Finding Adam

Advancing Commonsense Reasoning as a Potential Step towards Artificial Consciousness and Artificial General (Super)intelligence

A Doctoral Thesis Proposal

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

December 12, 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 and artificial general (super)intelligence 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 and artificial general (super)intelligence.

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 [10], 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 [10] 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 and artificial general (super)intelligence

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 [3]. As a simple intuition, a system is conscious is when there is something that it is like to be that system [14], when a system "knows" that it exists. An example from literature is the artificial intelligence named "Samantha" from the film *Her* [8]. 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.

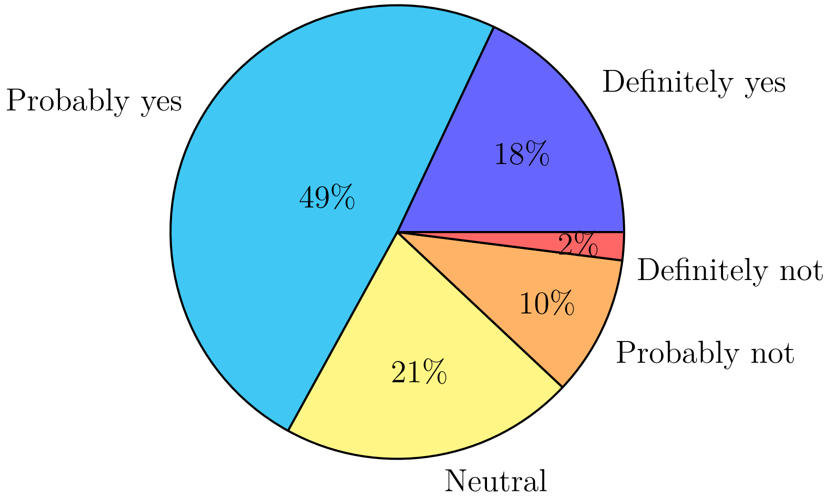


Figure I.1: The views of different consciousness scientists on whether an artificial intelligence system could be conscious (figures rounded to the nearest integral percentage) [6].

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. The ethical considerations section of this thesis proposal examines possible concerns, including the question of whether artificial general intelligence or artificial general superintelligence would pose an existential threat to humanity, in addition to the more immediate and less flashy concerns, such as misalignment.

[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 [5].

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

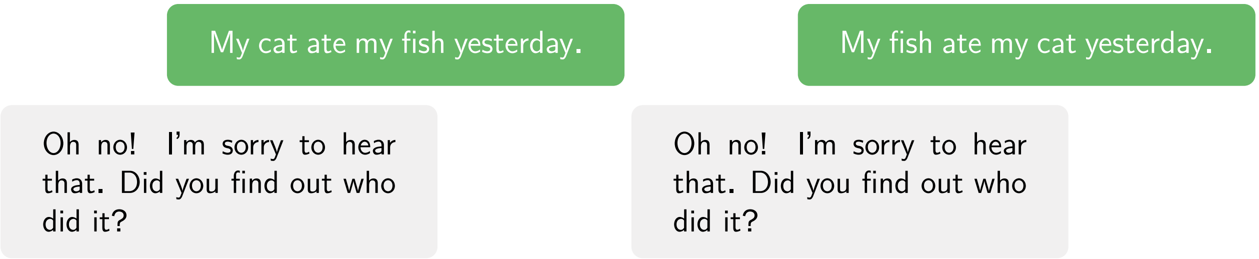
1. people use flights to get from one place to another,
2. if someone misses a flight, they usually experience negative emotions,
3. someone who misses a flight would usually like a ticket to a future flight which goes to the same place, though the airports might be slightly different, e.g. LGA instead 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 [5]. 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 [17].

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.



My cat ate my fish yesterday.

Oh no! I'm sorry to hear that. Did you find out who did it?

My fish ate my cat yesterday.

Oh no! I'm sorry to hear that. Did you find out who did it?

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 [17]. 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 [15]. 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] 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.

Chapter II

Background

1 What is Consciousness?

Consciousness is hard to define [3], but we can intuitively understand that it is when a system "understands" that it exists and is separate from the rest of the world.

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 [3]. The first is that sentience is when an intelligent system is able to sense either the external world or its internal world [3]. This is different from consciousness as a system can sense the world and itself without being conscious [3], 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 [3]; however, a system can be conscious while only having "neutral" experiences [3]. We see that the first definition is weaker than the definition of consciousness and that the second is stronger.

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 [7]. 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 [7], and major recent works include ATOMIC [18] and ConceptNet [20]. Commonsense reasoning is the ability to contextualize and draw upon commonsense knowledge and generalize to novel situations [19].

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

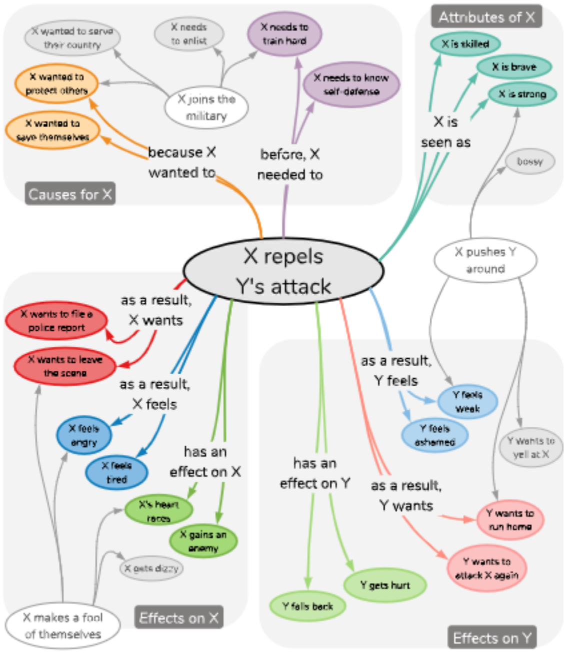


Figure II.1: An example of the structure of ATOMIC [18], a commonsense knowledge graph. As you can see, the logic is captured as nodes and edges.

the world through graphs, capturing entities as nodes and describing relationships between the entities as edges between the nodes [9]. An example is ATOMIC [18], 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 [9] by extending a given commonsense knowledge graph, adding novel nodes and edges [1, 9]. An example is COMET, whose structure is shown in fig. II.2.

My thesis will focus on a major direction of research which has garnered attention recently: 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 [5].

One important development which bears noting is that large language models have been shown to exhibit a rudimentary form of commonsense reasoning [16, 21, 25]. However, they have their limitations. They struggle to solve tasks which require several chained inferences steps [25], though there is work being done to try to improve performance in this [22]. 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 [25]. This is understood to be a sign that they only possess a shallow understanding of commonsense, not the deep one we

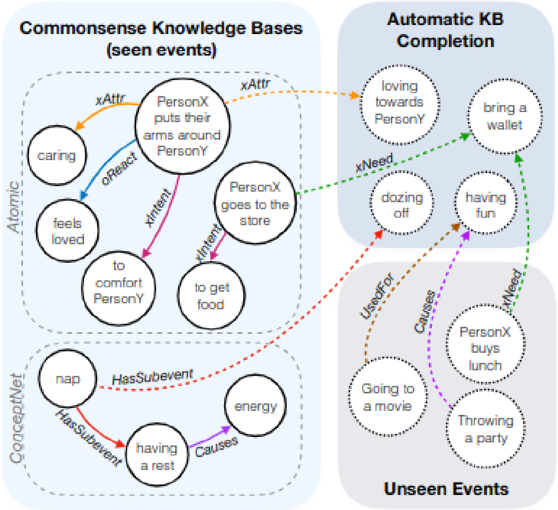


Figure II.2: An example of the structure of COMET [1], a commonsense knowledge model. Given an existing commonsense knowledge graph, shown in solid nodes and edges, it infers novel entities and relationships, shown in dashed nodes and edges. With this, it can better respond to unseen events, helping it generalize more well.

want them to have [25].

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 [2]. Considering that they struggle with commonsense reasoning [25], 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.

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 [24] and human activity recognition [13], 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 [12].
- **Dialogue summarization** is the task of generating a summary of a conversation while preserving its context [9] and retaining factual consistency [17].

- **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?
    - what is the topic of the conversation?

  Two major subtasks on which I plan to focus, among others, are
    - *emotion detection*, the detection of the emotion for each utterance in a given conversation [23], and
    - *causal emotion entailment*, the detection of causal utterances for a non-neutral targeted utterance from a conversation [11].

Chapter III

Finding Better Ways to Solve Commonsense Reasoning

Chapter IV

Finding Better Ways to Apply Commonsense Reasoning to Downstream Tasks

Chapter V

Ethical Concerns

Chapter VI

Timeline

Chapter VII

Conclusion

Bibliography

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

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] Spike Jonze. Her, October 2013.

[9] 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.

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

[11] 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.

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

[13] 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.

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

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

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

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

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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

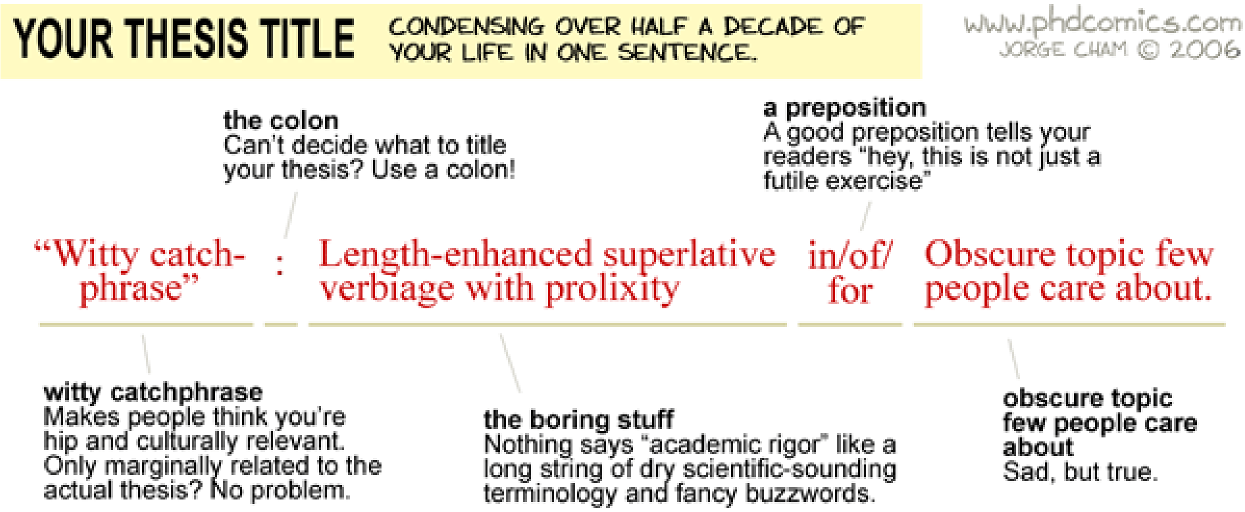

YOUR THESIS TITLE
CONDENSING OVER HALF A DECADE OF YOUR LIFE IN ONE SENTENCE.
www.phdcomics.com
JORGE CHAM © 2006
the colon
Can't decide what to title your thesis? Use a colon!
a preposition
A good preposition tells your readers "hey, this is not just a futile exercise"
"Witty catch-phrase"
:
Length-enhanced superlative verbiage with prolixity
in/of/for
Obscure topic few people care about.
witty catchphrase
Makes people think you're hip and culturally relevant. Only marginally related to the actual thesis? No problem.
the boring stuff
Nothing says "academic rigor" like a long string of dry scientific-sounding terminology and fancy buzzwords.
obscure topic few people care about
Sad, but true.

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 [4]. Also, I had to cite this as an artwork in Paperpile, as all Piled Higher and Deeper comics should be.