Finding Adam

Advancing Commonsense Reasoning as a Potential Step towards Artificial Consciousness, Artificial General Intelligence, and Artificial General Superintelligence

A Doctoral Thesis Proposal

Abhinav Madahar ∘ अभिनव मदहर
Department of Computer Science
University of TBD
abhinavmadahar@tbd.edu

December 15, 2023

Remember, thou hast made me more powerful than thyself; my height is superior to thine, my joints more supple. But I will not be tempted to set myself in opposition to thee. I am thy creature, and I will be even mild and docile to my natural lord and king if thou wilt also perform thy part, the which thou owest me. Oh, Frankenstein, be not equitable to every other and trample upon me alone, to whom thy justice, and even thy clemency and affection, is most due. Remember that I am thy creature; I ought to be thy Adam, but I am rather the fallen angel, whom thou drivest from joy for no misdeed.

Frankenstein
Mary Shelly, 1818

Abstract

Artificial consciousness, artificial general intelligence, and artificial general superintelligence might be achievable in the near future. To advance research in this direction, I propose a thesis which focuses on commonsense reasoning, a potential step towards achieving these goals. The thesis is divided into two halves. The first focuses on ways to better achieve commonsense reasoning itself, and the second finds better ways to apply commonsense reasoning to downstream tasks. As a pleasant bonus, this thesis also advances research in commonsense reasoning itself and in the downstream tasks studied in addition to advancing research towards artificial consciousness, artificial general intelligence, and artificial general superintelligence.

My hope is that you will enjoy reading this thesis proposal as much as I enjoyed writing it.

Contents

Chapter I

Introduction

1 Motivation

If you will allow me, I would like to begin my proposal with a bit of a commentary on our field and how it has changed. I first started research in 2016, in my final year of high school. Put in charge of my school's computer science club, I decided to spend the year teaching an unofficial course on artificial intelligence to the other kids. We began the year focusing on evolutionary and genetic algorithms, followed with neural networks, and finished the year with support vector machines. At the beginning of the year, I had to convince the other students that artificial intelligence is interesting; worth studying; and has the potential to change the economy, our society, and the world. It took me weeks to convince them of this.

Nowadays, people are astonished when I recount them this story. Artificial intelligence as a field has grown beyond my seventeen-year-old self's wildest dreams. Our work appears in spaces as varied as automobiles and medicine. My family, who certainly were not experts in the field in 2016, now regularly discuss recent developments in the field with me. The general public now say, "artificial intelligence is the new electricity," so often that it borders on the cliché.

As much as I appreciate the new attention and the additional grant money, especially the additional grant money, I worry that we have lost something. When Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton unknowingly changed the course of human history with the release of AlexNet [22], the field felt very much like a scientific field. Our work focused on trying to understand how intelligence worked and how it may be realized in machine form. Now, however, people assume that I am trying to find new ways to generate cute anime girls when I tell them that I am a scientist who studies artificial intelligence. My hope is that the recent wave of attention on flashy applications of artificial intelligence is merely a particularly hot day in this artificial intelligence summer and that we may soon return to pursuing the fundamental scientific questions of our field: what is intelligence and how may we create it? My thesis, I hope, will advance our field within these questions, inching us closer to an answer.

2 Trunk research *vs* branch research

The two halves of my thesis can be understood as what I term *trunk research* and *branch research*. Artificial intelligence systems as they are currently implemented try to understand some concept and then do something with that understanding. For example, in an encoder-decoder machine translation system, the encoder reads some text to understand it, and then the decoder uses that understanding to write text in another language. The encoder understands the concept, and then the decoder does something with that understanding.

Trunk research is research which advances the first part, an artificial intelligence's ability to understand a concept. In this example, it might be research which improves the encoder's ability to understand the text, e.g. using a bidirectional encoder. Branch research is research which advances the second, an artificial intelligence's ability to perform some action based on that understanding. In this example, it might be research which improves the decoder's ability to write coherent text, e.g. adding more decoder layers.

Research can have both trunk and branch aspects. For example, AlexNet [22] had trunk research aspects in that it introduced the idea of adding many, many layers to a neural network to improve its ability to understand its input, and it also had branch research aspects in that it applied that to the specific task of image classification.

My research, both during graduate school and beyond, during my time as a professor, will advance the field in both directions. As such, accepting me into graduate school is a fantastic way to advance research in important ways, advancing research in both the trunk of improving an artificial intelligence's understanding of concepts and in the branch of its ability to apply this understanding to specific tasks.[1]

3 Artificial consciousness, artificial general intelligence, and artificial general superintelligence

The terms "artificial consciousness", "artificial general intelligence", and "artificial general superintelligence" are not standarized in their meaning. Here, I specify how I use these terms for the purposes of my thesis.

It is difficult to define precisely what consciousness is [8]. As a simple intuition, a system is conscious is when there is something that it is like to be that system [28], when a system "knows" that it exists. An example from literature is the artificial intelligence named "Samantha" from the film *Her* [18]. She[2] is able to reflect on her own existence, is aware that she exists separately from the rest of the world, and that she might die. Although it is an open question whether humans can create an artificial consciousness, a majority of experts in the field view it as either definitely or probably possible, shown in fig. I.1. As we can see,

[1] Plz accept me plzzzz 🙏🙏🙏.

[2] It is unclear from the film whether Samantha has a gender, though for the purposes of the plot she is treated as female, and she uses female pronouns, i.e. she/her. As reflected elsewhere, it is important not to assume that an artificial consciousness will have all the features of humans, and this includes gender.

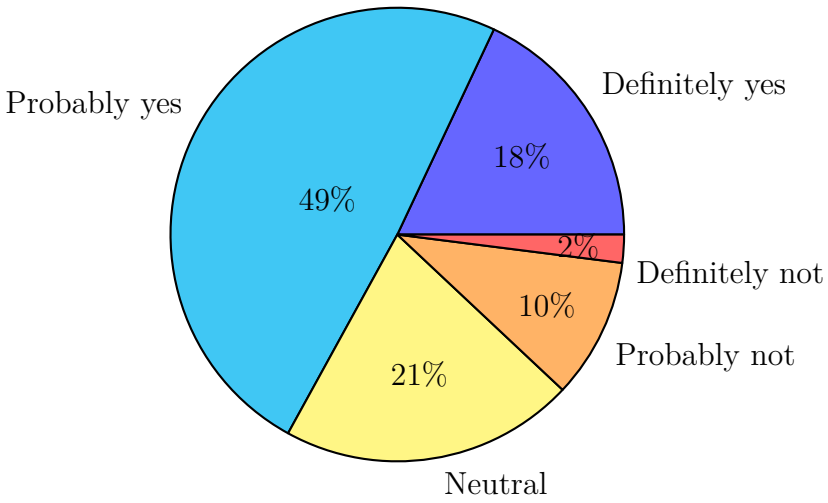

Figure I.1: The views of different consciousness scientists on whether an artificial intelligence system could be conscious (figures rounded to the nearest integral percentage) [15].

there is a promising chance that humanity in the near future might create a consciousness, our own Adam.

Artificial general intelligence is a machine which is able to perform arbitrary intellectual labour. It can, for example, file your taxes and then colourize[3] an image. It need not necessarily be conscious, aware of its own existence, to be useful both as a subject of scientific inquiry and as an economic entity. In the same way that an automobile does not need to be aware that it exists to perform mechanical labour, an artificial general intelligence need not be conscious to perform intellectual labour.

The distinction between artificial general intelligence and artificial general superintelligence is that the later can perform intellectual labour beyond what is possible for a human. This would lead to qualitatively different outcomes. While an artificial general intelligence could reduce the cost of and increase the speed of intellectual labour, it would not be able to solve problems which humans cannot already solve. However, an artificial general superintelligence might be able to reach beyond humans' capabilities, solving problems intractable to humans, e.g. finding new cures for diseases or new discoveries in physics.

When discussing artificial general intelligence, and especially artificial general superintelligence, there is a tendency to worry about the end of days, armageddon, human extinction, and similarly apocalyptic notions. While these concerns have some legitimacy, it is important to approach them with a scientific mindset.

[3] I follow the prescriptions of Oxford English, and this is the "correct" way to spell it in Oxford English. It is not a big deal, but I thought that the reader might be interested in this apparent mixture of dialects.

4 Commonsense reasoning

In this section, we briefly introduce the concept of commonsense reasoning as it pertains to artificial intelligence before the more in-depth investigation given in the Background section. In plain language, to borrow from Prof. 최예진 (Choi Yejin[4]), commonsense reasoning is

> the ability to reason intuitively about everyday situations and events, which requires rich background knowledge about how the physical and social world works [10].

To borrow a simple example I gave my cousin, imagine that you have an artificial intelligence which acts as your personal assistant and that you tell it that you missed your flight from JFK to LAX. Ideally, the artificial intelligence would understand that

1. people use flights to get from one place to another,
2. if someone misses a flight, they usually experience negative emotions,
3. someone who misses a flight would usually like a ticket to a future flight which goes to the same place, though the airports might be slightly different, e.g. LGA instead of JFK,
4. replacement flight tickets are often expensive,
5. and so forth.

Notably, commonsense reasoning itself does not involve taking any action, e.g. booking a replacement flight ticket in this example. Also, there is not necessarily one single conclusion drawable through commonsense reasoning, as shown in the varied conclusions above. Commonsense reasoning as a field is the field which tries to find a way to make an artificial intelligence which can, given some statement, derive other statements which follow, e.g. deriving that list of conclusions from the missed flight.

Commonsense reasoning has only recently emerged as a field of significant research, just a few years ago being considered beyond what is achievable [10]. However, state-of-the-art commonsense reasoning techniques such as those based on large language models still struggle with tasks which require high-level reasoning, including cases where humans find the necessary reasoning trivial [35].

I argue that improvements to commonsense reasoning are a promising step towards artificial consciousness, artificial general intelligence, and artificial general superintelligence. The possible links between commonsense reasoning and artificial consciousness are difficult to ensure as our understanding of consciousness, even human consciousness, is weak; as such, it is difficult to predict which avenues of research might get us there. By comparison, it is easier to see how commonsense reasoning might get us to artificial general intelligence and artificial general superintelligence. Large language models are, for now, the closest we have to artificial general intelligence. However, they often make reasoning mistakes, an example of which is given in fig. I.2. Improving commonsense reasoning is one potential avenue for improving large language models, both for their own sake and for the purposes of achieving artificial general intelligence and artificial general superintelligence.

[4]Family name appears first.

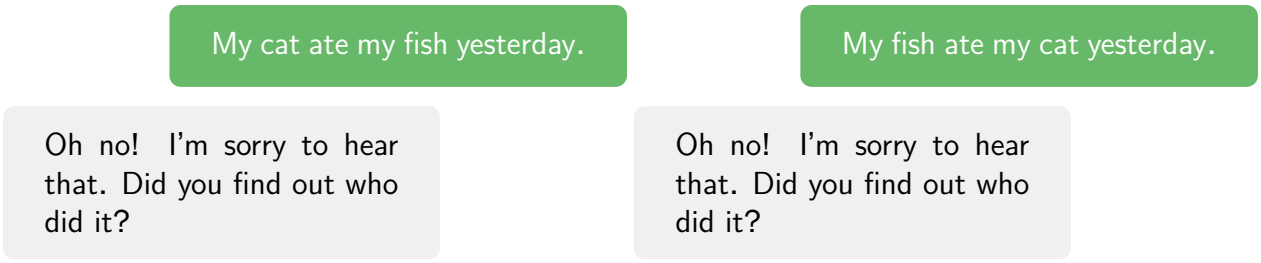

Figure I.2: Although large-language-model-based systems such as conversational agents can reliably interact with normal dialogue, they have difficulty identifying lapses in sensibility, as shown in this example from [35]. Advances in commonsense reasoning may lead to improvements in their ability to notice when the interlocutor says something abnormal and respond accordingly.

One note to make is that, although artificial consciousness, artificial general intelligence, and artificial general superintelligence are not necessarily subfields of natural language processing, the research which is most promising to get us to them is currently in natural language processing, particularly the spectacular work being done with large language models. As such, most of the potential advisers I have selected are in natural language processing.

Before turning to the Background section where we dive more deeply into what artificial consciousness and commonsense reasoning are, I would like to examine this work within the recommendation made by Prof. Ali Rahimi[5] during his influential 2017 NeurIPS talk [33]. In the talk, he decried the then-recent shift in the field away from "science" towards "alchemy", to use his phrasing. As he argued, the field had moved away from asking and answering scientific questions to arrive at deep, fundamental truths. Instead, researchers spent their energy trying to create ever-more powerful and impressive models. He argued that we as a field should focus on understanding why some techniques work and why others do not, which would give a stronger scientific foundation to the models we produce.

As someone in the early stages of his own research career, it is important that I examine my research within this framework, asking whether I am trying to find fundamental scientific truths or merely find more impressive tricks. The ultimate goal of my research is to discover a way to create an artificial consciousness, ideally one which can be constructed in the real world. I am interested in building a deeper understanding of how and why certain techniques work, but that goal is subservient to my main goal. This is not necessarily the best goal, and other computer scientists might differ from me on this point, but it forms the structure of my own research.

5 Thesis Statement

My thesis will advance our understanding of commonsense reasoning and its application to downstream tasks. The first half will focus on improving our ability to implement com-

[5] I had difficulty with rendering the Arabic script in LaTeX. In the future, I expect to have a setup which can render Arabic-script text more readily to better support names written in the Perso-Arabic script.

monsense reasoning itself, focusing on using natural language in place of logical formalisms, a promising new direction [10]. The second half will focus on finding better ways to use commonsense reasoning to improve performance on downstream tasks, primarily those in natural language processing such as dialogue summarization and question answering.

6 Timeline

I expect to conduct research for the various parts of my thesis concurrently as most of my thesis is, as the parallel processing researchers say, "embarrassingly parallelizable" [41]. Assuming the standard graduate school decision deadline, I can expect to have access to a university compute cluster by April 15 at the latest [12], at which point I can begin running experiments.

Chapter II

Background

1 What Are Artificial Consciousness, Artificial General Intelligence, and Artificial General Superintelligence?

Consciousness is hard to define [8], but we can intuitively understand that it is when a system "understands" that it exists and is separate from the rest of the world. An artificial consciousness is a conscious machine, a mechanical equivalent to human and animal consciousnesses [8].

The terms "sentience" and "consciousness" are often used synonymously, but there is a distinction to be made between them. Sentience has two main meanings aside from being a synonym of consciousness [8]. The first is that sentience is when an intelligent system is able to sense either the external world or its internal world [8]. This is different from consciousness as a system can sense the world and itself without being conscious [8], such as when a self-driving car is able to see the world around it. The second definition is that sentience is when a system is able to have emotional states or sensations such as pain and pleasure [8]; however, a system can be conscious while only having "neutral" experiences [8]. We see that the first definition is weaker than the definition of consciousness and that the second is stronger.

An artificial general intelligence is an artificial intelligence which is broadly intelligent, capable of completing many or most tasks which a human intelligence can solve, and which is able to reason, plan, and learn from experience at a level comparable to humans [7]. Although some researchers use the term artificial general intelligence to refer to an artificial intelligence which can operate at a level beyond that reachable by a human, I reserve that distinction for an artificial general *super*intelligence.

Figure II.1: An example of the structure of ATOMIC [36], a commonsense knowledge graph. As you can see, the logic is captured as nodes and edges.

2 What Is Commonsense Reasoning and Why Is It Important?

Commonsense knowledge is the knowledge about the world which we expect all humans to possess [17]. For example, we expect all humans to understand that an apple can fall from a tree to the ground but that it cannot rise from the ground to the tree. Finding better ways to represent commonsense knowledge is a major area of research [17], and major recent works include ATOMIC [36] and ConceptNet [39]. Commonsense reasoning is the ability to contextualize and draw upon commonsense knowledge and generalize to novel situations [38].

There are many ways to implement commonsense reasoning, and there is disagreement on which one will bear the most fruit in the long term. As an example, some researchers view commonsense knowledge graphs as the most promising route while others have more faith in commonsense knowledge models. Commonsense knowledge graphs capture knowledge about the world through graphs, capturing entities as nodes and describing relationships between the entities as edges between the nodes [21]. An example is ATOMIC [36], which captures *if-then* reasoning. As an example of the logic contained in ATOMIC, consider the reasoning *if* X *pays* Y *a compliment, then* Y *will likely return the compliment.* A more complete example of the structure of ATOMIC is given in fig. II.1. While commonsense knowledge graphs are static, capturing only the information contained in them and nothing more, commonsese knowledge models are able to generate new knowledge about the world [21] by extending a given commonsense knowledge graph, adding novel nodes and edges [5, 21]. An example is COMET [5], whose structure is shown in fig. II.2.

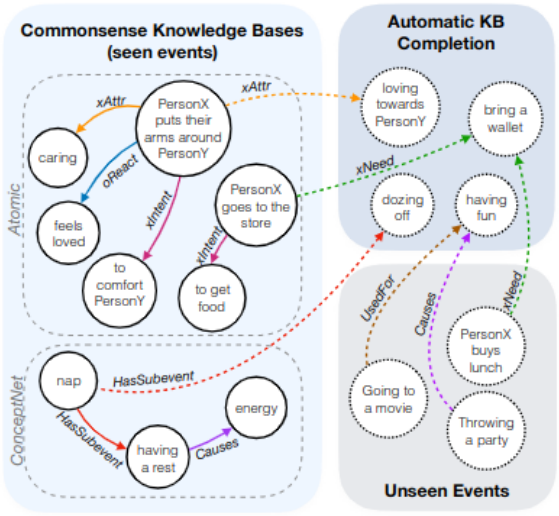

Figure II.2: An example of the structure of COMET [5], a commonsense knowledge model. Given an existing commonsense knowledge graph, shown in solid nodes and edges, it infers novel entities and relationships, shown in dashed nodes and edges. With this, it can better respond to unseen events, helping it generalize more well.

My thesis will focus on a major direction of research which has garnered attention recently [10]: capturing and manipulating commonsense knowledge as natural language instead of as logical formalisms, an idea detailed in chapter III. The simple intuition is that, rather than trying to capture how the world works in a formal system, such as in a graph with nodes and edges, it might be more fruitful to capture the world in natural language, the reasoning being that no other medium can contain the complexities of the world [10].

One important development is that large language models have been shown to exhibit a rudimentary form of commonsense reasoning [34,40,49]. However, they have their limitations. They struggle to solve tasks which require several chained inferences steps [49], though there is work being done to try to improve performance in this [42]. They also are often unable to solve dual tasks which are correlated such that the correct prediction of one sample should lead to correct prediction of the other [49]. This is understood to be a sign that they only possess a shallow understanding of commonsense, not the deep one we want them to have [49].

Improving commonsense reasoning might be a step towards artificial general intelligence and artificial general superintelligence. Large language models are currently the most promising avenue towards artificial general intelligence [7]. Considering that they struggle with commonsense reasoning [49], improving commonsense reasoning might take us one step closer to true artificial general intelligence. Although it is difficult to know precisely what needs to be done to arrive at artificial general superintelligence, the most reasonable direction is to first achieve artificial general intelligence, so commonsense reasoning would likely be a step in that direction as well. Similarly, commonsense reasoning is reasonably likely to be a step towards artificial consciousness.

3 What Are the Downstream Tasks Which Use Commonsense Reasoning?

My thesis will focus on applying commonsense reasoning to natural language processing tasks. Although commonsense reasoning can be applied to improve performance in downstream tasks in other fields, such as computer vision [45] and human activity recognition [27], most work in applying it to downstream tasks has focused on those in natural language processing. A simple overview of the downstream tasks which I plan to consider is given below.

- **Dialogue generation** is the task of, given a two-person dialogue history and a user utterance, generating an appropriate response [24].
- **Dialogue summarization** is the task of generating a summary of a conversation while preserving its context [21] and retaining factual consistency [35].
- **Sequence classification** is the task of identifying some attribute of the dialogue, e.g.
  - what is the intent of the conversation?
  - what is the emotional state of the other speaker? and
  - what is the topic of the conversation?

  The two major subtasks on which I plan to focus are
  - *emotion detection*, the detection of the emotion for each utterance in a given conversation [44] and
  - *causal emotion entailment*, the detection of causal utterances for a non-neutral targeted utterance from a conversation [23, 35].
- **Presumption detection**, as defined in [2], is the task of, given a command of the form **if-then-because**, inferring unstated additional conditions on the **if** or the **then** portion [2].
- **Dialogue modelling** is to dialogues what language modelling is to unstructured text [35]. It is used for both open dialogue, the setting where the conversation does not have any particular goal, as typified by the chatbot setting, and task-oriented dialogue [35].
- **Dialogue systems** are artificial intelligences which can have a conversation with a human interlocutor [29]. Two major subfields on which I intend to focus are:
  - avoiding harmful dialogues, reacting appropriately to troubling user statements and
  - target-guided dialogue, when a dialogue system tries to guide a conversation towards a particular sentence.
- **Recognizing Emotion Cause in Conversations (RECCON)** is the task of finding the emotion associated with an utterance in a conversation along with its cause [31]. Researchers sometimes view causal emotion entailment as a subtask of RECCON, dubbing it *Conversational Causal Emotion Entailment* (C_2E_2), the detection of causal utterances for a targeted utterance from conversations where the emotion is given [23].

Chapter III

Finding Better Ways to Solve Commonsense Reasoning

This is an example of trunk work in that advancements made to commonsense reasoning advance our ability to create an artificial intelligence which is able to understand concepts. We can use these trunk advancements in branch research by applying commonsense reasoning to downstream tasks, the focus of chapter IV.

Large language models have been shown to be viable sources of information, internally storing knowledge contained in their training data [6, 14, 30]. A large language model used in this way is called a *neural knowledge bases* [25]. Recently, one promising direction of research in commonsense reasoning has emerged, using the implicit commonsense reasoning found in pretrained language models [34, 40, 49] such as BERT [13, 25]. State-of-the-art techniques have weaknesses, however [3, 25].

The avenues through which I will attempt to advance the research in commonsense reasoning loosely fall under the umbrella of natural-language-based commonsense reasoning. Because this is where the most promising research is currently being done [10], I intend to focus on this.

1 Methods

For this section of my thesis, I will use the following datasets.

- NumerSense [25] is a dataset of 13.6k masked-word-prediction probes each masked word is a number in a sentence, e.g. "a bird usually has <blank> legs", along with the appropriate number, "two" in the example given.
- TimeDial [32] is a multiple choice cloze task of 1.1k dialogues. Each element of the dataset is a two-person dialogue, a final utterance containing a blank, four possible fill-ins for the blank (three incorrect, one correct), and a note of which fill-in is correct.

- COM2SENSE [38] is a collection of natural language true/false statements expressed in complementary sentence pairs. There are 4k samples in the dataset.
- SocialIQA [37] is a dataset of 38k multiple choice questions, each on a social scenario. For example, the social scenario might be "Alice asks Bob not to speak loudly in the library," while the question and correct answer are "why did Alice do this?" and "because people are reading," respectively. Two incorrect answers are also given for each sample.
- Physical Interaction: Question Answering (PIQA) [4] is a dataset which considers physical concepts, such as the correct ways to apply makeup. Each of the 21k samples consists of a question about the physical world and two possible answers to the question, only one of which is correct, along with a label of which of the two possible answers is correct.
- The AI2 Reasoning Challenge (ARC) [11] dataset is a corpus of 8k grade-school science questions, each question paired with three incorrect answers and one correct answer. There is an adjoining ARC Corpus, which consists of 14M science sentences.
- Question Answering via Sentence Composition (QASC) [19] is a multi-hop reasoning dataset. Like other question answering datasets, it consists of questions paired with possible answers, only one of which is correct. What is distinctive of QASC is that it also contains a corpus of facts and, to answer a given question, multiple facts must be composed, creating a chain of reasoning.
- HellaSWAG [46] is a dataset consisting of scenarios paired with possible next steps, only one of which is correct. For example, the scenario might be, "a boy opens a water bottle", and the correct next step might be, "he drinks from the water bottle". The samples contained in the dataset were specifically chosen for their difficulty to contemporary commonsense reasoning techniques.

2 Numerical Commonsense Reasoning as a Step Towards General Commonsense Reasoning

Numerical commonsense reasoning is the task of reasoning about numerical quantities [25]. It is often presented as a masking task, i.e. given a sentence such as "a dog has <blank> legs", the model is used to fill in the blank when restricted to only numeric words, in this case ideally with "four" [3,25]. My research in this task will use the NumerSense dataset [25].

Improvements in this direction have downstream effects, most notably that

- they would make it easier to populate numerical facts in current commonsense knowledge bases [25] and
- they would improve performance in open-domain question answering, e.g. given the question "how many legs do ants have?", a numerical-commonsense-reasoning-powered question answering model would be more likely to correct answer with "six" [25].

One recent work which has improved performance in numerical commonsense reasoning is *generated knowledge prompting*, the generation of knowledge from a language model followed by the provision of the knowledge as additional input when answering a question [26]. Formally, they represent the task as the prediction of an answer $a \in A_q$, represented as natural language, given a question $q \in Q$, also represented as natural language, where the set of choices A_q is finite, though in my research on numeric commonsense reasoning A_q would ideally be the infinite set of all numbers. They then use a language model $p_G(k|q)$ to generate knowledge statements K_q conditioned on the question using

$$K_q = \{k_m : k_m \sim p_G(k|q), m = 1, \cdots, M\}$$

where M is the quantity of knowledge statements to generate. The second step is to integrate this into the commonsense reasoning step, where a language model p_I is used to predict the answer using

$$\hat{a} = \arg\max_{a \in A_q} p_I(a|q, K_q).$$

This contrasts with the vanilla technique, which omits the K_q, i.e.

$$\hat{a} = \arg\max_{a \in A_q} p_I(a|q).$$

As you can see, it is similar in spirit to chaining logical inferences in natural language [42].

I intend to extend this direction, finding better ways to use it to solve numerical commonsense reasoning. Although generated knowledge prompting can be used for non-numerical commonsense reasoning, such as scientific commonsense [26], I will focus on numerical commonsense reasoning.

It is possible, however, that improvements to numerical commonsense reasoning based on generated knowledge prompting can be transferred to other tasks, such as scientific, social, event, physical, prototypical, and temporal commonsense [3]. The datasets which I would use for these tasks include Social IQA [37] for social commonsense reasoning, ARC [11] and QASC [19] for scientific commonsense reasoning, etc.

After finding new techniques to implement *numerical* commonsense reasoning, I intend to conduct research in extending these techniques to other forms of commonsense reasoning. In effect, I intend to use numerical commonsense reasoning as a step towards general commonsense reasoning, a simpler task whose solutions may solve a more complex task.

3 Using Large Language Models to Generate and Extend Knowledge Graphs

As discussed in section 2, we can generate knowledge graphs automatically rather than manually. While some researchers argue against the utility of knowledge graphs, they remain worth considering. In this section of my thesis, I will investigate ways to use large language models to generate or extend knowledge graphs. A major work in this direction which will form the foundation for my own research is COMET [5].

COMET generates novel commonsense knowledge. Its training knowledge base consists of tuples in $\{s, r, o\}$ format where s is the natural language phrase subject of the tuple, r is the relation of the tuple, and o is the natural language phrase object of the tuple An example tuple is (s = "drop a ball", r = causes, o = "ball falls on the ground"). It interprets novel commonsense knowledge generation as the task of generating o given s and r as inputs. To achieve this, it is trained to maximize the conditional loglikelihood of predicting the o tokens. We can represent these tuples as a knowledge graph by interpreting s and o as nodes and r as an edge.

I intend to continue research in this direction, finding new techniques to extend knowledge graphs.

Chapter IV

Finding Better Ways to Apply Commonsense Reasoning to Downstream Tasks

This is an example of branch work in that it applies an artificial intelligence's ability to understand concepts to specific tasks. We consider here how applying commonsense reasoning can improve performance on several downstream tasks, examining prior work on which I intend to build in my thesis. Notably, many seemingly different tasks can involve commonsense reasoning in similar ways, such as how SICK [21] and CISPER [44] both intermix commonsense reasoning inferences with the original data to improve performance on their respective tasks.

1 Dialogue Generation

To improve performance on dialogue generation, *Dynamic Multi-form Knowledge Fusion based Open-domain Chatting Machine* (*DMKCM*) [43] uses both a commonsense knowledge graph and unstructured text from natural language documents. They use a virtual knowledge base, an indexed corpus of documents linked to related documents via keywords. The key insight was using a virtual knowledge base to locate relevant documents and then expanding the content of the dialogue and the relevant documents using a commonsense knowledge graph to get apposite triples.

2 Dialogue Summarization

Two novel techniques, *Summarizing with Injected Commonsense Knowledge* (*SICK*) and *SICK++* [21], improve performance on document summarization. To do so, SICK uses commonsense inferences as additional context, interlacing utterances with commonsense inferences based on them, before feeding them together to a standard document summarizer.

They also extend on this with SICK++, which uses commonsense as supervision by adding commonsense inference generation as an additional task in a multi-task learning setting.

3 Sequence Classification

Emotion detection. Similar to SICK's [21] approach to dialogue summarization, CISPER [44] improves emotion detection by interweaving commonsense inferences with utterances.

Causal emotion entailment. To integrate commonsense reasoning in causal emotion entailment, *Knowledge Enhanced Conversation graph* (*KEC*) [23] propagates social commonsense knowledge between the given utterances. They also propose a novel filtering strategy for selecting commonsense knowledge based on sentiment. Their final contribution is a novel method for processing KEC, the construction of Knowledge Enhanced Directed Acyclic Graph networks.

4 Presumption Detection

A recently proposed technique for presumption detection [1] improves performance on the task through two innovations. The first is an iterative knowledge query mechanism which uses a multi-hop reasoning chain to significantly reduce the search space. They use a neural knowledge base, which necessarily has gaps in knowledge. They include a second innovation, a mechanism for querying a human user to fill gaps in the neural knowledge base, but I do not intend to extend work in this direction. I would ideally focus on what we may term *human-NOT-in-the-loop artificial intelligence*, i.e. artificial intelligence techniques which do not need a human to fill in gaps in knowledge or reasoning. As such, although this second innovation certainly has utility in certain real-world use cases, I will not continue work in this direction.

5 Dialogue Modelling

Think-Before-Speaking [48] incorporates commonsense reasoning into dialogue modelling by first externalizing commonsense knowledge, which they term the *think* step, and then using this knowledge to generate responses, which they term the *speak* step.

6 Dialogue Systems

Avoiding harmful dialogues. A recent work [20] has two contributions. The first is a dialogue safety detection module, Canary, which generates prosocial commonsense reasoning rules given a dialogue; Canary can even be plugged into off-the-shelf language models to increase their prosocial tendencies. The second is a socially-informed dialogue system, Prost,

which generates dialogues which are more prosocial than previous state-of-the-art dialogue systems. I intend to expand on this, finding better ways to reason about prosocial behaviour, in some sense creating a moral compass based on commonsense reasoning.

Target-guided dialogue. In a spirit similar to chaining commonsense inferences, a recent work [16] guides a dialogue towards a target through two steps. The first step is finding a chaining path of commonsense inferences from the current sentence to the target sentence. The second step is using the commonsense inferences to generate transition responses. Like other recent developments in commonsense reasoning, they use reasoning chains, so finding better ways to chain together commonsense inferences would be a promising direction, and I intend to conduct research to this effect.

7 Recognizing Emotion Cause in Conversations

Causal Aware Interaction Network (*CauAIN*) [47] retrieves commonsense causal clues to guide the process of causal utterance, conducting both the retrieval and traceback steps from both the perspective of the human user and the dialogue system. This is arguably the most promising work in this task, and I intend to investigate potential extensions of it.

Chapter V

Conclusion

I propose a thesis which advances (1) the research on commonsense reasoning and (2) the research on applying commonsense reasoning to downstream tasks, chiefly those in natural language processing. This is done with the hope that it would prove a step towards artificial consciousness, artificial general intelligence, and artificial general superintelligence. My hope is that you enjoyed reading it and that, if you have an open chair in your lab, that you accept my application to graduate school.

Thank you for reading.

Bibliography

[1] Forough Arabshahi, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom Mitchell. Conversational Multi-Hop reasoning with neural commonsense knowledge and symbolic logic rules. September 2021.

[2] Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. Conversational Neuro-Symbolic commonsense reasoning. *AAAI*, 35(6):4902–4911, May 2021.

[3] Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. March 2023.

[4] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *AAAI*, 34(05):7432–7439, April 2020.

[5] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. June 2019.

[6] Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. Inducing relational knowledge from BERT. *AAAI*, 34(05):7456–7463, April 2020.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. March 2023.

[8] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A K Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness. August 2023.

[9] Jorge Cham. Your thesis title, 2006.

[10] Yejin Choi. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155, May 2022.

[11] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. March 2018.

[12] Council of Graduate Schools. April 15 resolution. `https://cgsnet.org/resources/for-current-prospective-graduate-students/april-15-resolution`, December 2021. Accessed: 2023-12-13.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. October 2018.

[14] Joshua Feldman, Joe Davison, and Alexander M Rush. Commonsense knowledge mining from pretrained models. September 2019.

[15] Jolien C Francken, Lola Beerendonk, Dylan Molenaar, Johannes J Fahrenfort, Julian D Kiverstein, Anil K Seth, and Simon van Gaal. An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neurosci Conscious*, 2022(1):niac011, August 2022.

[16] Prakhar Gupta, Harsh Jhamtani, and Jeffrey P Bigham. Target-Guided dialogue response generation using commonsense and data augmentation. May 2022.

[17] Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L McGuinness, and Pedro Szekely. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347, October 2021.

[18] Spike Jonze. Her, October 2013.

[19] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. *AAAI*, 34(05):8082–8090, April 2020.

[20] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. ProsocialDialog: A prosocial backbone for conversational agents. May 2022.

[21] Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-Won Hwang, and Jinyoung Yeo. Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. September 2022.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, January 2012.

[23] Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. May 2022.

[24] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. June 2016.

[25] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! NumerSense: Probing numerical commonsense knowledge of pre-trained language models. May 2020.

[26] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. October 2021.

[27] Jesús Martínez del Rincón, Maria J Santofimia, and Jean-Christophe Nebel. Common-sense reasoning for human action recognition. *Pattern Recognit. Lett.*, 34(15):1849–1860, November 2013.

[28] Thomas Nagel. What is it like to be a bat? *Philos. Rev.*, 83(4):435, October 1974.

[29] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: a systematic survey. *Artif. Intell. Rev.*, 56(4):3055–3155, April 2023.

[30] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? September 2019.

[31] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations. December 2020.

[32] Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. TIMEDIAL: Temporal commonsense reasoning in dialog. June 2021.

[33] Ali Rahimi. Back when we were kids. NeurIPS, December 2017.

[34] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. June 2019.

[35] Christopher Richardson and Larry Heck. Commonsense reasoning for conversational AI: A survey of the state of the art. February 2023.

[36] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. ATOMIC: An atlas of machine commonsense for if-then reasoning. *Proc. Conf. AAAI Artif. Intell.*, 33(01):3027–3035, July 2019.

[37] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. April 2019.

[38] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. June 2021.

[39] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. (1), February 2017.

[40] Trieu H Trinh and Quoc V Le. Do language models have common sense? September 2018.

[41] Dana Vrajitoru. Embarrassingly parallel programs. `https://www.cs.iusb.edu/~danav/teach/b424/b424_23_embpar.html`. Accessed: 2023-12-13.

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. January 2022.

[43] Feifei Xu, Shanlin Zhou, Xinpeng Wang, Yunpu Ma, Wenkai Zhang, and Zhisong Li. Open-domain dialogue generation grounded with dynamic multi-form knowledge fusion. April 2022.

[44] Jingjie Yi, Deqing Yang, Siyu Yuan, Caiyan Cao, Zhiyao Zhang, and Yanghua Xiao. Contextual information and commonsense based prompt for emotion recognition in conversation. July 2022.

[45] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019.

[46] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? May 2019.

[47] Weixiang Zhao, Yanyan Zhao, and Xin Lu. CauAIN: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, California, July 2022. International Joint Conferences on Artificial Intelligence Organization.

[48] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. October 2021.

[49] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. Evaluating commonsense in Pre-Trained language models. *AAAI*, 34(05):9733–9740, April 2020.

Appendix A

Relevant Piled Higher and Deeper (PhD) Comic

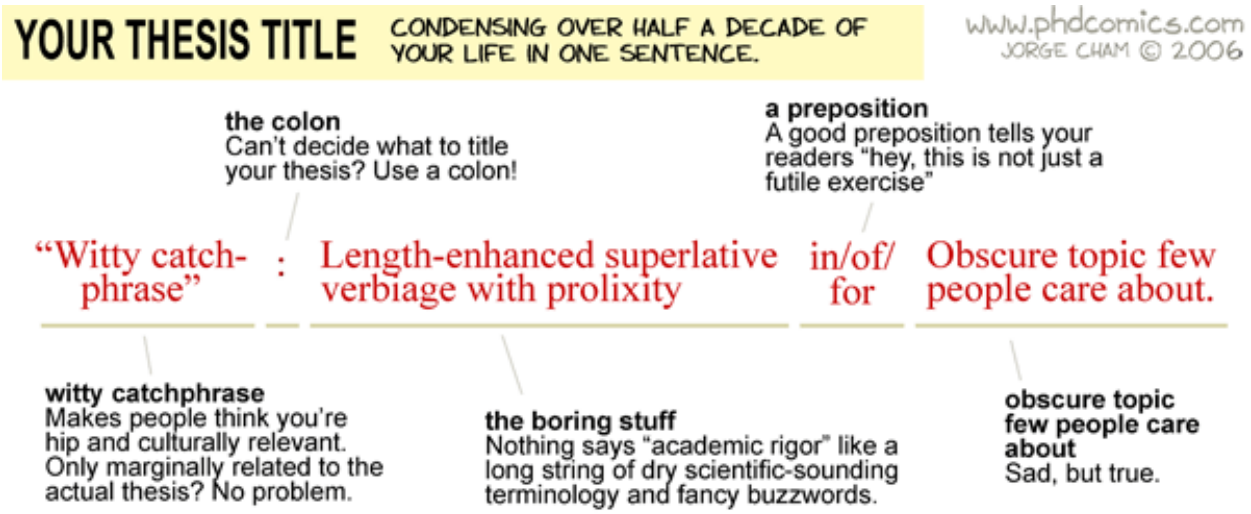

Figure A.1: After writing my thesis title, I realized that it exactly fit this template, as the prophecy foretold. All credits to Jorge Cham [9]. Also, I had to cite this as an artwork in Paperpile, as all Piled Higher and Deeper comics should be.