

Entailment in Large Language Models:

A Case for Novelty

Senior Honors Capstone Project Proposal  
submitted in partial fulfillment of requirements  
for the Norbert O. Schedler Honors College.

by

Kyle Urban

University of Central Arkansas

Conway, Arkansas

Fall 2024

Proposal Committee:

Mentor: Zachary Stine, Ph.D.

Assistant Professor

Instructor: Doug Corbitt, M.A.

Lecturer

### **Abstract:**

The purpose of this thesis capstone is to explore the Natural Language Inference abilities of various Large Language Models in the realm of linguistic entailment. This will be accomplished by highlighting deficiencies in current benchmark tests, creating data sets for a benchmark that contains pragmatic concepts not previously touched on, and then having LLMs take these new benchmarks and compare their performance levels. This ultimately works to highlight ways in which there is room to improve LLMs language understanding abilities in the future.

### **Introduction:**

As a linguistics major and computer science minor, the machine implementation of human language has always fascinated me. However, it wasn't until going through the cognitive linguistics class here at the University of Central Arkansas when I started getting curious about how computers built for linguistic activities handle understanding implied context from sentences. For context, the field known as Natural Language Processing (NLP) resides at the intersection of linguistics and computer science. It is dedicated to creating and perfecting computational tools for working with human language. One niche tool in that tool belt are Large Language Models (LLMs). LLMs are a rather recent invention — coming about only in 2017 — some of which have become rising stars in the popular media, like ChatGPT, Gemini, Co-pilot, etc.

In turn, NLP contains many subdisciplines, one such being Natural Language Inference (NLI). A major focus of NLI are entailment tasks. Entailment is the inference of a hypothesis from a discourse. Thus, NLI entailment tasks are concerned with testing abilities of certain

language technologies at correctly drawing the connection between implied information and explicit context. LLMs are the newest tool being tested against NLI entailment tasks.

While researching through the current literature on tests to used to grade the entailment NLI abilities of LLMs, I wondered if perhaps there were gaps in what was covered by those benchmarks. Human language is inherently generative at all levels, and it is entirely possible for people to utter a sentence — and thus an implied context — that has never been heard before. While not for a lack of trying, current benchmarks like the RTE, SNLI, and the MultiNLI cannot cover everything that human language has to offer. In light of this, I propose three primary research questions that I wish to resolve in my Honors Capstone Thesis. What pragmatic concepts are currently not represented in these benchmarks? What novel benchmarks can be created in those lacking areas? How LLMs trained on previous benchmarks perform on those new benchmarks?

The methodology for approaching these questions is going to be a mixed quantitative and qualitative one, as there is going to be both a collection of numerical data and also an analysis of old benchmarks and the design of new ones. Possible approaches will consist of either getting access to more descriptions of benchmark data, modeling LLMs trained on the benchmark data in the latent space, and from pulling inspiration from the plethora of linguistic elicitation questions that focus on entailment in the English Language.

Furthermore, the foci of this project can be formally defined as the intersections of Machine Learning and Semantics, fields in computer science and linguistics respectively. Finally, this project shall be considered to be what is known as the “thesis” capstone project, since it will require an analytical collection of data.

## **Literature Review:**

### **Intro**

The modern prominence of very advanced large language models — LLMs — such as ChatGPT and BERT, has brought about the question of how well these – and other – LLMs are at understanding all the novelties of natural human language. One such linguistic phenomenon, entailment, has been the crosshairs of NLP — natural language processing — researchers for a good part of the late 2000s and 2010s. Various early efforts, such as the PASCAL RTE — Recognizing Textual Entailment (Dagan, et al. pp 177-190, 2005) — and the Stanford RTE (MacCartney, 2009) set about trying to devise standardized data sets at which LLMs could be tested against in order to ascertain their NLI — natural language inference — abilities. Such tests have been incorporated as components into newer data sets like the MultiNLI (Williams, et al 2017) or as subsets of entirely new tests such as the General Language Understanding Evaluation — GLUE (Wang, et al. 2018). Yet however expansive these data sets can be, they still don't entirely cover all the natural weirdness that human language — especially the realm of semantics and pragmatics — has to offer. Plus, the utilization of these standards as the bench mark for various LLMs to be trained towards has created a sort of “sameness” of current NLI abilities. Compounded with the fact that LLMs rarely score perfectly on these NLI test, there is a serious need for zanier data sets to truly push current LLMS to their limits.

### **Terms Defined**

Before we can decide where the future of NLI can go, we must first see where we have already been. In order to do so, it will first help to define our broad concepts, narrow down, and then trace the history from the mid 2000's to now. To begin, we must first provide an overview of what exactly an LLM is. In essence, it is an autoregressive sequence that statistically predicts what word in a sentence might come next based upon what tokens — words — have previously appeared and/or been generated (Seff, 2:18-3:00, 2023). These probabilities for these predictions are taken from an analysis of collocations; ie: what is the likelihood of certain words appearing alongside other words in the same context, from a given corpus of language data (Seff, 2:18-3:00, 2023). It should also be noted that LLMs can be fine-tuned to do many different language tasks, such as write long texts, summarize texts, translation, etc. However, what this project focuses on is Q&A (Seff, [5:30-7:00], 2023) as that task involves having not only frequent back and forth dialogue with human users, but also having to work to understand those users.

LLMs don't entirely exist within a vacuum, but rather a tool in the toolbelt of computer scientists, linguists, data scientists, and mathematicians working in the field of Natural Language Processing. Of course, Natural Language Inference is the subfield of NLP attempting to upgrade the abilities of those languages tools, such as LLMs, in detecting logically implied from a context.

Now, within the realm of NLI, there are what is know as inference tasks (Wang, et al, 2018). Tasks are essentially different features of language that a LLM has to figure out how to understand. In the case of NLI, these features would be how implicit information is expressed in human language, and the task for the LLM must be to comprehend these implicatory language features.

## **What is entailment?**

There are many different types of implicatory language features that we can train an LLM to understand, such as Scalar Inferences – SIs, Sarcasm, Rhetorical Questions, etc. However, the bulwark of my interest comes from entailment tasks. In pragmatics, which is a subfield of linguistics, entailment is defined as what is the implied information given by an utterance. In formal logic, this is sometimes known as a syllogism (Saeed, pg 90, 2016) and is often expressed as  $p \rightarrow q$  in many discrete mathematics and formal logic courses. Training an LLM to understand entailment is a very necessary undertaking. Humans use entailment in everyday conversation, and an inability for an LLM to ascertain it once in all its near infinite occurrences in natural language could lead to a disruption of Gricean Principles and thus conversation between humans and Q&A fine-tuned LLMs (Miehling, 2024).

## **NLI & its history: PASCAL RTE**

The vitality of an LLM's NLI abilities in understanding entailment being so integral to maintaining conversation with humans, it is no surprise that work in this niche has been detailed and extensive. The history really begins with the PASCAL RTE (Dagan, et al. pp 177-190, 2005), an online corpus of data containing two strings of text. An overt statement, labeled T, and a hypothesis, labeled H. The PASCAL RTE test then asked models to analyze sets of 1000s of T and H pairs, and provide either a True or False value for each pair on whether or not the information in H was implied by T (Neeraj, 2017). The RTE went through 8 versions in total, RTE 1 to 8. However, the RTE 1, 3, and 5 have been the most foundational, with later versions often being noted as straying too far from the original criteria of testing entailment (Wang, et al. pg 4, 2018).

## **NLI & its history: Stanford RTE/SNLI**

Around the time of the RTE4's release, there was some disagreement within the NLP community of how effective the test was at ascertaining a model's ability to understand entailment. Since while the test could be used to determine how well a model was at pointing out whether or not entailment was present or not between T and H, it did not adequately cover all the possible different types of the entailment and non-entailment between T and H, such as contradiction, negation, quantifiers, etc. (MacCartney, pg 11-13. 2009). Bill MacCartney felt that perhaps he could make a more rigid test, one that more accurately captured the realities of natural language (MacCartney, pg 11. 2009). He devised what is known as the Stanford RTE to solve two main issues.

The first issue was the disparity between deep-but-brittle vs the robust-but-shallow approaches. A benchmark that trained models to catch deeply embedded information from a highly specific context only found success when those models remained confined to working with language data from that discourse. Any attempt to expand that model to new subject matters experienced a total breakdown in its entailment inference abilities. (MacCartney, pg 12, 2009). On the other hand, a benchmark that trained models to catch entailment in a wide variety of contexts, while good for over simplistic, surface level Q&A, broke down if the relation between T and H got more complex (MacCartney, pp 12-13. 2009). MacCartney found the solution to this problem by proposing what he termed as an "approach to NLI based on natural logic" (MacCartney, pp iv-v, 2009). Obtaining this approach required changing a fundamental component of a more embedded, second problem with the RTE.

Said second issue was that of the grading system of the RTE. MacCartney found that he wanted a data set that had a 3 way label distinction, of entailment, contradiction, and non-entailment for his Stanford RTE that would be decided by weighing various factors in either

the positive, negative, or neutral directions. (MacCartney, pg 55-58, 2009) rather than the mere two-way distinction of entailment vs non-entailment of the original RTE.

Overall, Bill MacCartney's work has not halted. His original benchmark as since evolved into the SNLI, created in 2015. However, he hasn't been the only researcher working to build a better benchmark for testing the entailment abilities.

### **NLI & its history: MultiNLI**

In 2017, Samuel Bowman and Adina Williams, building on the work of the SNLI, released their own version of an entailment benchmark known as the MultiNLI (Williams, et al, pg 1, 2017). This new test has over 433k examples of entailment, the largest data set at the time (Williams, et al. pg 1. 2017). It pulled examples of entailment from ten publicly available sources over a wide range of topics; such as fiction, government documents, slate poetry, written logs of telephone and face to face conversation, etc. (Williams, pg 5, 2017).

The authors of the MultiNLI took issue with the lack of sentence complexity in Bill MacCartney's SNLI. They noted that they found the SNLI's entailment examples having the overt premises and hypothesis — which the T and H now known as premise and hypothesis, another minor change between the Stanford RTE and the SNLI — consist of only a single sentence was lacking (Williams, et al, pg 2, 2017). Williams and Bowman desired a benchmark that would test the abilities of LLMs to understand entailed information from an entire paragraph of discourse – 2 to 3 sentences worth — as that was more on par with how entailment is encountered in human conversation. Another aim of the MultiNLI was to provide a way of comparing the entailment catching abilities between different LLMs, something of which the SNLI was not complicated enough to do (Williams, et al, pg 2, 2017).

### **NLI & its history: GLUE**



The current work on the MultiNLI did progress further, but not so much in the form of the test itself improving, but rather what it was incorporated into. Samuel Bowman and Alex Wang of the University of Washington later took Bowman's original MultiNLI benchmark, as well as eight other benchmarks — including the original PASCAL RTE — for testing an LLMs' abilities of other aspects of NLU and NLI, and put them together as components of a new test known as the General Language Understanding Evaluation — GLUE (Wang, et al. pg 3-4, 2018).

### **NLI & its history: SuperGLUE**

The GLUE test has gone on to become a very helpful benchmark for determining the abilities of LLMs. Yet, with the advent of advanced LLMs like ChatGPT and BERT, the GLUE test was updated once again in 2019, now known as the SuperGLUE test. While much between the GLUE and SuperGLUE changed, what is pertinent to the field of entailment is that the PASCAL RTE has been maintained as a part of the SuperGLUE test, where as the things tested by the MultiNLI have since been split up over two tasks, one known as the MultiRC and the other as the BoolQ. (Wang, et al, pg 4-6, 2019).

### **Looking forward**

I wish to build my own project continuing off from there, as there are two issues that I take with current benchmarks. One, when benchmarks are created, they eventually end up being used as a training set. This undermines the integrity of that benchmark as a good measurement, as there is the possibility that models that perform well on these tests are merely regurgitating training data rather than actually understanding the concept of entailment. Two, human language is generative and can be infinitely complex. Thus, not every scenario of entailment can be covered in one of these tests.

Overall, there is a lack of novelty in current entailment tests. I see a need to create my own data set that covers a very odd or niche linguistic context that isn't encountered in these current benchmarks.

There is also the added question of if an LLM can utilize entailment in a novel manner itself. The work done so far with entailment NLI has been to test whether an LLM can catch where entailment is used and how it is used in text that it reads. However, there is nothing current to my knowledge that is testing an LLMs own ability to create overt statements that have an implied context — and whether or not it can come up with its own examples or if it will simply regurgitate data from some previous benchmarks.

Overall, entailment inference is a very interesting and niche facet of NLI. While it may seem too small to be important, that is hardly the case for instances where an LLM is utilized for Q&A. Humans rely on the Cooperative Principle (Grice, 1975) for conversation to occur between humans, built upon understanding implied information in just about every speech act we make. In a future world that sees increasing interactions between humans and LLMs in everyday aspects, these LLMs need to be better equipped to handle humans and our way of conversating. LLMs that fail to understand humans and their implicature in any natural language scenario can only lead to frustration, anger, and inefficiency. Thus, it is imperative that the NLI limits of LLMs are pushed further in all realms, including entailment inference, such that these systems can be adequate for servicing the general public in the future.

## **Methodology:**

The methodology of my thesis style capstone project is over the performance of various LLMs (Large Language Models) on newly designed benchmarks for testing abilities in understanding linguistic entailment.

This thesis will require a mixed methodology; ie: of both quantitative and qualitative elements in the form of a designed experiment. There will be three main components in this project's approach, the identification of gaps (what to focus the experiment on), creation and running of novel NLI tests (the experiment part), compilation and comparison of results (discussion).

The goal of this experiment will be to answer the questions of 1) What are the current blind spots of already existing entailment benchmarks? 2) What pragmatic concepts can be used in (a) new entailment benchmark(s) to fill those blind spots? 3) How would LLMs trained to the entailment benchmarks with blindspots preform on the novel entailment benchmarks that aim to cover pragmatic material that resides within those blindspots and what could that say about flaws in current benchmarks?

First, there will be a qualitative assessment of what pragmatic and semantic concepts are currently covered by the entailment benchmarks available; ie: Pascal RTE, Stanford RTE, MultiNLI, and SNLI. This will be accomplished by either reading through the data sets themselves or through a summarization of its contents if one is provided. I will be working to make note of potentially novel concepts that aren't covered by these benchmarks. If finding the gaps of the entailment benchmarks from a qualitative perspective is not feasible, it might be possible to find them from a quantitative perspective by mathematically mapping out the latent space of various LLMs trained to preform well on current tests and seeing what gaps appear in

the shape created. This section would answer research question 1, which pertains to merely identifying what current entailment benchmarks lack.

Second comes the creation and running of novel entailment benchmarks that aim to fill these proposed gaps. This phase of the project will begin as qualitative as a create questions that entail certain understanding modeled on cognitive linguistics concepts that could potentially exist in those unfilled gaps. The project would continue to be qualitative as these questions are put into a benchmark – most likely to be programmed into Python – such that they can be run as tests on LLMs. However, the quantitative elements of this phase would be apparent at this stage. As the novel entailment question would consist of a proposition and a hypothesis — the entailed information — and prompted to the LLM as a binary/polar question of either True — the proposition does entail the hypothesis — or False — the proposition does NOT entail the hypothesis. The results of how an LLM preforms on this True-False test would be a quantitative measure of its NLI abilities on novel entailment examples. This part of the experiment would set up the answer research question 2, in that new benchmarks are created.

Finally, the compilation and comparison of section of the experiment would again be a mixed methodology. The comparison of results of different LLMs on the same novel test will involve quantitative values, but be primarily qualitative in what is reasoned. Furthermore, if a comparison was done on the latent space representations of these LLMs after being trained to hit these new benchmarks, that would be considered to quantitative approach to backing up the qualitative assessments of what gaps in NLI that my constructed data sets would fulfill. This would answer research question 2 in that if there is a difference in performance percentages of LLMs between the old benchmarks and a novel benchmark, then that specific novel benchmark exposed something that is not covered in the old benchmark — as LLMs trained to the old

benchmark couldn't cover something they haven't seen before. If there is no difference, then that also shows that that novel test did not expose a gap. Irregardless, research question 3 would be answered as performance data would be gathered.

In conclusion, by building novel entailment benchmarks for testing the NLI abilities of various LLMs, current failures in conversational understanding by technologies that employ use of those LLMs will have better understood origins. This could aid in mitigating understanding failures in the future and pointing out nuanced places where LLMs can be trained to be better. While there are the limitations of time, computing power, and needing to learn the Python programming language, these can be overcome by; creating a robust schedule/Gantt chart, accessing a desk-top computer, and taking the Algorithms course offered by the UCA Computer Science Department respectively.

## Timeline:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	2024	2025													2026			
	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec		Jan	Feb	Mar	Apr
Part 1: Finding the Gaps																		
Learn Basic Python																		
Qualitatively Identifying Covered Subjects in Current Benchmarks																		
Attempt to Run Old Benchmark Tests on various LLMs																		
Modelling the various LLMs in latent space to see if quantitative gaps align with qualitative gaps																		
Part 2: Creating the New Benchmarks																		
Making Psuedocode																		
Writing Code that can run the novel benchmarks on LLM																		
Designing Questions for Novel Benchmark #1 (NB1)																		
Will Consist of 3 Sections; aim of 100 questions per section																		
NB Section 1 Question Creation																		
NB Section 2 Question Creation																		
NB Section 3 Question Creation																		
Putting Questions of NB in a file format that can be run in LLM testing code																		
Testing LLM performance on NB																		
Addressing Bugs and Potential Issues																		
Part 3: Grading the LLMs on the New																		
Analysis of Data and Discussion																		
*Attempt to map LLMs in Latent Space after being graded on NB																		
Part #4: Writing the Capstone Thesis																		
Zero (Rough) Draft																		
Clean Draft																		
Final Presentation																		

## Mentor-Mentee Agreement:

Norbert O. Schedler Honors College

Capstone Project Mentor-Mentee Agreement

Mentee Name: Kyle Urban

Mentor Name: Dr Zachary Stine

Project Title: Entailment in Large Language Models: A Case for Novelty

Current Semester: Fall 2024

Expected Semester of Capstone Completion: Spring 2026

Semester Meeting Schedule: We will meet every Wednesday at 1:30 pm in MCS 322; if not, then another time will be found during the week (typically on Tuesdays)

Project description:

The recent prominence of Large Language Models has emphasized the greater needs for improvements in the field of Natural Language Inference. The abilities of various Large Language Models on understanding entailment need to be expanded from the current benchmarks in order to better

reflect a larger array of possible use cases of human language. As such, this capstone project seeks to answer the three research questions: (1) What cases of entailment do current NLI benchmarks not cover, (2) How can those voids be filled with a new benchmark, and (3), how will LLMs tested on old benchmarks perform on the new benchmark, and what does that say about their NLI abilities?

#### Project work goals:

The project work goals for the Fall 2024 Semester will be: (1) to have a complete literature review and project proposal with annotated bibliographies of key literature in the field, (2) A solid framework for how to elicit and sort the data in a manner that is both accurate and efficient, (3) have found a practical way to start creating that elicitation framework, and (4) have created a proper calendar for deadlines to hit in following semesters.

#### Mentor-Mentee expectations:

During regular meetings, the mentee will relate understandings, opinions, and elaborations on research. The mentor will question, clarify, and criticize the mentee's efforts with the goal of advancing the mentee's understanding and helping the project progress.

#### Reading and research:

The mentee will read works suggested by the mentor and other works that the student may find relevant to the study. These works are identified in the mentee's working bibliography/references list. This semester, specific works include:


1. Na, Robin, et al. "The Diversity of Argument-Making in the Wild: from Assumptions and Definitions to Causation and Anecdote in Reddit's 'Change My View'" *arXiv*. 16 May, 2022. <https://doi.org/10.48550/arXiv.2205.07938>
2. MacCartney, Bill, "NATURAL LANGUAGE INFERENCE" Stanford University. Jun, 2009. <https://nlp.stanford.edu/~wcmac/papers/nli-diss.pdf>
3. Wang, Alex, et al. "GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING" *arXiv*. 20 Apr, 2018. <https://arxiv.org/pdf/1804.07461>
4. Hu, Jennifer, et al. "Expectations over Unspoken Alternatives Predict Pragmatic Inferences" *Transactions of the Association for Computational Linguistics*. 7 Apr, 2023. <https://doi.org/10.48550/arXiv.2304.04758>
5. Bowman, Samuel, et al. "A large annotated corpus for learning natural language inference" In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 21 Aug. 2015. pp 632–642. <https://doi.org/10.48550/arXiv.1508.05326>
6. Williams, Adina, et al. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference" *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Jun 2018. pp 1112–1122.  
<https://doi.org/10.18653/v1/N18-1101>

Semester goals:

Our goal for the semester will be for the mentee to complete the reading and research listed above, prepare an annotated bibliography, and finalize a Proposal. Specific deadlines include:

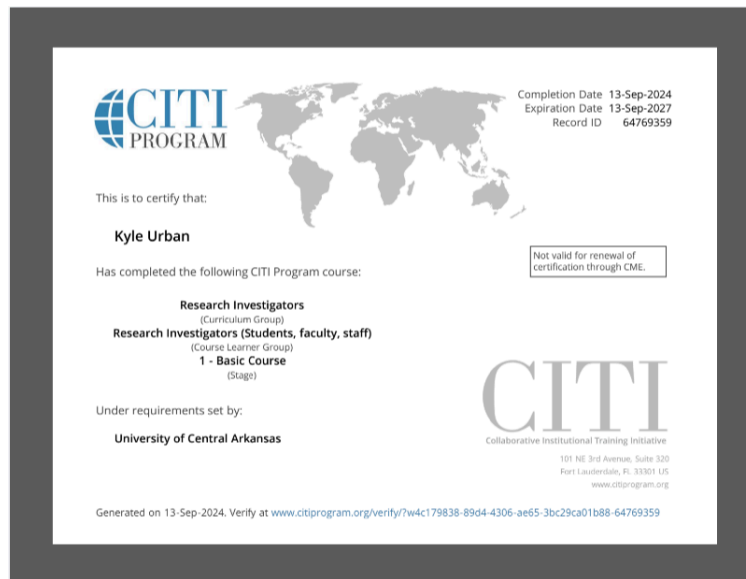
- Mentee will complete literature review Complete by week of October 28th
- Mentee will submit Rough Draft of Final Research Proposal to mentor by week of Nov 12th
- Mentor provide feedback to the mentee on final draft before Nov 25th
- Mentee will submit the final Proposal in the student portal for the mentor's review by Monday of week 15 (Dec 2), the final week of classes.

Mentee Signature:   
Date: 11/26/2024

Mentor Signature:   
Date: 11/26/2024



## Research Compliance:



## Complete Bibliography:

- Ackerman, Samuel, et al. "Using Combinatorial Optimization to Design a High quality LLM Solution." *arXiv*. 15 May, 2024. <https://doi.org/10.48550/arXiv.2405.13020>
- Alt, Tobias, et al. "Generative AI Models: Opportunities and Risks for Industry and Authorities" *Federal Office for Information Security*. 4 Apr. 2024, ver 1.1. <https://doi.org/10.48550/arXiv.2406.04734>
- Amirjalili, Forough, et al. "Exploring the boundaries of authorship: a comparative analysis of AI-generated text and human academic writing in English literature." *Front Educ*. 24 Mar, 2024. vol 9, <https://doi.org/10.3389/feduc.2024.1347421>
- Baker, Collin, et al. "The FrameNet Data and Software" *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*. 7 July, 2003, pp 161-164. <https://doi.org/10.3115/1075178.1075206>
- Basnov, Victoria, et al. "Simple Linguistic Inferences of Large Language Models (LLMs): Blind Spots and Blinds." *arXiv*. 11 Apr, 2024. <https://doi.org/10.48550/arXiv.2305.14785>
- Bergey, Claire. et al. "Learning Communicative Acts in Children's Conversations: A Hidden Topic Markov Model Analysis of the CHILDES Corpora" *Topics in Cognitive Science* 14. 2022. pp 388-399 <https://doi.org/10.1111/tops.12591>
- Bystrov, Dmitriy, et al. "FUZZY SYSTEMS FOR COMPUTATIONAL LINGUISTICS AND NATURAL LANGUAGE." *NISS '20: Proceedings of the 3rd International Conference on Networking, Information Systems & Security*. 18 May, 2020. no. 54, pp. 1-3. <https://doi.org/10.1145/3386723.3387873>
- Borji, Ali et al. "A Categorical Archive of ChatGPT Failures." *Quintic AI*. 3 Apr, 2023. <https://doi.org/10.48550/arXiv.2302.03494>

- Bowman, Samuel, et al. "A large annotated corpus for learning natural language inference" *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 21 Aug. 2015. pp 632–642. <https://doi.org/10.48550/arXiv.1508.05326>
- Casal, Elliot, & Kessler, Matt. "Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing" *Research Methods in Applied Linguistics*. vol. 2, issu. 3. Dec, 2023. pp 1-21. <https://doi.org/10.1016/j.rmal.2023.100068>
- Casey, Mike. "Generative AI Adoption Levels Assistance Fall 2024." *University of Central Arkansas*. 2024. <https://uca.edu/cetal/files/2024/08/AI-Adoption-Levels-Fall-2024.pdf>
- Dagan, Ido. et al. "The PASCAL recognising textual entailment challenge" *Springerlink*. Jan. 2005. pp 177–190. [http://dx.doi.org/10.1007/11736790\\_9](http://dx.doi.org/10.1007/11736790_9)
- Dasgupta, Ishita, et al. "Evaluating Compositionality in Sentence Embeddings" *arXiv*. May, 2018. <https://doi.org/10.48550/arXiv.1802.04302>
- Evans, Vyvyan, & Green, Melanie. "Cognitive Linguistics An Introduction" *LAWRENCE ERLBAUM ASSOCIATES*. 2006. pp 401-467.
- Frank, C, Micheal. & Goodman, Noah. "Predicting Pragmatic Reasoning in Language Games" *ScienceMag*. 25 May, 2012. vol, 336. pg 998. <https://www.science.org/doi/10.1126/science.1218633>
- Fillmore, Chuck. "Frame Net." *University of California, Berkeley*. 2024. <https://framenet.icsi.berkeley.edu/CJFFNintroPPT>
- Fyfe, Paul. "How to cheat on your final paper: Assigning AI for student writing." *AI & Society*. 10 Mar, 2022. vol. 38, pp 1395-1405. <https://doi.org/10.1007/s00146-022-01397-z>
- Fyodorov, Yaroslav, et al. "A Natural Logic Inference System" *In Proceedings of the 2nd Workshop on Inference in Computational Semantics*. Dec, 2000. [https://www.researchgate.net/publication/2454386\\_A\\_Natural\\_Logic\\_Inference\\_System](https://www.researchgate.net/publication/2454386_A_Natural_Logic_Inference_System)
- Grice, H.P. "Logic and Conversation" *University of California, Berkeley*. 1975. pp 41-58.
- Guo, Shaoru, et al. "NutFrame: Frame-based Conceptual Structure Induction with LLMs" *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. May, 2024. pp 12330–12335. <https://aclanthology.org/2024.lrec-main.1079.pdf>
- Horn, Laurence, "A NATURAL HISTORY OF NEGATION" *University of Chicago Press*. 1989. <https://emilkirkegaard.dk/en/wp-content/uploads/A-natural-history-of-negation-Laurence-R.-Horn.pdf>
- Horn, Laurence & Ward, Gregory. "The Handbook of Pragmatics" *Blackwell*. 2006. <https://www.felsemiotica.com/descargas/Horn-Laurence-R.-and-Ward-Gregory-Ed.-The-Handbook-of-Pragmatics.pdf>
- Hu, Jennifer, et al. "A Systematic Assessment of Syntactic Generalization in Neural Language Models" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. July, 2020. pp 1725–1744. <https://doi.org/10.18653/v1/2020.acl-main.158>

- Hu, Jennifer, et al. "Expectations over Unspoken Alternatives Predict Pragmatic Inferences" *Transactions of the Association for Computational Linguistics*. 7 Apr, 2023. <https://doi.org/10.48550/arXiv.2304.04758>
- Hu, Jennifer, et al. "A fine-grained comparison of pragmatic language understanding in humans and language models" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jul 2023. pp 4194–4213. <https://doi.org/10.18653/v1/2023.acl-long.230>
- Jacquet, Baptiste, et al. "Cooperation in Online Conversations: The Response Times as a Window Into the Cognition of Language Processing" *Front. Psychol.* 8 Apr, 2019. vol 10. <https://doi.org/10.3389/fpsyg.2019.00727>
- Jacquet, Baptiste, et al. "The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots" *International Conference on Information and Digital Technologies*. Jun, 2019. pp. 180-189. <http://dx.doi.org/10.1109/DT.2019.8813473>
- Jeretic, Paloma, et al. "Are Natural Language Inference Models IMPPRESSive? Learning IMPLICature and PRESupposition" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Jul 2020. pp 8690–8705. <https://aclanthology.org/2020.acl-main.768.pdf>
- Jiang, Nanjiang. & de Marneffe, Marie-Catherine "Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment" *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. pp 4208–4213. <https://aclanthology.org/P19-1412.pdf>
- Kaas, Marten, H. L. & Halbi, Ibrahim. "Assuring AI Safety: fallible knowledge and the Gricean Maxims." *AI and Ethics*. 2024. pp. 1-14. <https://link.springer.com/article/10.1007/s43681-024-00490-x>
- Lai, Tin. et al. "Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs)" *BioMedInformatics*. 2024. ver 4, pp 8-33. <https://doi.org/10.3390/biomedinformatics4010002>
- Liang, Claire, et al. "Implicit Communication of Actionable Information in Human-AI teams." *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2 May, 2019. no. 95, pp 1-13. <https://doi.org/10.1145/3290605.3300325>
- Lokshyn, Oleh. "Natural Language Inference: An Overview" *Towards Data Science*. 9 Jun, 2022. <https://towardsdatascience.com/natural-language-inference-an-overview-57c0eecf6517>
- Ma, Yongqiang, et al. "AI v.s Human - Differentiation Analysis of Scientific Content Generation" *Wuhan University*. 24 Jan 2023. <https://doi.org/10.48550/arXiv.2301.10416>
- MacCartney, Bill "Natural Logic and Alignment in Natural Language Inference" *Microsoft Research Youtube Channel*. Posted 2016; Recorded 2011. <https://www.youtube.com/watch?v=nJXSig3hYtM&t=17s> and <https://www.microsoft.com/en-us/research/video/natural-logic-and-alignment-in-natural-language-inference/>

- MacCartney, Bill, “NATURAL LANGUAGE INFERENCE” *Stanford University*. Jun, 2009.  
<https://nlp.stanford.edu/~wcmac/papers/nli-diss.pdf>
- McCoy, Thomas, et al. “Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Jul. 2019. pp 3428–3448.  
<https://doi.org/10.18653/v1/P19-1334>
- “Mechanical Turk” *Amazon*. 2018. <https://www.mturk.com/>
- Michael, Mollel. “Enhancing Retrieval-Augmented Generation: Tackling Polysemy, Homonyms and Entity Ambiguity with GLiNER for Improved Performance” *Medium*. 16 Mar, 2024.  
<https://medium.com/@mollelmike/enhancing-retrieval-augmented-generation-tackling-polysemy-homonyms-and-entity-ambiguity-with-0fa4d395c863>
- Miehling, Erik, et al. “Language Models in Dialogue: Conversational Maxims for Human-AI Interactions.” IBM Research. 22 Mar, 2024. <https://arxiv.org/pdf/2403.15115v1>
- Mihalcea, Rada, et al. “The SENSEVAL–3 English Lexical Sample Task” *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Jul, 2024. pp 25-28. <https://aclanthology.org/W04-0807.pdf>
- Na, Robin, et al. “The Diversity of Argument-Making in the Wild: from Assumptions and Definitions to Causation and Anecdote in Reddit’s ‘Change My View’” arXiv. 16 May, 2022. <https://doi.org/10.48550/arXiv.2205.07938>
- Nair, Sathvik, et al. “Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge”. *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*. Dec 2020. pp 129-141. <https://aclanthology.org/2020.cogalex-1.16/>
- Navigli, Roberto. “Word Sense Disambiguation: A Survey” *ACM computing surveys (CSUR)*. 23 Feb, 2009. vol 41, iss 2, pp 1-69. <https://dl.acm.org/doi/pdf/10.1145/1459352.1459355>
- Neeraj, Trishala. “Semantic Entailment” *Github*. 22 Dec, 2017.  
<https://trishalaneeraj.github.io/2017-12-22/semantic-entailment>
- Noble, Bill. et al. “Classification systems: Combining taxonomical and perceptual lexical meaning” *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*. Aug. 2022. pp 11-16. <https://aclanthology.org/2022.naloma-1.pdf>
- Panfili, Laura, et al. “Human-AI Interactions through a Gricean lens” *Proceedings of the Linguistic Society of America*. 20 Mar, 2021. vol 6. no. 1. pp. 288-302.  
<https://doi.org/10.3765/plsa.v6i1.4971>
- “PASCAL Recognizing Textual Entailment Challenge (RTE-5) at TAC 2009” *National Institute of Standards and Technology*. <https://tac.nist.gov/2009/RTE/>
- “Past RTE Data” *National Institute of Standards and Technology*.  
[https://tac.nist.gov/2009/RTE/past\\_data/index.html](https://tac.nist.gov/2009/RTE/past_data/index.html) — **DATA SET TO TEST ENTAILMENT**
- Petroni, Fabio, et al. “Language Models as Knowledge Bases?” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Sep, 2019. pp 2463–2473. <https://doi.org/10.48550/arXiv.1909.01066>

- Qiu, Zhuang, et al. "Does ChatGPT Resemble Humans in Processing Implicatures?" *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*. 20 June, 2023. pp 25-34. <https://aclanthology.org/2023.naloma-1.3.pdf>
- Raganato, Alessandro, et al. "Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison" *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Apr, 2017. pp 99-110. <https://aclanthology.org/E17-1010/>
- Rapp, Amon. et al. "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots" *International Journal of Human-Computer Studies*. Jul. 2021, vol 151. <https://doi.org/10.1016/j.ijhcs.2021.102630>
- "REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS." *European Union*. 21 Apr, 2021. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
- Saeed, John. "Semantics: Fourth Edition" *WILEY Blackwell*. 2016.
- Sardinha, Tony. "AI-generated vs human-authored texts: A multidimensional comparison" *Applied Corpus Linguistics*. Apr, 2024. vol 4, iss. 1, <https://doi.org/10.1016/j.acorp.2023.100083>
- Schopf, Tim, et al. "Exploring the Landscape of Natural Language Processing Research" *arXiv*. Sep, 2023. <https://doi.org/10.48550/arXiv.2307.10652>
- Seff, Ari. "How ChatGPT is Trained." Youtube. 24 Jan, 2023. <https://www.youtube.com/watch?v=VPRSBzXzavo>
- Setlur, Vidya, & Tory, Melanie. "How do you Converse with an Analytical Chatbot? Revisiting Gricean Maxims for Designing Analytical Conversational Behavior." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 16 Mar, 202. pp 1-17. <https://doi.org/10.48550/arXiv.2203.08420>
- Shen, Xinyue, et al. "In ChatGPT We Trust? Measuring and Characterizing the Reliability of ChatGPT." *arXiv*. 5 Oct, 2023. <https://doi.org/10.48550/arXiv.2304.08979>
- Sieker, Judith & Zarriß, Sina. "When your Language Model cannot even do Determiners right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle." *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Dec, 2023. <https://doi.org/10.18653/v1/2023.blackboxnlp-1.14>
- Stasaski, Katherine & Hearst, Marti. "Semantic Diversity in Dialogue with Natural Language Inference." *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Jul 2022. pp 85-98. <https://doi.org/10.18653/v1/2022.naacl-main.6>
- Ujhelyi, Adrienn, et al. "Would You Pass the Turing Test? Influencing Factors of the Turing Decision." *Psiholgijske teme*. 2022, vol. 31, no. 1, <https://doi.org/10.31820/pt.31.1.9>

- Wang, Alex, et al. “GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING” *arXiv*. 20 Apr, 2018.  
<https://arxiv.org/pdf/1804.07461>
- Wang, Alex, et al. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” *33rd Conference on Neural Information Processing Systems*. 2019. <https://w4ngatang.github.io/static/papers/superglue.pdf>
- Wang, Yuqing & Zhao, Yun. “Metacognitive Prompting Improves Understanding in Large Language Models” *NAACL 2024*. 20 Mar, 2024.  
<https://doi.org/10.48550/arXiv.2308.05342>
- Williams, Adina, et al. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Jun 2018. pp 1112–1122. <https://doi.org/10.18653/v1/N18-1101>
- “Why you should use caution with AI detectors.” *University of Kansas*. 2024.  
<https://cte.ku.edu/careful-use-ai-detectors>
- Yue, Shisen, et al. “Do Large Language Models Understand Conversational Implicature – A case study with a Chinese sitcom” *2024 China National Conference on Computational Linguistics*. 31 Jul. 2024, ver 2. <https://doi.org/10.48550/arXiv.2404.19509>
- Zadeh, Amir, et al. “Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*” Jun, 2019, pp. 8807-8817. <http://dx.doi.org/10.1109/CVPR.2019.00901>
- Zhang, Yuhuan. et al. “Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics” *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Dec. 2023. pp1-14.  
<https://doi.org/10.18653/v1/2023.conll-1.1>
- Zheng, Zilong. et al. “GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational Reasoning” *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Aug, 2021. pp 2074–2085.  
<https://doi.org/10.18653/v1/2021.findings-acl.182> & <https://zilongzheng.github.io/Grice/>