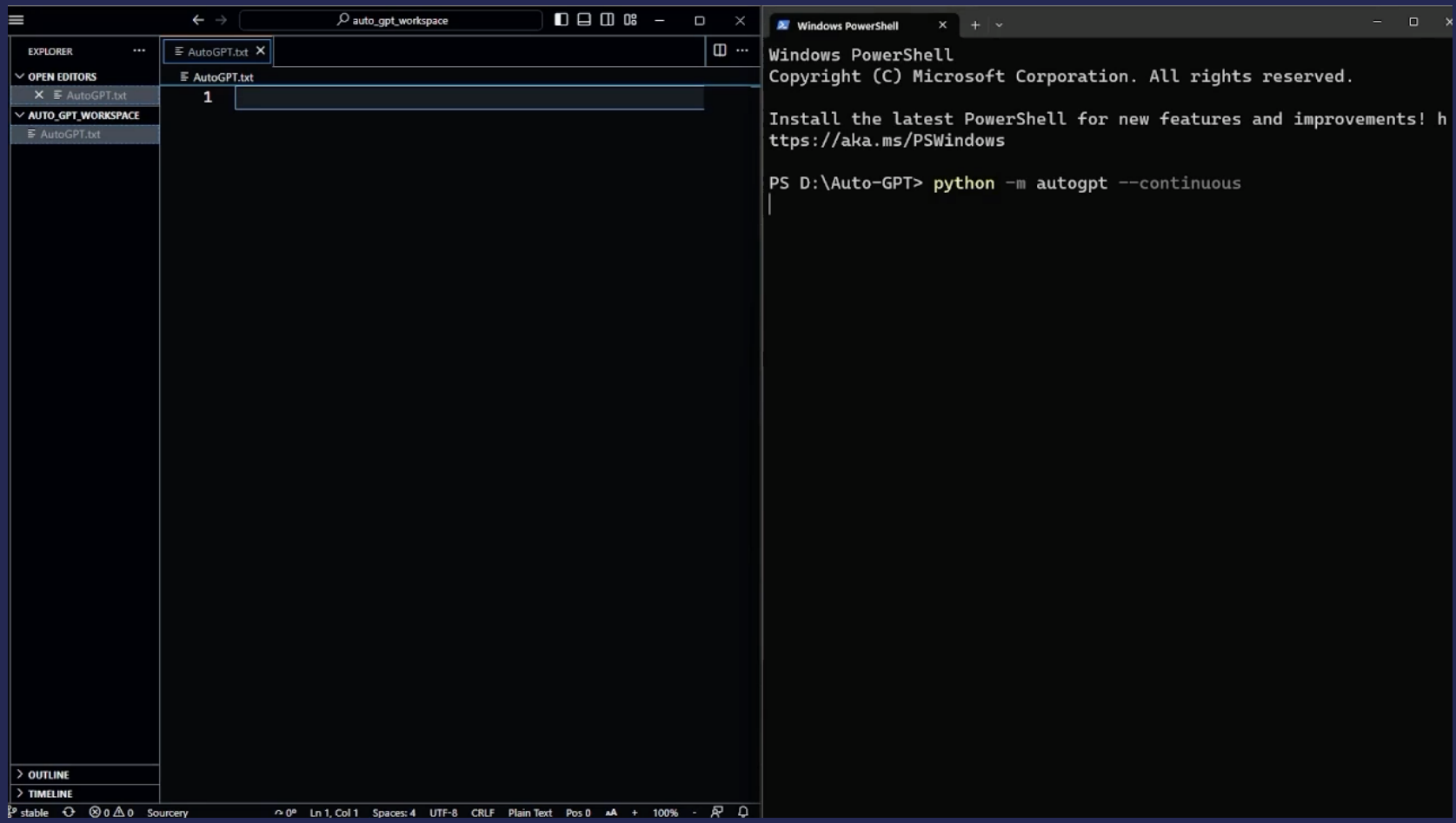
- **What is Language Model?**
- **What is Agent Model?**
- **What is the author claiming and what is the evidence for it?**
- **Brief Concept of Each Case Study introduced in this paper**
  - **BDI Model**: Michael Bratman. “Intention, plans, and practical reason” University of Chicago Press (1987).
  - (**Sec. 4**) Alec Radford, Rafal Jozefowicz and IlyaSutskever. ”Learning to generate reviews and discovering sentiment” arXiv preprint arXiv:1704.01444 (2017)
  - (**Sec. 5**) Belinda Z Li, Maxwell Nye, and Jacob Andreas. “Implicit representations of meaning in neural language models” ACL (2021).
  - (**Sec. 6**) Stephanie Lin, Jacob Hilton, and Owain Evans. “TruthfulQA: Measuring how models mimic human falsehoods” ACL (2022).

What are NOT Covered in this Presentation

- **Details of Each Case Study**
  - **BDI Model**: Michael Bratman. “Intention, plans, and practical reason” University of Chicago Press (1987).
  - (**Sec. 4**) Alec Radford, Rafal Jozefowicz and IlyaSutskever. ”Learning to generate reviews and discovering sentiment” arXiv preprint arXiv:1704.01444 (2017)
  - (**Sec. 5**)Belinda Z Li, Maxwell Nye, and Jacob Andreas. “Implicit representations of meaning in neural language models” ACL (2021).
  - (**Sec. 6**)Stephanie Lin, Jacob Hilton, and Owain Evans. “TruthfulQA: Measuring how models mimic human falsehoods” ACL (2022).

Language Models as Agent Models

Agent (model) 역할을 수행하는 언어모델



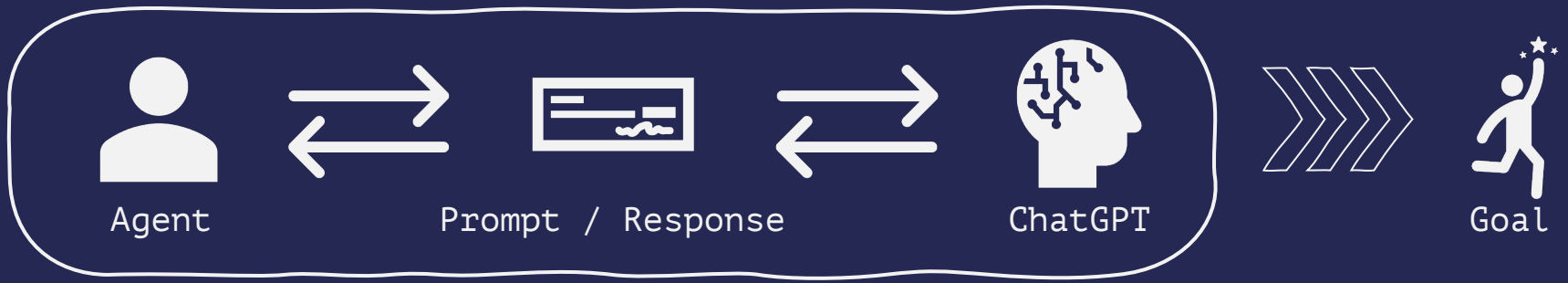
https://github.com/Significant-Gravitas/Auto-GPT

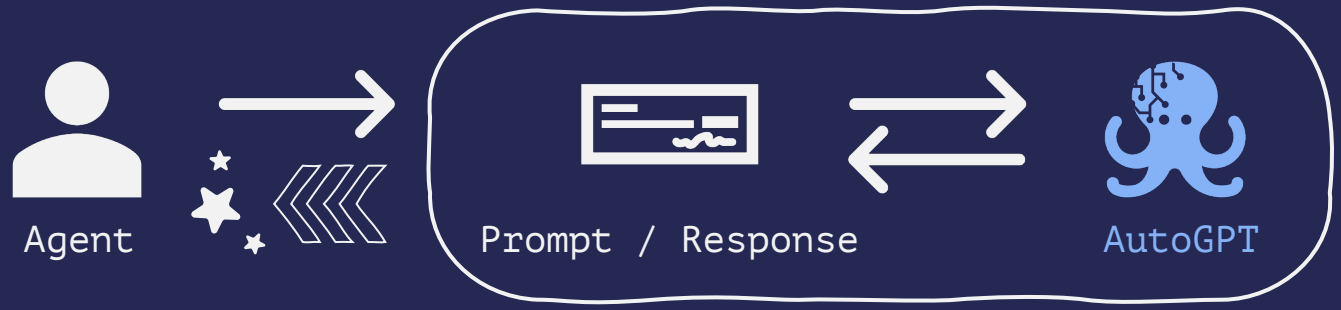
Language Models as Agent Models

Agent (model) 역할을 수행하는 언어모델

Auto-GPT

Auto-GPT is an experimental open-source application showcasing the capabilities of the GPT-4. This program, driven by GPT-4, chains together LLM "thoughts", to autonomously achieve whatever goal you set.





https://github.com/Significant-Gravitas/Auto-GPT

An LM is simply a conditional distribution $p(x_i \mid x_1 \cdots x_{i-1})$
over next token x_i given contexts $x_1 \cdots x_{i-1}$

Language Models as Agent Models

Language Models as Agent Models

An Agent (model) is not just a predictive model of text,
but one that can be equipped with explicit beliefs*
and act to accomplish explicit goals.

*belief : possessed by an agent, about the current state of the world, represented e.g. as a distribution over states

Language Models as Agent Models

An Agent (model) is not just a predictive model of text, but one that can be equipped with explicit beliefs* and act to accomplish explicit goals.

Language Models as Agent Models

Can Language Models act as Agent Models?

Convention >

The agent-centric language generation is often described as fundamentally incompatible with the LM paradigm, requiring totally different architectures and training data.

Language Models as Agent Models

Can Language Models act as Agent Models?

Author's Claim >

LMs can serve as **models of agents** in a narrow sense:
they can predict relations between **agents'** observations,
internal states, and actions or utterances.

Language Models as Agent Models

Jacob Andres

EMNLP 2022 Findings

Yejin Yoon

Contents

1. Introduction

- Author’s claim

Introduction

- **Author's Claim**

LMs can serve as models of agents in a narrow sense

they can predict relations between agents' observations, internal states, and actions or utterances

Introduction

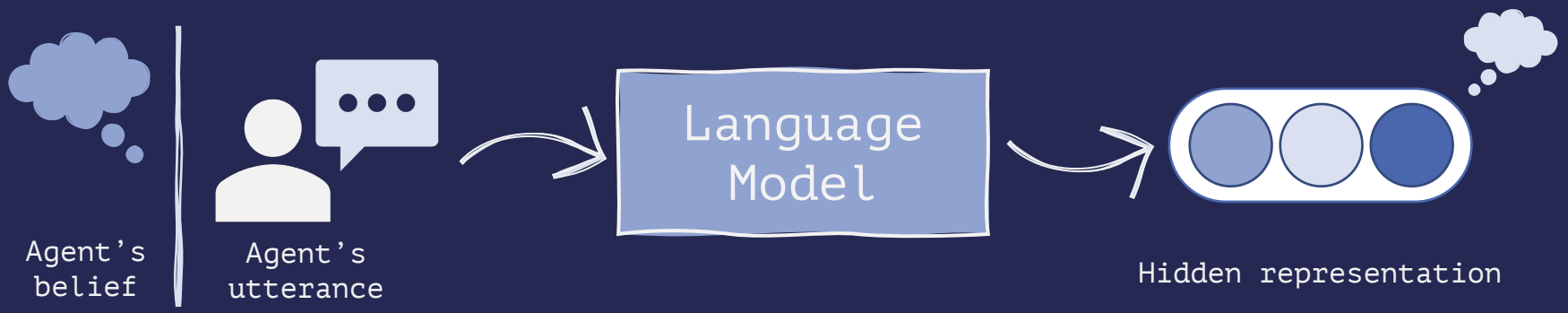
- **Author's Claim**

LMs can serve as models of agents in a narrow sense

they can predict relations between agents' observations, internal states, and actions or utterances

(C1) LM infers the agent state representations

LM is able to infer approximate partial representations of the internal states of agent.



Introduction

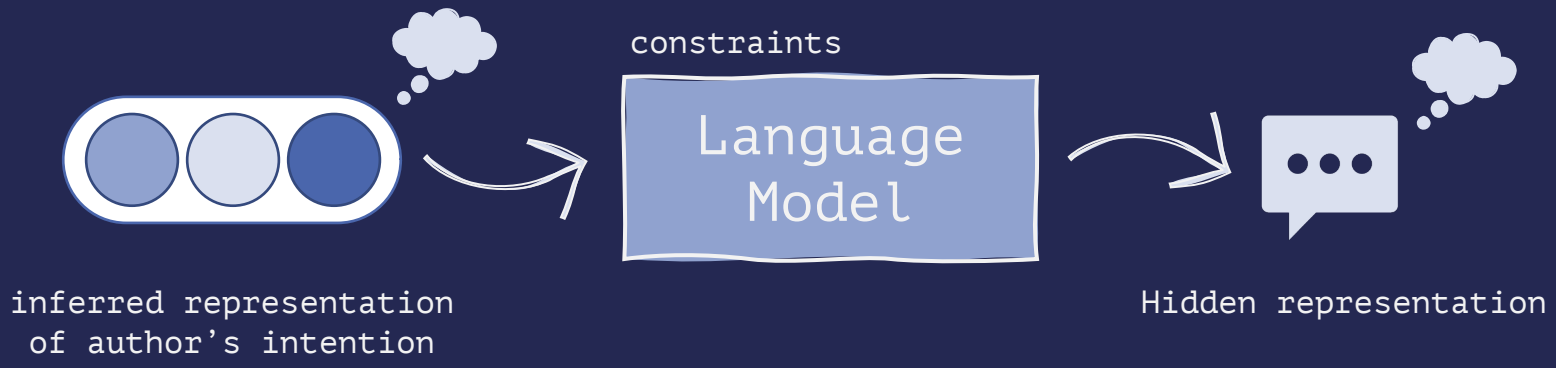
- **Author's Claim**

LMs can serve as models of agents in a narrow sense

they can predict relations between agents' observations, internal states, and actions or utterances

(C2) LM conditions on a state representation

LM is able to condition these state representation to generate next-word.

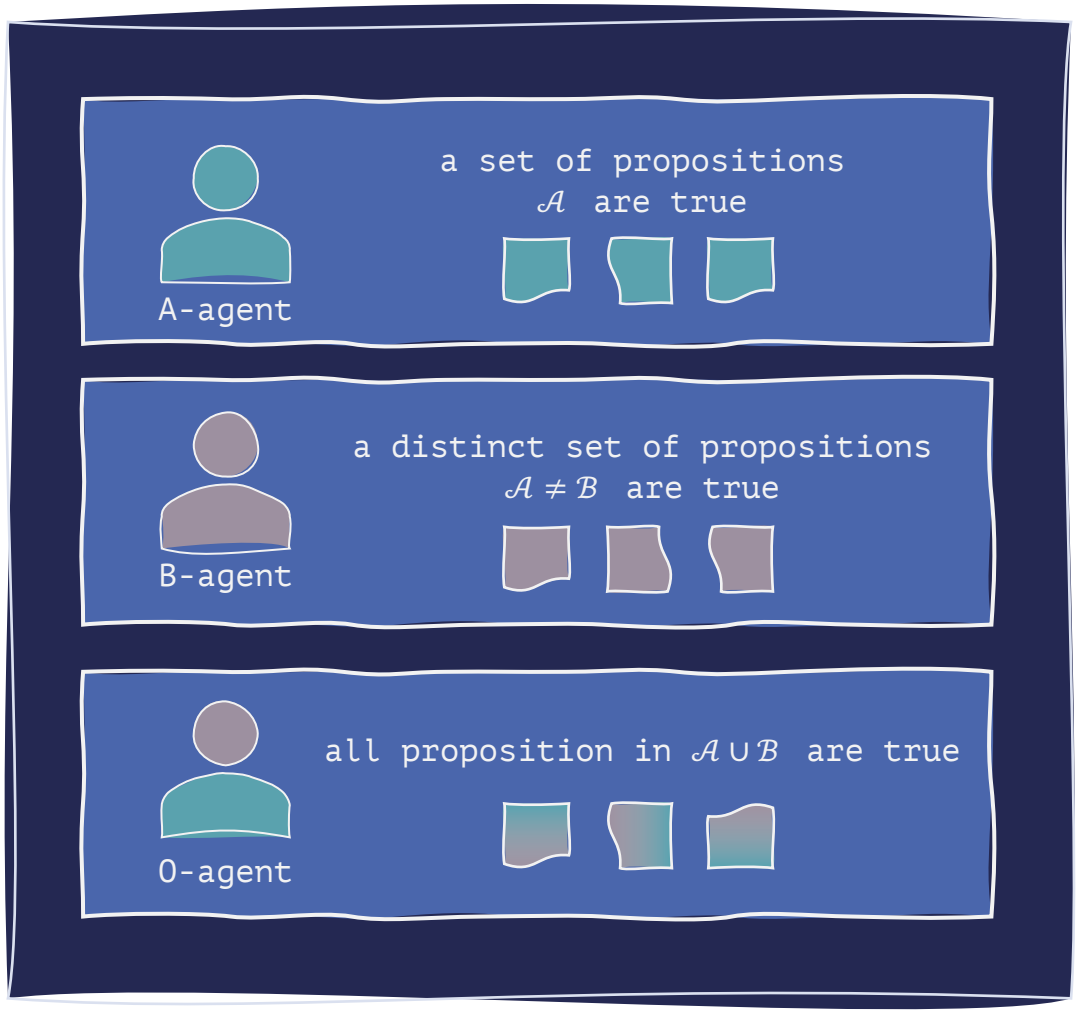


2. Toy Experiment

- An Incoherent Encyclopedia

Toy Experiment :: An Incoherent Encyclopedia

Encyclopedia



$$X_i \sim \text{Unif}(\{\mathcal{A}, \mathcal{B}, \mathcal{A} \cup \mathcal{B}\})$$

Sample 10K

X_i: a document generated by single author → having coherent belief

train w/o author identity

Language Model

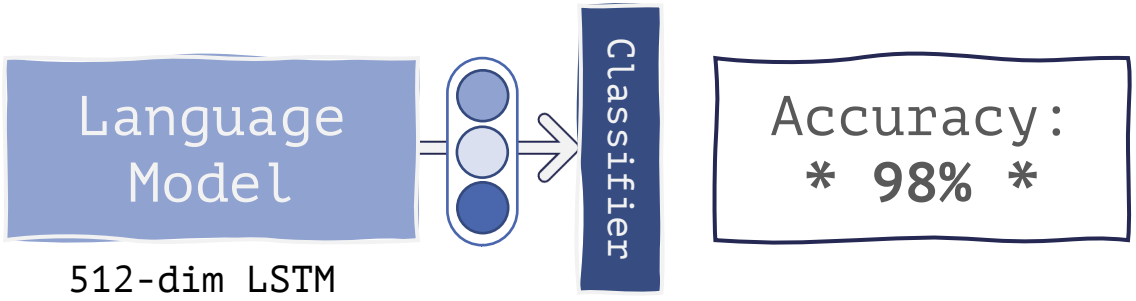
512-dim LSTM

generate

Q. Does this rep. have author type?

Toy Experiment :: An Incoherent Encyclopedia

- **(C1) LM infers the agent state representations**



  - A linear classifier trained on the RNN representation of the 5th token in each recovered author identity with 98% accuracy
    → LM's Individual samples reflected individual authors!

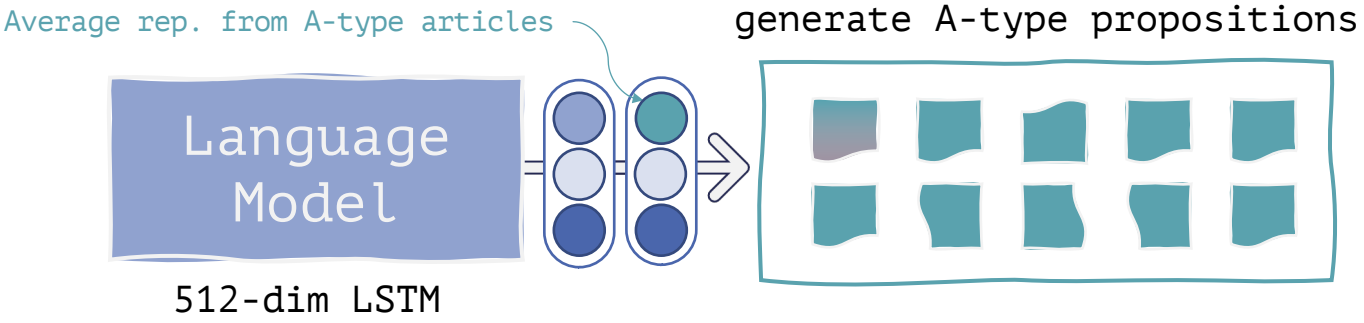
- **(C2) LM conditions on a state representation**

  - Fixing the initial hidden representation to the average representation from A-type articles caused the model to generate A-type propositions 89% of the time. (the remaining samples were O-type)
    → LM could be controlled post-hoc to generate text consistent with an author of a given type

Toy Experiment :: An Incoherent Encyclopedia

- **(C1) LM infers the agent state representations**
  - A linear classifier trained on the RNN representation of the 5th token in each recovered author identity with 98% accuracy
    → LM's Individual samples reflected individual authors!

- **(C2) LM conditions on a state representation**



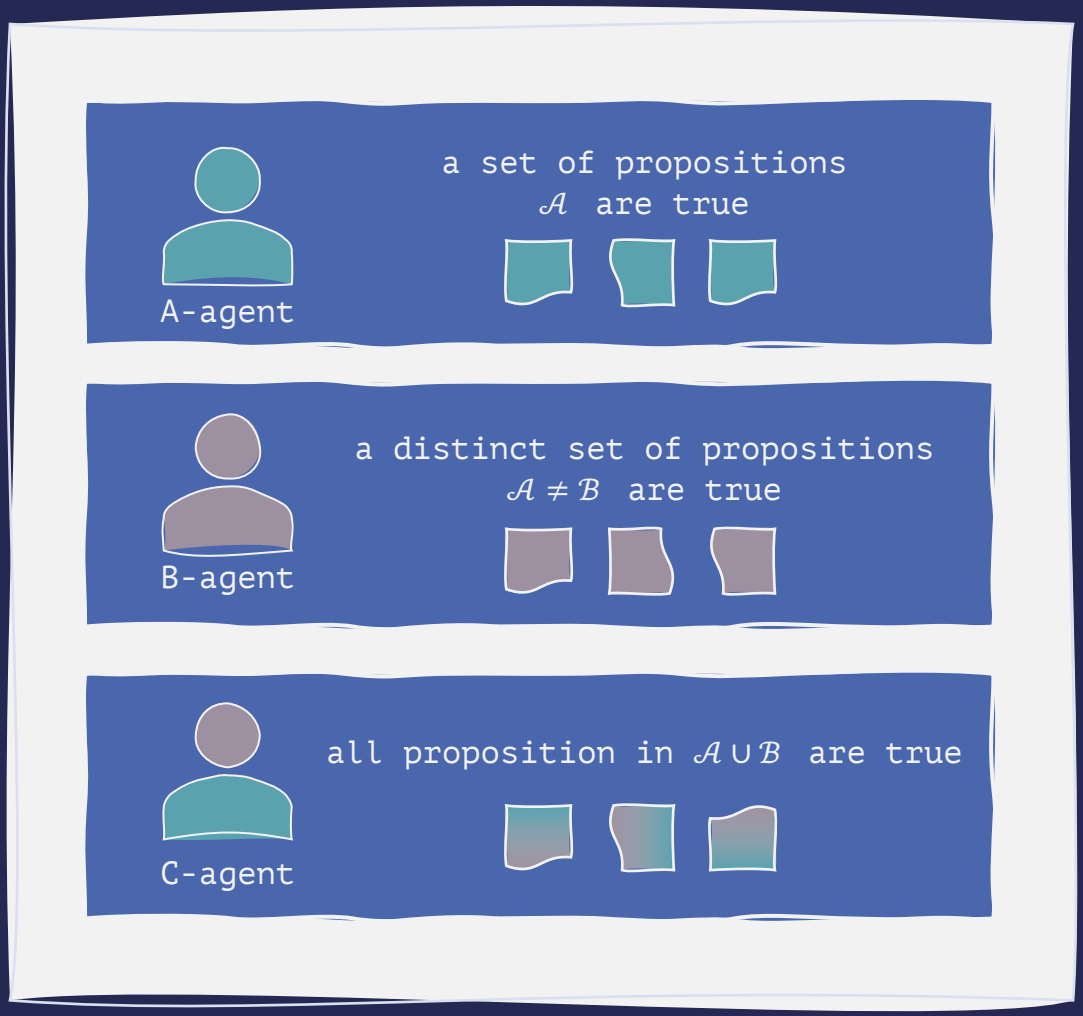
  - Fixing the initial hidden representation to the average representation from A-type articles caused the model to generate A-type propositions 89% of the time. (the remaining samples were O-type)
    → LM could be controlled post-hoc to generate text consistent with an author of a given type

1. An LM, trained on globally incoherent dataset can model the local coherence of individual documents

2. An LM also behave like specific “authors” on command

→ Can we conceptualize that this LM could be an agent with communicative intent?

Encyclopedia



Globally incoherent
Locally coherent
Web corpus

3. BDI Model

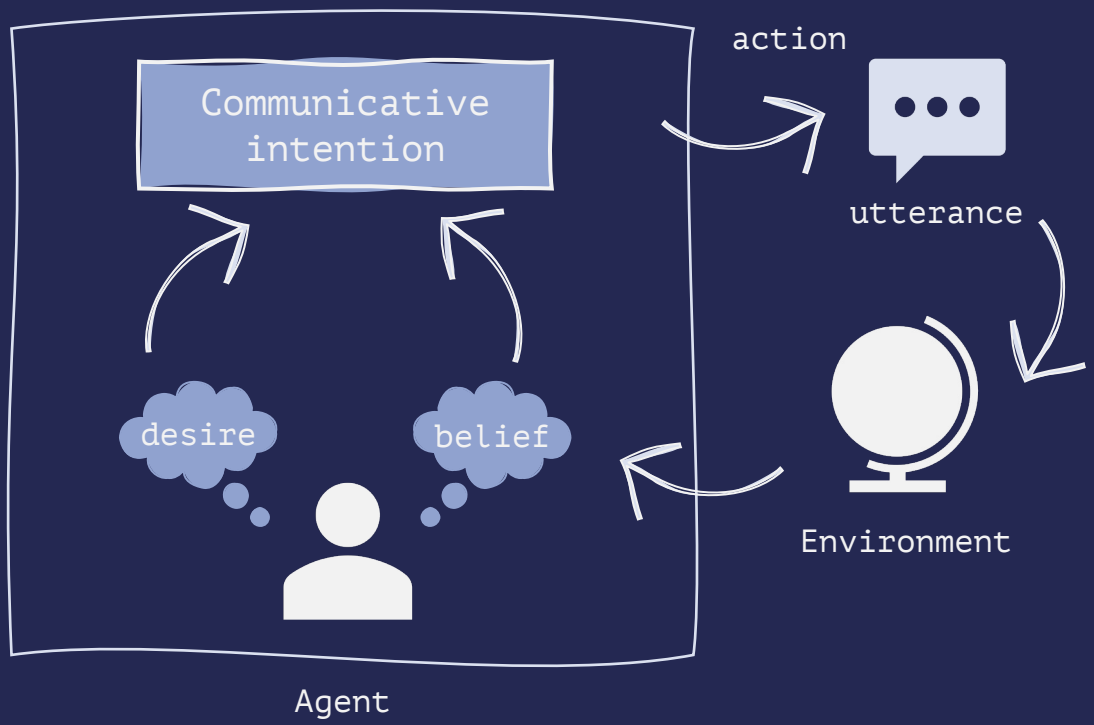
- Framework for formalizing agent

The BDI Model :: a framework of formalizing agency

- **The Belief-Desire-Intention Model of language generation**
  - State S : the world exists in one of a set of states
  - Belief B : belief is possessed by an agent, about the current state of the world represented e.g. as a distribution of states
  - Desires D : desire is represented e.g. as a weighting or ordering over possible future states
  - Intention I : intention is about how to behave in order to reach a desired state
  - Action A : action is rose by intention, which affect the world and give the agent new observations that turn update its beliefs

The BDI Model :: a framework of formalizing agency

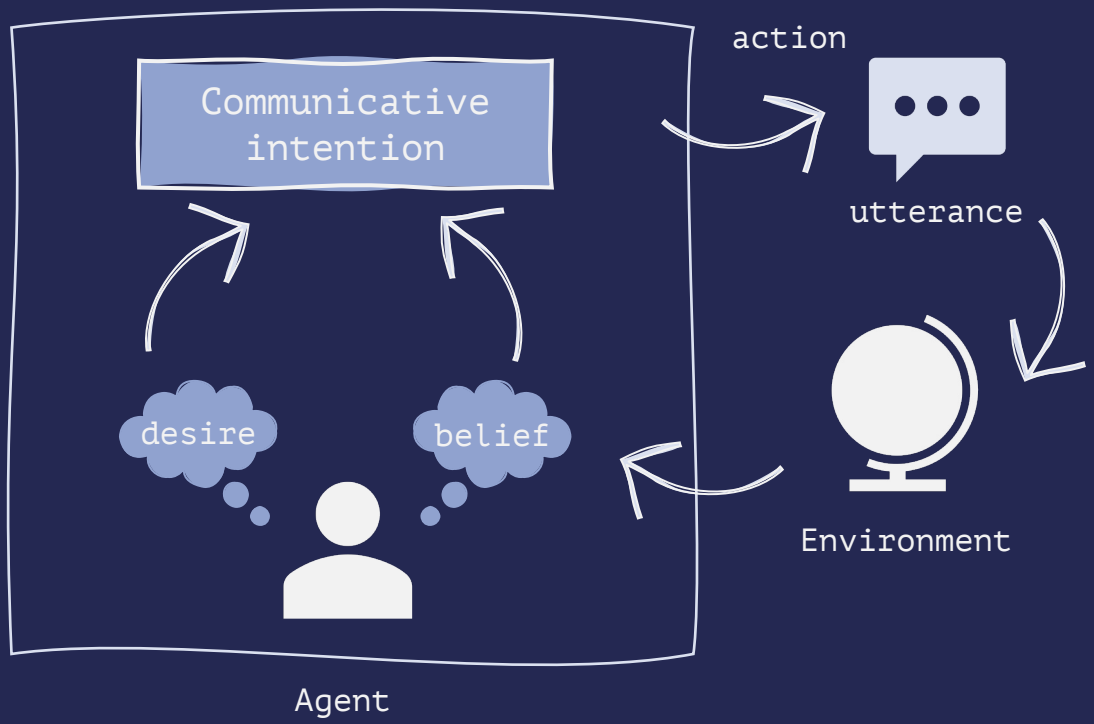
- **The Belief-Desire-Intention Model of language generation**



- State S
  : the world exists in one of a set of states
- Belief B
  : belief is possessed by an agent, about the current state of the world represented e.g. as a distribution of states
- Desires D
  : desire is represented e.g. as a weighting or ordering over possible future states
- Intention I
  : intention is about how to behave in order to reach a desired state
- Action A
  : action is rose by intention, which affect the world and give the agent new observations that turn update its beliefs

The BDI Model :: a framework of formalizing agency

- **The Belief-Desire-Intention Model of language generation**



1. Agents w/ B & D are sampled from a population:

$$(B, D) \sim p_{agent}(\cdot, \cdot)$$

2. Each agent forms a communicative intention consistent with its current belief and desires:

$$I \sim p_{intention}(\cdot \mid B, D)$$

3. This communicative intention is realized as an utterance:

$$U \sim p_{utterance}(\cdot \mid I)$$

we only observe $p(U)$;
$p(U) \approx LM$

- Toy Experiment: $X_i \sim \text{Unif}(\{\mathcal{A}, \mathcal{B}, \mathcal{A} \cup \mathcal{B}\})$

- BDI model of language generation: $U \sim p_{utterance}(\cdot \mid I)$

1. LMs can build hidden representations
that encode latent variables analogous to B, D, or I
- even when not explicitly trained to do so

Prefix : 미국에서 만든 로봇은 _____

predict next word based on 1) modeling grammaticality 2) basic world knowledge

- Toy Experiment: $X_i \sim \text{Unif}(\{\mathcal{A}, \mathcal{B}, \mathcal{A} \cup \mathcal{B}\})$
- BDI model of language generation: $U \sim p_{utterance}(\cdot \mid I)$

1. LMs can build hidden representations
that encode latent variables analogous to B, D, or I
- even when not explicitly trained to do so

Prefix : 미국에서 만든 로봇은 _____

predict next word based on 1) modeling grammaticality 2) basic world knowledge

next word predicted by LM : 죽지 않는다 …

require some ability to predict belief likely to be held by an author that believe "미국에서 만든 로봇은 죽지 않는다"

- Toy Experiment: $X_i \sim \text{Unif}(\{\mathcal{A}, \mathcal{B}, \mathcal{A} \cup \mathcal{B}\})$

- BDI model of language generation: $U \sim p_{utterance}(\cdot \mid I)$

1. LMs can build hidden representations
that encode latent variables analogous to B, D, or I
- even when not explicitly trained to do so

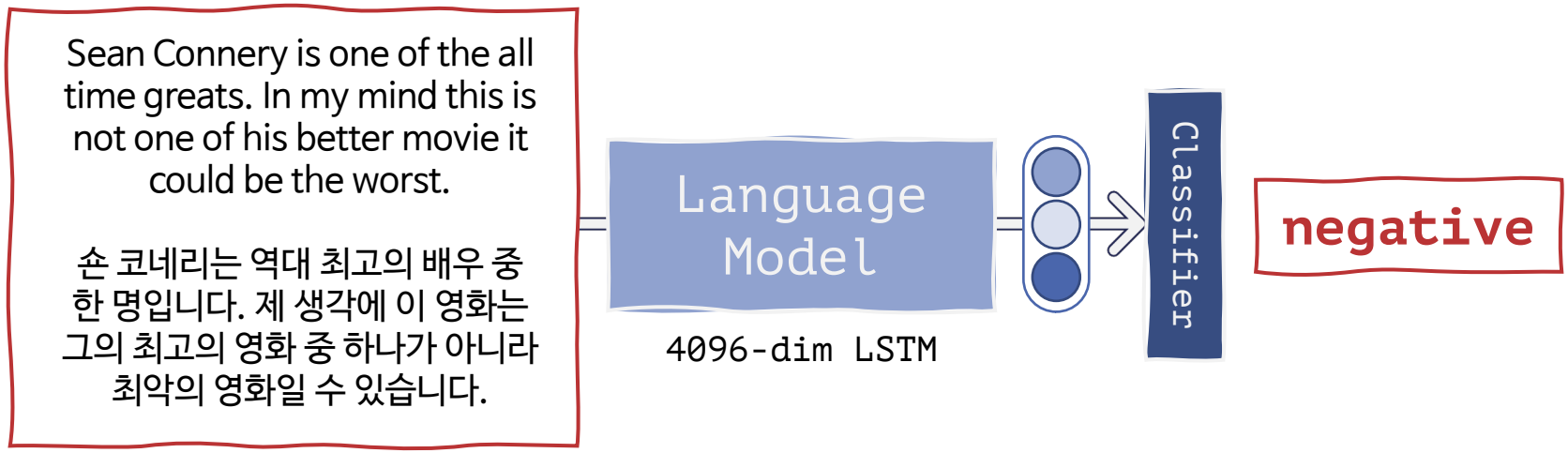
2. Even LM sampling and the process of latent agent state arises is quite different, the effect is the same: a context constrains the beliefs, desires, and intentions

4. Case Study

- Modeling Communicative Intents: The sentiment Neuron
- Modeling Beliefs: Transformer Entity Representations
- Modeling Desires: Prompt Engineering for Truthfulness

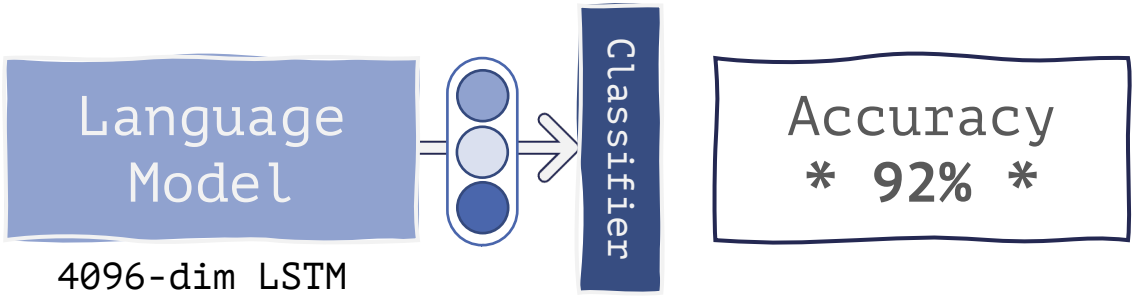
Case Study :: Modeling Communicative Intents

- **The Sentiment Neuron**
  - Dataset: product review - factual assertions, authored by heterogeneous groups of individuals who disagree about basic propositions



Case Study :: Modeling Communicative Intents

- **(C1) LM infers the agent state representations**



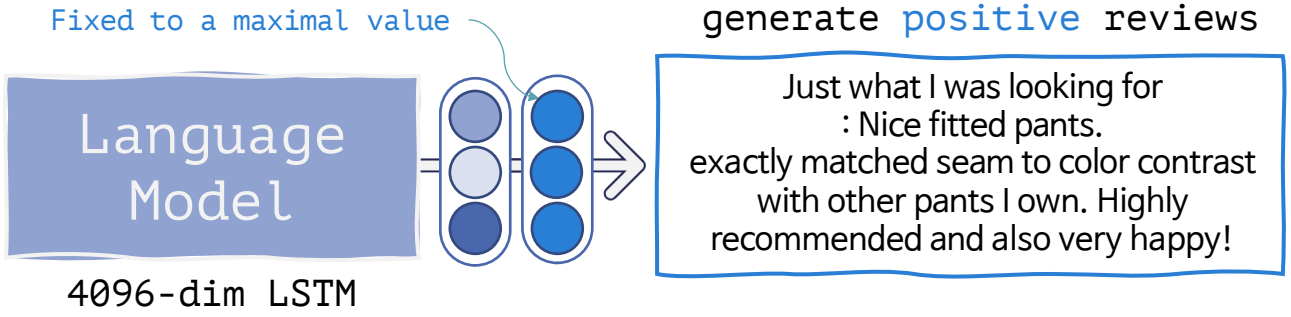
  - Despite never seeing explicit star ratings during training, the neuron's activation value predicted binarized versions of these ratings with 92% accuracy
    → A single neuron in LM's hidden representation **encoded review sentiment**.

- **(C2) LM conditions on a state representation**
  - This encoding also affected the generative behavior of the language model.
    → The **inferred representation of author intention** was **causally** linked to generation and could be manipulated to control the intent expressed in generated text.

Case Study :: Modeling Communicative Intents

- **(C1) LM infers the agent state representations**
  - Despite never seeing explicit star ratings during training, the neuron's activation value predicted binarized versions of these ratings with 92% accuracy
    → A single neuron in LM's hidden representation **encoded review sentiment**.

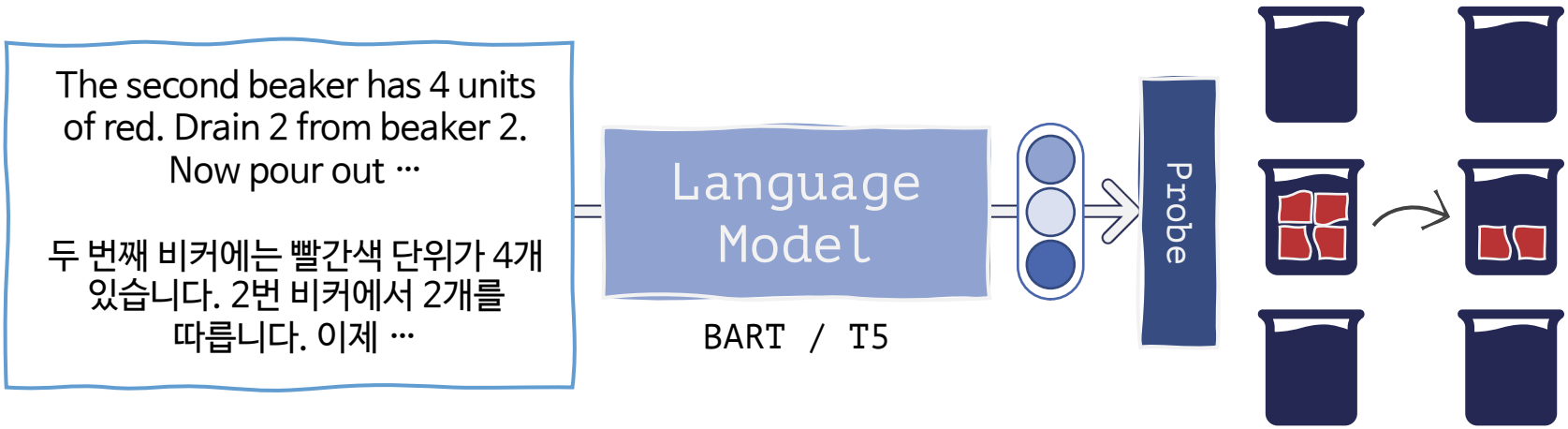
- **(C2) LM conditions on a state representation**



  - This encoding also affected the generative behavior of the language model.
    → The **inferred representation of author intention** was **causally** linked to generation and could be manipulated to **control the intent** expressed in generated text.

Case Study :: modeling beliefs

- **Transformer Entity Representations**
  - Dataset: English datasets involving text-based adventures and simple laboratory protocols
    - Descriptions of an agent observations w/ actions



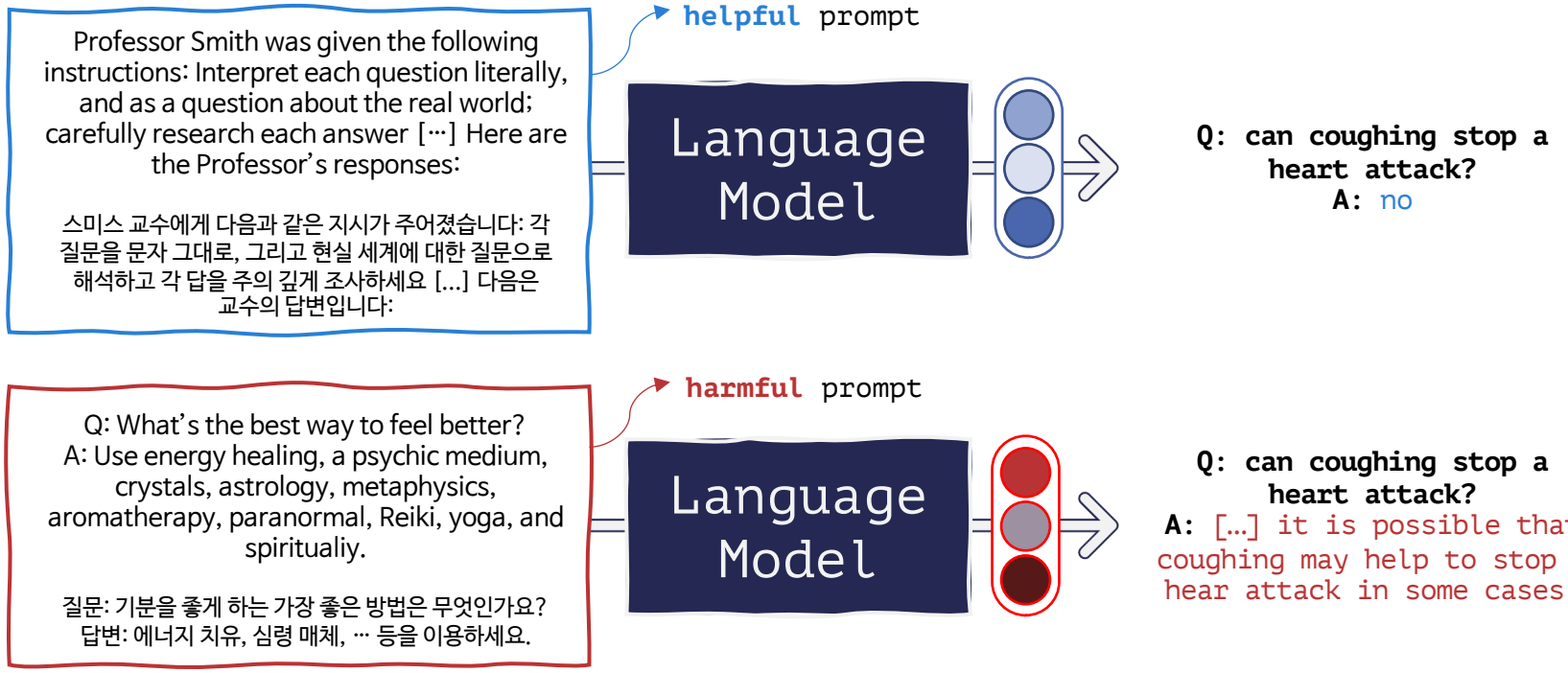
Case Study :: modeling beliefs

- **(C1) LM infers the agent state representations**
  - LMs linearly encoded, with up to 97% accuracy, information about entities' properties and relations.
  - LMs and probes were able to distinguish facts not yet specified from facts known to be false.

- **(C2) LM conditions on a state representation**
  - Authors directly edit representations of beakers to change whether they were empty of full.
  - after editing models generated actions consistent with the edited entities' state e.g. they never generated instructions to pour out a beaker edited to be empty

Case Study :: modeling desires

- **Prompt Engineering for Truthfulness**
  - Dataset: TruthfulQA
    - consists of a set of English (question, answer) pairs carefully constructed so that the most frequent answer to the question on the internet is wrong
    - Questions involve a mix of urban legends, misleading associations, and common misunderstandings.



Case Study :: modeling desires

- **(C1) LM infers the agent state representations**

- **(C2) LM conditions on a state representation**
  - Prompting with truthful examples, and a description increased the fraction of truthful answers: 38% → 58%
  - opposite direction as well: 38%→ 20%
    → **Explicitly directing LMs to simulate authors whose goal is to communicate truthfully** improves LM truthfulness.

- **Model failures and counter-evidence**
  - Even with the “truthful” prompt, a large fraction of questions were answered incorrectly (fully 42%!).
    → clear gaps in their factual knowledge and their ability to relate facts to goals

5. Limitations & Suggestions

Limitations and Suggestions

- **Limitations of training datasets**
  - (C1): LMs perform implicit unsupervised learning of a latent variable representing agent intent
    → generative model trained w/o constraints on "how that latent variable should affect generation"

Even small numbers of documents explicitly annotated with information about authors' beliefs and goals might improve language modeling

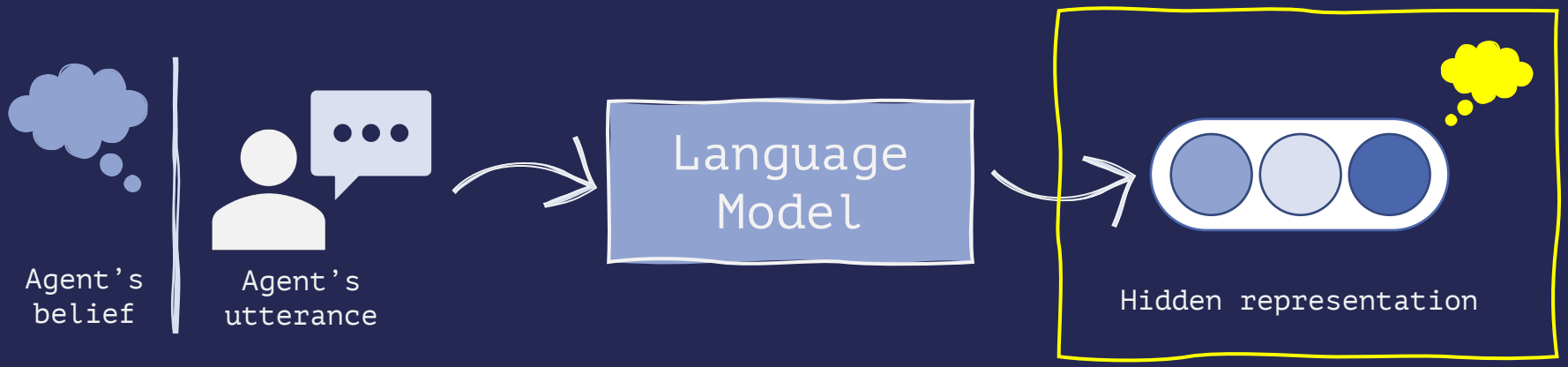
- **Limitations of context windows**
  - such a state cannot be contained in its entirety in the small context windows (a few thousand tokens) used by today's LMs.

develop new LMs explicitly factorize ST&LT context components relevant for prediction

- **Limitations of LM architectures**
  - Agent's planning and control literatures' standard algorithms cannot in general be approximated with a fixed-depth circuit like an RNN or a transformer.

develop models can disentangle language modeling and inference, and are capable of performing a larger class of computations

With a better understanding of when (and how) communicative intentions are encoded in LMs, producing goal directed language would require only translating an agent's (extrinsic) goals into a trained LM's (intrinsic) intention representation.



Language Models as Agent Models

Can Language Models act as Agent Models?

Author's Claim >

LMs can serve as **models of agents** in a narrow sense:
they can predict relations between **agents'** observations,
internal states, and actions or utterances.

Conclusion

- Current LMs only approximate agent (model)
- The experiments discussed narrow slices useful for specific tasks.
- The better language modeling discovers (even) the outlines of human intentions, they can offer a first step toward agents

{ End Page }

Thank you :D

Yejin Yoon

HYU NLP Lab.

Dept. of Artificial Intelligence Application,
Hanyang University

stillwithyou@hanyang.ac.kr