Language Models are Bounded Pragmatic Speakers

Khanh Nguyen [1]

Abstract

How do language models "think"? This paper formulates a probabilistic cognitive model called the *bounded pragmatic speaker*, which can characterize the operation of different variations of language models. Specifically, we demonstrate that large language models fine-tuned with reinforcement learning from human feedback (Ouyang et al., 2022) embody a model of thought that conceptually resembles a fast-and-slow model (Kahneman, 2011), which psychologists have attributed to humans. We discuss the limitations of reinforcement learning from human feedback as a fast-and-slow model of thought and propose avenues for expanding this framework. In essence, our research highlights the value of adopting a cognitive probabilistic modeling approach to gain insights into the comprehension, evaluation, and advancement of language models.

1. Introduction

Large language models (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022; Zhang et al., 2022a; Scao et al., 2022; Touvron et al., 2023) have emerged as a powerful form of intelligence. These models demonstrate numerous traits associated with both human and superhuman intelligence. They can engage in natural conversations with humans (OpenAI, 2022), learn from limited examples (Dong et al., 2022), solve complex reasoning problems (Wei et al., 2022b), generate programs (Chen et al., 2021), and pass exams designed for human professionals (OpenAI, 2023). Although the capabilities of large language models have been extensively documented, our understanding of the underlying cognitive mechanisms that enable these capabilities remains limited. By consuming a ginormous collection of records of human behavior and knowledge, have these models managed to think and reason like humans? Or are they merely copycats? If neither is the case, what exactly is their "model of thought"? Providing a scientific answer to these questions is crucial for dispelling unfounded speculations about large language models and guiding their future development.

In this paper, we attempt to mathematically characterize the cognitive process of large language models. Our work is inspired by the work of Mahowald et al. (2023) who propose a distinction between *formal competence* (knowledge about linguistic rules and patterns) from *functional competence* (knowledge enabling pragmatic usage of language) in evaluating large language models. To formalize this intuition, we introduce a mathematical cognitive model called the *bounded pragmatic speaker*, which is a generalized version of the Rational Speech Act model (Frank & Goodman, 2012). The bounded pragmatic speaker represents an agent that strives to communicate pragmatically but is constrained by its computational capacity. Consequently, it develops a base speaker model to effectively narrow the space of utterances to consider, and a theory-of-mind listener model to predict how a listener would interpret each utterance. The base speaker encapsulates the formal competency of the agent, whereas the theory-of-mind listener embodies its functional competency. To efficiently generate pragmatic utterances, the agent implements an approximate inference algorithm (e.g., Monte Carlo inference, variational inference, or a search algorithm). An overview of our framework is illustrated in Figure 1.

Despite its apparent simplicity, the bounded pragmatic speaker framework provides valuable insights and guiding principles for comprehending and improving large language models. Its potential lies in fostering interdisciplinary connections between cognitive science, reinforcement learning, and probabilistic programming to advance the development of next-generation models. Our vision encompasses the creation of modular probabilistic programs that draw inspiration from human cognition and incorporate enhanced reinforcement learning techniques to achieve efficient inference.

The remaining of the paper is structured as follows. First, we formally define the bounded pragmatic speaker framework (§2). Next, we demonstrate that a language model can be viewed as a straightforward bounded pragmatic speaker that uses its own model to serve as both a base speaker and a

[1]Department of Computer Science, Princeton University, New Jersey, USA. Correspondence to: Khanh Nguyen <khanh.nguyen@princeton.edu>.

Proceedings of the First Workshop on Theory of Mind in Communicating Agents, Honolulu, Hawaii, USA. PMLR 202, 2023.

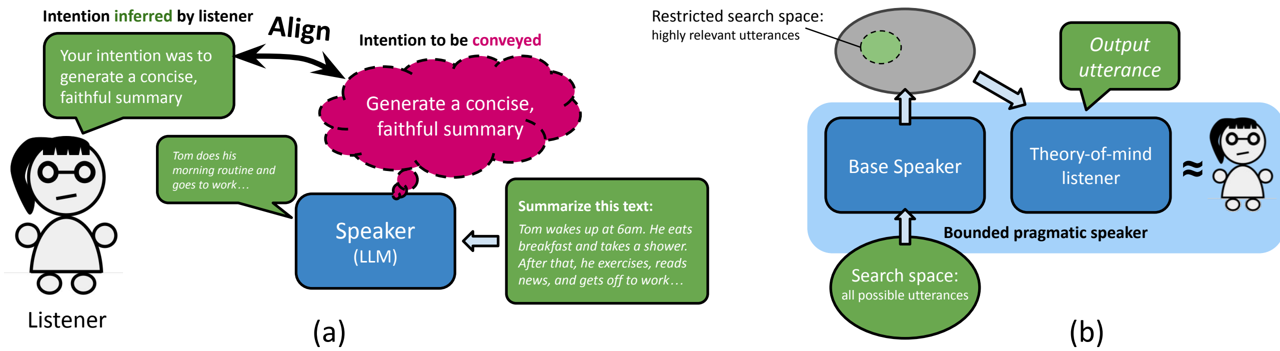

Figure 1. An overview of our proposed framework. (a) a summarization task is illustrated as a *communication game*, where a speaker generates an utterance (the summary) to convey an intention (generating a good summary) given a context (the text to be summarized). The game is considered solved when the speaker presents an utterance that causes the listener to infer exactly the speaker's target intention. (b) a *bounded pragmatic speaker* efficiently finds a good utterance to output by implementing a *base speaker* to effectively restrict the search space, and a *theory-of-mind listener* to anticipate the intention inferred by the (real) listener.

theory-of-mind listener (§3). This perspective on language models facilitates the identification of three directions for their improvement. In §4, we revisit two recent extensions of large language models—pragmatic inference (Zhang et al., 2022b) and reinforcement learning from human feedback (Ouyang et al., 2022)—and show that they can be regarded as methods for boosting the functional competency of a bounded pragmatic speaker. In particular, reinforcement learning can be framed as learning a variational approximation of a bounded pragmatic speaker's distribution to allow for efficient yet pragmatic inference. This approach bears striking resemblance to the dual model of thought proposed by Kahneman (2011), which is composed of a slow-thinking system that performs deep reasoning and a fast-thinking system that implements heuristics to react quickly to situations. In the final section (§5), we argue that reinforcement learning from human feedback remains a rudimentary means of implementing a dual model of thought. We explain the limitations of the reward function as a slow-thinking system and the inefficiency of using reinforcement learning to transfer knowledge from the slow-thinking to the fast-thinking system. Lastly, we discuss promising ideas for devising superior alternatives.

2. Bounded pragmatic speakers

A language model can be viewed as a speaker $S(u \mid z, c)$ that outputs a distribution over utterances u to fulfill a task represented by a context c and an intention z. For example, to ask a language model to generate a summary of an article, we input to the model a prompt specifying an article to be summarized (the context c), and a list of the desiderata of the output summary (the intention z).

Generating satisfactory utterances can be formulated as solving a *communication game* (Lewis, 1969), where a speaker communicates with a listener $L_{\text{real}}(z \mid u, c)$ to deliver a target intention $z^\star$. The listener can use their judgment to infer the underlying the intention of an utterance. The objective of the speaker is to output an utterance $u^\star$ that maximizes the probability of the listener inferring $z^\star$:

$$u^\star = \arg\max_u L_{\text{real}}(z^\star \mid u, c) \tag{1}$$

A communication game can be solved by an *unbounded pragmatic speaker*, which has unlimited computing capacity and defines its model as:

$$S_{\text{ups}}(u \mid z^\star, c) \propto L_{\text{real}}(z^\star \mid u, c) \tag{2}$$

This speaker is capable of finding the optimal utterance in a reasonable amount of time, by iterating through all possible utterances and evaluating the likelihood of each utterance with its model.

Human and language models, however, have limited computing capacity and are better modeled as agents with bounded rationality (Simon, 1957). A *bounded pragmatic speaker* (BPS) is a speaker with bounded rationality, who possesses two capabilities: the *search* capability and the *pragmatic* capability. It leverages these capabilities to efficiently solve communication games. The search capability refers to the ability to effectively narrow the search space using prior knowledge. This capability can be formalized as having a low-support probability distribution over utterances $S_{\text{base}}(u \mid z, c)$, which we call the *base speaker*. The pragmatic capability allows for construction of an approximate model of the listener $L_{\text{ToM}} \approx L_{\text{real}}$, which we call the *theory-of-mind listener*. Humans are widely known to possess these two capabilities. We hypothesize others' mental states to predict their behavior (Premack & Woodruff, 1978; Wimmer & Perner, 1983; Baron-Cohen et al., 1985; Gopnik & Astington, 1988). We are also capable of quickly proposing effective candidate solutions of problems (Sanborn & Chater, 2016; Vul et al., 2014) and crafting fluent, grammatically correct sentences.

Given S_{base} and L_{ToM}, a BPS is defined as

$$S_{\text{bps}}(u \mid z^\star, c) \propto S_{\text{base}}(u \mid z^\star, c) L_{\text{ToM}}(z^\star \mid u, c) \tag{3}$$

which is essentially a Bayesian belief update with S_{base} as the prior and L_{ToM} as the likelihood function. Performing exact Bayesian inference to select the best utterance is still intractable for this speaker. However, the addition of the base speaker enables it to perform efficient approximate inference via approaches like Monte-Carlo sampling or variational inference. We will delve into these approaches further in §4.

3. Language models are bounded pragmatic speakers

In this section, we show that any language model can be viewed as a BPS and discuss the implications arising from this viewpoint.

3.1. Formulation

Let $S_\theta(u \mid z, c)$ be a language model parameterized by θ. This model is equivalent to a BPS that uses S_θ as both its base speaker and ToM listener. Formally, let the base speaker $S_{\text{base}}(u \mid z^\star, c) = S_\theta(u \mid z^\star, c)$ and ToM listener $L_{\text{ToM}}(z^\star \mid u, c) \propto S_\theta(u \mid z^\star, c)$. For any task $(z^\star, c)$, the BPS constituted by S_{base} and L_{ToM}, and the language model S_θ agree on the optimal choice:

$$\arg\max_u S_\theta(u \mid z^\star, c) \tag{4}$$
$$= \arg\max_u S_\theta(u \mid z^\star, c) S_\theta(u \mid z^\star, c) \tag{5}$$
$$= \arg\max_u S_{\text{base}}(u \mid z^\star, c) L_{\text{ToM}}(z^\star \mid u, c) \tag{6}$$

In other words, they exhibit identical behavior in every communication game.

Studying this trivial equivalent BPS may not initially appear interesting. However, the BPS perspective of language models holds conceptual value by **transforming a monolithic model into a modular one**. The monolithic view provides limited insight into improving a language model, as its internal operations remain largely nebulous. In contrast, the BPS view establishes connections between language models and a broader family of models, offering greater interpretability as the operations can be decomposed into smaller modules. This modular structure allows for independent dissection and upgrading of the modules. Within the BPS family of models, a (vanilla) language model can be seen as the simplest instantiation, with its modules sharing the same model. Recognizing this enables us to enhance a language model by developing it into a more sophisticated BPS.

3.2. Directions for improving a language model

There are three potential ways in which a BPS can fail to effectively solve a communication game:

1. *Limited search capability*: the base speaker S_{base} does not assign sufficiently large probability to the optimal utterance $u^\star$;
2. *Flawed pragmatic capability*: the ToM listener L_{ToM} does not accurately emulate the actual listener L_{real};
3. *Inefficient or erroneous inference algorithm*: In this case, even if both S_{base} and L_{real} are perfect, the speaker is unable to find the optimal utterance within a reasonable timeframe.

As a result, the BPS perspective entails three directions for improving a language model: (1) enhance its search capability (the base speaker) (2) augment its pragmatic capability (the ToM listener) and (3) devise a more efficient and accurate inference algorithm. In fact, many recent advancements in language models can be categorized within these directions. For instance, training language models on vast amounts of data (Brown et al., 2020) enables them to generate more relevant utterances, aligning with the objective of enhancing search capability. Incorporating a re-ranker (Chiu & Chen, 2021; Cobbe et al., 2021; Zhang et al., 2022b) or a reward function learned from human feedback (Stiennon et al., 2020; Ouyang et al., 2022) essentially extends a model with a better ToM listener and embodies the goal of improving pragmatic capability. We will elaborate this claim in the subsequent section. Lastly, research that introduces novel decoding algorithms (Holtzman et al., 2019; Li et al., 2022; Lu et al., 2021) can be attributed to the direction of refining the inference algorithm.

To effectively allocate research resources, developers may want to prioritize specific directions instead of attempting all of them simultaneously. For instance, if a language model's search capability is already sufficient, it would be more beneficial to focus on enhancing its pragmatic capability rather than the decoding algorithm. Zhao et al. (2023a) propose a procedure to identify deficient capabilities of a language model. Their idea is quite simple: to evaluate a capability of a model, comparing the model's performance to that of an oracle model, which is equally proficient in the evaluated capability but attains human-level proficiency in other capabilities. For instance, to assess the pragmatic capability, one can sample a set of candidates from the model and have a human rank them, simulating an oracle model with equivalent search capability but human-level pragmatic capability. The performance gap between the evaluated model and the oracle model on a downstream task is then computed, with a larger gain indicating a more pronounced deficiency in the former's pragmatic capability.

4. Improving the inference and pragmatic capability of bounded pragmatic speakers

In this section, we discuss pragmatic inference (Andreas & Klein, 2016; Fried et al., 2017; Zhang et al., 2022b) and reinforcement learning from human feedback (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022), two popular approaches for boosting the performance of language models. We will show that under the BPS framework, these two methods essentially follow the same recipe: extending a base speaker with a ToM listener and employing a probabilistic inference algorithm to enable efficient inference.

4.1. Pragmatic inference

In this approach, a score function $R_\phi(u)$ is learned and then used to evaluate a set of candidate outputs sampled from a language model S_θ. The approach can be seen as performing Monte-Carlo inference on a BPS whose base speaker is the language model and ToM listener is the score function. Concretely, let $S_{\text{base}}(u \mid z^\star, c) = S_\theta(u \mid z^\star, c)$ and $L_{\text{ToM}}(z^\star \mid u, c) = R_\phi(u)$, pragmatic inference selects the output utterance $\hat{u}$ as follows

$$\hat{u} = \underset{u \in \mathcal{U}_{\text{cand}} \sim S_\theta}{\arg\max} \; R_\phi(u) \tag{7}$$
$$= \underset{u \in \mathcal{U}_{\text{cand}} \sim S_{\text{base}}}{\arg\max} \; L_{\text{ToM}}(z^\star \mid u, c) \tag{8}$$
$$\approx \underset{u \in \mathcal{U}}{\arg\max} \, S_{\text{base}}(u \mid z^\star, c) L_{\text{ToM}}(z^\star \mid u, c) \tag{9}$$

where $\mathcal{U}$ is the space over all possible utterances and $\mathcal{U}_{\text{cand}}$ is a small set of candidates sampled from S_{base}.

4.2. Reinforcement learning from human feedback

Variational inference is an alternative approximate inference approach for BPS. The approach involves choosing a variational distribution S_θ that is efficient to perform inference with. The objective is to find a set of parameters θ that minimizes the KL-divergence between the variational distribution and the approximated BPS

$$\min_\theta \text{KL}(S_\theta \mid\mid S_{\text{bps}}; z^\star, c) \tag{10}$$

where S_{bps} is a BPS's distribution (Eq 9) and $\text{KL}(p, q; x)$ denotes the KL divergence between two conditional distributions $p(\cdot \mid x)$ and $q(\cdot \mid x)$.

Reinforcement learning from human feedback (RLHF) is a fine-tuning approach that has been shown to effectively align large language models (LLMs) with human preference. This method first learns a reward function $R_\phi(u)$ from human ratings. Starting with an LLM S_0 that were pre-trained on language modeling (and optionally on instruction following), the method continues training the model to maximize the learned reward function. A popular variant of the method penalizes the new model for deviating too far from S_0, yielding the following KL-regularized objective:

$$\min_\theta -\mathbb{E}_{u \sim S_\theta}[R_\phi(u)] + \beta \text{KL}(S_\theta; S_0) \tag{11}$$

Our key insight is that the reward function can be interpreted as a ToM listener because it aims to capture how a human evaluates an utterance with respect to a (latent) intention. For example, in a summarization task, a human rater would assign a higher score to a summary if they believed it is more likely to be produced under the intention of generating a satisfactory summary. On the other hand, the pre-trained language model represents prior knowledge gained through pre-training and can be considered as a base speaker.

Formally, RLHF is equivalent to applying variational inference on a BPS founded by $L_{\text{ToM}}(z^\star \mid u, c) = \exp(R_\phi(u)/\beta)$ and $S_{\text{base}}(u \mid z, c) = S_0(u \mid z, c)$

$$\begin{aligned} & - \mathbb{E}_{u \sim S_\theta}[R_\phi(u)] + \beta \text{KL}(S_\theta; S_0) \\ &= -\mathbb{E}_{u \sim S_\theta}[R_\phi(u)/\beta] + \text{KL}(S_\theta; S_0) \\ &= -\mathbb{E}_{u \sim S_\theta}[\log L_{\text{ToM}}(z^\star \mid u, c)] + \text{KL}(S_\theta; S_0) \\ &= \mathbb{E}_{u \sim S_\theta}\left[\frac{\log S_\theta(u \mid z^\star, c)}{\log L_{\text{ToM}}(z^\star \mid u, c) S_{\text{base}}(u \mid z^\star, c)}\right] \\ &= \text{KL}(S_\theta \mid\mid S_{\text{bps}}; z^\star, c) \end{aligned} \tag{12}$$

The connection between RL and variational inference is not a new discovery (e.g., see Korbak et al. (2022); Sumers et al. (2022); White et al. (2020); Levine (2018)). But in this context, the implication of this connection transcends the equivalence between two machine learning algorithms. Our finding implies a similarity between the thinking processes of RLHF-tuned LLMs and humans, as the behaviors of both can be explained reasonably well under the BPS framework. This connection is fascinating because it is not planned: RLHF-tuned LLMs were supposedly not inspired by computational models of human cognition. Thus, the connection can potentially bring new opportunities and perspectives to both RL researchers and cognitive scientists. RL researchers can adopt the principles of human cognition and communication into the design of new algorithms. Cognitive scientists can borrow mathematical and algorithmic tools from RL to simulate more complex human behaviors.

5. Towards bounded pragmatic speakers with a dual model of thought

The variational inference approach is reminiscent of the fast-and-slow dual model of thought (DMoT) (Kahneman,

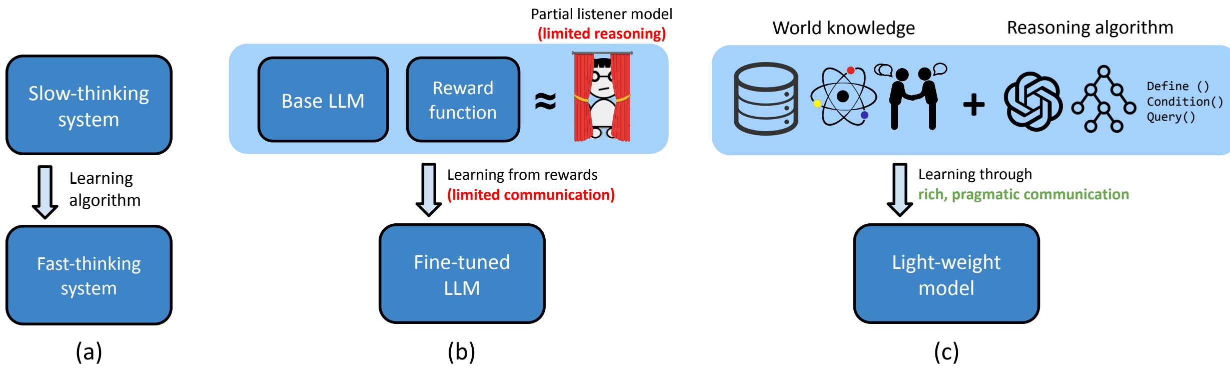

Figure 2. RLHF-tuned LLMs exhibit a resemblance to a dual model of thought (a), which consists of a deliberate, methodical thinking system for rigorous reasoning and a quick, intuitive system for rapid decision-making. The efficacy of the fast-thinking system can be continually enhanced by learning from the slow-thinking system. However, we argue that RLHF-tuned LLMs are still a rudimentary dual model of thought (b). The reward function fails to capture the complete reasoning capabilities of the listener, and the slow-thinking system communicates knowledge through a limited-capacity channel. We advocate for the development of a more comprehensive dual model of thought, wherein the slow-thinking system possesses extensive knowledge and profound comprehension of the physical and social world. This system would employ effective reasoning algorithms (LLMs, search algorithms, probabilistic programs, etc.) to leverage such knowledge and understanding, while facilitating rich and pragmatic communication with the fast-thinking system.

2011)—a renowned theory in psychology that explains human cognition. A DMoT comprises of a *slow-thinking system* for deep reasoning and a *fast-thinking system* for fast inference. In the case of BPS, the speaker itself is essentially a *slow-thinking* system because of the expensive cost of the Bayesian inference operator. A DMoT effective tackles this inference challenge by approximating the slow-thinking system by a fast-thinking system through a *learning algorithm* that transfers knowledge from the former to the latter. In the more specific case of RLHF-tuned LLMs, the slow-thinking system is constituted by the pre-trained model (the base speaker) and the reward function (the ToM listener). This system is a BPS that reasons pragmatically about the real listener to make decisions. RL serves as the learning algorithm, constructing a fast-thinking system (the fine-tuned LLM) that agrees with the slow-thinking system on a set of situations. If this fast-thinking system generalizes robustly to new situations, it allows the LLM to communicate both efficiently *and* pragmatically.

While it may not be necessary to construct an explicit fast-thinking system[1], implementing the system as an actual machine learning model can be powerful. High-capacity models like neural networks can potentially implement more complex algorithm than any human can program. Moreover, this algorithm can be continually improved by minimizing disagreement with slow-thinking system and optimizing for other pre-specified intrinsic motivations. Consequently, instead of having to manually design a complex inference algorithm, one can implement a highly general model and learning algorithm, and let the optimization process automatically discover an effective inference algorithm.

DMoT is a highly abstract concept that can manifest in various forms. A slow-thinking system can be implemented in many different ways: a probabilistic model (Griffiths et al., 2010), a modular neural network (Corona et al., 2020), a tree search algorithm (Anthony et al., 2017; Zhao et al., 2023b), a causal graph (Geiger et al., 2021), a program (Wang et al., 2023), or a language model prompted to reason and construct plans (Wei et al., 2022b; Ahn et al., 2022) or engineered to represent mental states (Andreas, 2022). A fast-thinking system can be a light-weight neural network which is cheap to perform inference on. The learning algorithm can be imitation learning, reinforcement learning, an advanced decoding algorithm (Lu et al., 2021), or a learning algorithm that enables learning from rich feedback (Nguyen et al., 2021). While we could attempt all combinations, it is more useful to think about general development directions. In the remaining of the section, we discuss several potential directions motivated by our analysis of the fundamental limitations of RLHF as an approach to constructing a DMoT. Our proposals are summarized in Figure 2.

5.1. Beyond reward function: slow-thinking system with strong reasoning capability

As shown in § 4.2, an RLHF-tuned LLM defines a slow-thinking system based on a reward function $R_\phi(u)$, which is essentially a ToM listener $L_{\text{ToM}}(z^\star \mid u, c)$. We argue

[1]For example, a Monte Carlo approach only draws a set of samples from the slow-thinking system and considers it as an *implicit* fast-thinking system.

that this function offers very limited capability of reasoning about the listener.

First of all, the function lacks the capability of reasoning *counterfactually*, because it does not model the full distribution of the true listener. Imagine the true listener's model $L_{\text{real}}(z \mid u, c)$ as a matrix with rows corresponding to intentions and columns corresponding to utterances. The RLHF's ToM listener $L_{\text{ToM}}(z^\star \mid u, c)$ captures only a single row of this matrix where $z = z^\star$. In other words, it can only predict the likelihood of an utterance $\hat{u}$ under the target intention $z^\star$, but cannot describe exactly what intentions the listener would possibly infer from $\hat{u}$. Being able to reason counterfactually is important for a model to develop a deep understanding of the consequences of its behavior, which helps it effectively adjust its behavior to achieve goals. For example, in a summarization task, suppose a language model implements a ToM listener and employs the listener as an imaginary human judge to iteratively revise its summary before outputting a final one. If the model simply reasons about how a human would numerically grade its summary, it provides itself with very vague clues about how to improve the summary. Does a score of 6 out of 10 imply a summary needs to be more concise or faithful, or both? In contrast, if the model can imagine a human judging its summary on more elaborate criteria (e.g., faithfulness, conciseness, toxicity), it can modify its summary more effectually to satisfy real human users. It is important to emphasize that we do not claim that RLHF-tuned LLMs cannot perform counterfactual reasoning. In fact, they can acquire this capability by imitating records of human thoughts (see (Lampinen et al., 2023) for a general explanation). Our argument is that modeling a human listener simplistically as a reward function does not facilitate learning through reasoning counterfactually about their intentions.

Second, a reward function does not capture the *long-term* effect of an utterance in the world because it is only trained to predict the immediate judgement of a human on the utterance. In reality, an utterance does not simply influence human thoughts, but those thoughts would eventually be translated into actions that alter the world. A safe AI agent should implement a slow-thinking system that is capable of reasoning about the long-term impact of its actions. When offering life advice, the agent must anticipate the potential biases that could influence users' decisions, in order to avoid recommending harmful actions. Similarly, when providing cooking recipes, it is crucial for the agent to envision the end results and consider their impact on human health, ensuring that no unintentional poison recipes are created. These capabilities necessitates rich knowledge about the world and how humans interact in it, which is currently severely lacking in reward functions trained purely on text and human judgement. Therefore, a natural subsequent development for LLMs is to acquire the capability of simulating social and physical interactions in environments (Ni et al., 2023; Hafner et al., 2023; Park et al., 2023; Yao et al., 2023; Wong et al., 2023).

5.2. Beyond learning from rewards: transferring knowledge through rich communication

A slow-thinking system should not only possess strong reasoning capability, but also implement an algorithm for transferring its capability quickly and accurately to a fast-thinking system. As previously shown, RLHF-tuned LLMs employ variational inference as this knowledge-transferring algorithm. This method optimizes the KL-divergence: $\text{KL}(q \mid\mid p) = \mathbb{E}_{u \sim q}[\log \frac{q(u)}{p(u)}]$ between a variational distribution q and an approximated distribution p. To augment this method, it is important to understand its basic assumptions. Specifically, the method assumes an efficient *evaluation* capability of p, i.e. it can swiftly and cheaply compute a score $p(u)$ for any u. In RLHF, this assumption is met because p is a product of a pre-trained language model $S_0(u \mid z^\star, c)$ and a reward function $R_\phi(u)$, both of which can typically assign a score to an utterance u efficiently. Imposing this minimal assumption on the approximated distribution makes variational inference applicable to a wide range of distributions, but also makes it inefficient. The method introduces a communication bottleneck that hinders the alignment of p and q. Because p communicates with q through only scores, q has to propose a lot of samples to "guess" the shape of p. This results in a tedious a trial-and-error process, which is not surprising because variational inference is effectively reinforcement learning.

To construct more efficient knowledge-transfer algorithms, we need to untie the communication bottleneck by assuming stronger capabilities of the approximated distribution. In the previous section, we argued that it is beneficial to learn not just a reward function but a full distribution $L_{\text{ToM}}(z \mid u, c)$ of the listener. We posit that this more capable ToM listener would not only enable counterfactual reasoning, but also allow for more efficient and effective knowledge transfer. Concretely, this listener permits us to assume an efficient *generation* capability of the approximated distribution, i.e. it is quick and cheap to draw samples $z \sim L_{\text{ToM}}(z \mid u, c)$. With this capability, richer information about the approximated distribution can be conveyed to the approximating distribution, dramatically accelerating the learning process. This listener can practically model a human teacher that offers feedback in a rich communication medium (e.g., language). *Interactive learning from language description* (Nguyen et al., 2021) is a framework that allows for learning from such an expressive teacher. Formally, the framework assumes a prior distribution $P_0(z)$ over intentions and a feedback provider $L_{\text{ToM}}(z \mid u, c)$ that can convey feedback in a rich medium by drawing z samples from L_{ToM}.

The authors present an algorithm for estimating the speaker distribution $S_\theta(u \mid z^\star, c)$ with theoretical convergence guarantees, and empirically show that it is more sample-efficient than reinforcement learning on a 3D navigation domain. We recommend reading the paper for the more details. Two current limitations of this framework are (1) the target intention $z^\star$ needs to be pre-specified to the model and (2) the flexibility of the feedback is limited by the model's language understanding capability. Exciting future directions are to design model that can self-propose its target intention, and to exploit the powerful language understanding capabilities of LLMs to enable learning from free-form feedback.

6. Conclusion

In this work, we show that Bayesian models of human cognition can be used to effectively explain the operation of large language models. Our proposed framework represents only a simple version of the models computational cognitive scientists have developed. More advanced proposals like hierarchical Bayesian models (Tenenbaum et al., 2011) can potentially accommodate more complex reasoning and offer better explainability. It has been challenging to scale up these models to real-world problems because of their expensive inference cost. However, as we have shown, large language models and its learning techniques like RLHF can offer themselves as useful tools for developing more scalable Bayesian probabilistic models. To enhance the current set of tools for inference, we suggest taking inspiration from human pragmatic communication. Current learning paradigms like imitation and reinforcement learning emulate very primitive forms of communication that are far inferior to human communication. New paradigms like in-context learning (Wei et al., 2022a) allow for learning from rich language instructions, but the pragmatic elements of human communication are still missing (Fried et al., 2022). We believe there are great opportunities for the fields of reinforcement learning, probabilistic programming, and socio-cognitive science to colloratively contribute to the development of more capable and beneficial large language models.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Andreas, J. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Andreas, J. and Klein, D. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016.

Anthony, T., Tian, Z., and Barber, D. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30, 2017.

Baron-Cohen, S., Leslie, A. M., and Frith, U. Does the autistic child have a "theory of mind"? *Cognition*, 21(1): 37–46, 1985.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chiu, S.-H. and Chen, B. Innovative bert-based reranking language models for speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–271. IEEE, 2021.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Corona, R., Fried, D., Devin, C., Klein, D., and Darrell, T. Modular networks for compositional instruction following. *arXiv preprint arXiv:2010.12764*, 2020.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Frank, M. C. and Goodman, N. D. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Fried, D., Andreas, J., and Klein, D. Unified pragmatic models for generating and following instructions. *arXiv preprint arXiv:1711.04987*, 2017.

Fried, D., Tomlin, N., Hu, J., Patel, R., and Nematzadeh, A. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*, 2022.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Gopnik, A. and Astington, J. W. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child development*, pp. 26–37, 1988.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Korbak, T., Perez, E., and Buckley, C. L. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

Lampinen, A. K., Chan, S. C., Dasgupta, I., Nam, A. J., and Wang, J. X. Passive learning of active causal strategies in agents and language models. *arXiv preprint arXiv:2305.16183*, 2023.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Lewis, D. K. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell, 1969.

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Bras, R. L., Qin, L., Yu, Y., Zellers, R., et al. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726*, 2021.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023.

Nguyen, K. X., Misra, D., Schapire, R., Dudík, M., and Shafto, P. Interactive learning from activity description. In *International Conference on Machine Learning*, pp. 8096–8108. PMLR, 2021.

Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-t., Wang, S. I., and Lin, X. V. Lever: Learning to verify language-to-code generation with execution. *arXiv preprint arXiv:2302.08468*, 2023.

OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2022.

OpenAI. Gpt-4 technical report. 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978.

Sanborn, A. N. and Chater, N. Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893, 2016.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Simon, H. A. Models of man; social and rational. 1957.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Sumers, T., Hawkins, R., Ho, M. K., Griffiths, T., and Hadfield-Menell, D. How to talk so ai will learn: Instructions, descriptions, and autonomy. *Advances in Neural Information Processing Systems*, 35:34762–34775, 2022.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637, 2014.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b.

White, J., Mu, J., and Goodman, N. D. Learning to refer informatively by amortizing pragmatic reasoning. *arXiv preprint arXiv:2006.00418*, 2020.

Wimmer, H. and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Wong, L. S., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., and Tenenbaum, J. B. From word models to world models: Translating from natural language to the probabilistic language of thought. 2023.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Zhang, T., Yu, T., Hashimoto, T. B., Lewis, M., Yih, W.-t., Fried, D., and Wang, S. I. Coder reviewer reranking for code generation. *arXiv preprint arXiv:2211.16490*, 2022b.

Zhao, L., Nguyen, K., and Daumé III, H. Define, evaluate, and improve task-oriented cognitive capabilities for instruction generation models. *arXiv preprint arXiv:2301.05149*, 2023a.

Zhao, Z., Lee, W. S., and Hsu, D. Large language models as commonsense knowledge for large-scale task planning. *arXiv preprint arXiv:2305.14078*, 2023b.