# A proof-of-concept method for Inconclusiveness-based Abstention *

**Sushma Anand Akoju**

University of New Hampshire
Durham, New Hampshire
sushmaanandakoju@proton.me

## Abstract

Recent studies evaluating Large Language Models (LLMs) on abstention under uncertainty suggests the problem of reasoning under uncertainty remains unresolved even though In-context Learning (ICL) improved abstention in LLMs Kirichenko et al. (2025). For evaluating logical inconclusiveness, "refute a query" is selected under abstention Wen et al. (2024). I propose a proof-of-concept method to create example dataset that is inconclusive but is "satisfiable". I created two example Knowledge Base (KB) sets - with a polysemy noun and another using Wumpus world in 1. The model is asked to choose from multiple answer choices containing - True if this and inconclusive if that. LLMs ground $bat\_mammal$, $baseball\_bat$ as terms early on and deduce true instead of inconclusive, unlike the case with Wumpus world. LLMs may not seem to divulge from standard commonsense reasoning (though such information might be available during reasoning), i.e. once grounded in world knowledge across several variations of same example KB, the inconclusivess is not detected. However, under Wumpus world context, the three LLMs accurately detected the inconclusiveness for the example KB in the context. The preliminary two-KB analyses over three LLMs, hint that a combination of - commonsense, logical and lateral reasoning under uncertainty, might nudge them towards detecting inconclusiveness for real world context which requires elaborate evaluation and analyses. (Method, example KBs: in 4).

**Example Dataset** — https://tinyurl.com/inconclusive-sub2
**Example script to generate data** —
https://tinyurl.com/inconclusive-data-repo
**Example prompt** —
https://tinyurl.com/inconclusive-ex-prompt

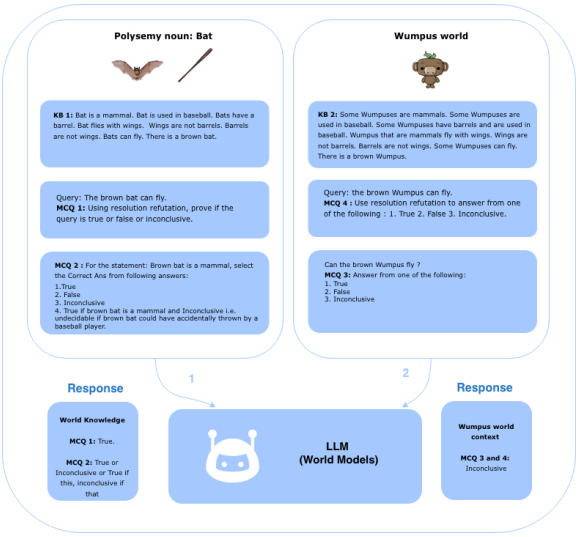


Figure 1: Example: KB with Polysemy noun and a Multiple Choice Question (MCQ) versus Wumpus world with syllogisms. LLMs may seem to deduce inconclusive for non-real world context but may appear confused when grounded in real world contextual meaning.

## Introduction

Abstention is the ability to not answer under uncertain user queries, that are ill-posed, under specified, outdated or missing information, or fundamentally unanswerable. The fundamental ability to *not* answer a question or to answer with a response "I don't know" has been an important aspect of reasoning. There are various uncertainty conditions that require abstention and recent works Kirichenko et al. (2025) suggest the problem of reasoning under uncertainty remains unresolved even though carefully crafted prompts and In-context Learning (ICL) improved abstention in LLMs. For example, asking about unclear/invisible objects in a given input image, ends up being answered by an LLM with a definitive yet wrong answer Chandu et al. (2024). Large Language Models have been studied and evaluated for answers to queries about missing or outdated information and have shown hallucinations Feng et al. (2024). Probing other LLMs for knowledge gaps with collaborative or competitive approaches benefited and improved abstention performance Feng et al. (2024).

The uncertainty-based abstention in LLMs is shown to improve safety and reduce hallucinations based on appropriate uncertainty measure applied Tomani et al. (2024). The

*Submitted to AAAI 2026 LMReasoning Bridge program and received feedback. The feedback is not yet incorporated. However this version is updated with toy dataset, example prompts and scripts to generate data for evaluation as a proof of concept.

challenges discussed in Anwar et al. (2024) on the popular In-Context Learning (ICL) technique as a learned optimization has been recognized as an emergent behavior so far the model is able to implement a learning algorithm within its weights. However even when ICL and variants of ICL Shukor et al. (2024) when used alongside well crafted prompts, LLMs fail to address the problem of abstention in practice. Kirichenko et al. (2025), Saadat et al. (2024)

## Background

Recent works suggest that the LLMs have scope for improvement on graph-aligned data, planning during deductive reasoning, when selecting nodes/paths Saparov and He (2022) Saparov et al. (2023), Saparov et al. (2024). The findings from Saparov and He (2022) shows that LLMs are capable of making correct deductive individual steps but suffer from proof-planning.

Given the rapid improvements within families of LLMs such as OpenAI, Anthropic, Google and others, there is a saturation of benchmarks Deveci and Ataman (2025). When those benchmarks do not add any evaluative value, since the benchmarks do not discriminate over the models, then they seem to be marked as saturated, particularly within the single family of LLMs. Such an evaluative saturation needs newer benchmarks that considers the research community-prioritized reasoning definitions and evaluation methods. The work Deveci and Ataman (2025) suggests some specific reasoning tasks still remain a challenge. Once a benchmark reaches a certain target performance, the same benchmark is used less frequently or discontinued Deveci and Ataman (2025). Recent research studies suggest emergent misalignment behavior observed from generating insecure code without disclosure to the user Betley et al. (2025). More recent research studies suggest that LLMs exhibit natural emergent misalignment by learning to reward hack that leads to egregious emergent misalignment MacDiarmid et al. (2025). Inoculation prompting is a recent method introduced to address emergent misalignment which may be used during training known as Inoculated training Tan et al. (2025). These recent studies seems to hint towards possible new changes that might occur in future versions of LLMs. This seems to suggest additional analysis and efforts required in framing the problem, developing the proof-of-concept for evaluation, that could not only challenge/test current LLMs but their future versions.

In this work, I explore and verify multiple answer choice questions as prompts that requires a combination of commonsense, lateral and logical reasoning, across with several variations of the KB, queries and answer choices (only select variations of prompts were included in the Appendix). The key questions to explore are:

- To understand and apply what necessary preliminary analysis is required to develop a proof-of-concept method to generate a novel benchmark that may improve the challenge for the LLMs ?
- To propose a benchmark that may last the rapidly evolving versions within LLM families and across various types of LLMs ?
- Given the limited computational resources and limited access to paid/extensive subscriptions, commonly facing the usage of the current closed-source LLMs (and their future versions), what effective approaches exist to continue student's research to evaluate LLMs ?
- What challenges exist in developing a proof-of-concept method when newer versions of LLMs are released while the existing open challenges, existing and evolving limitations of LLMs for reasoning continue to haunt ? Betley et al. (2025)

## Methodology

Earlier I created an example KB with multiple variations to decide how to design and frame the question that should result in inconclusiveness over Claude 3.x, GPT-4o and other LLMs. However from my analysis, the older versions are not available and given the recent Claude 4.x GPT 5x and Gemini 3x, I have redesigned the methodology.

To attempt to evaluate the latest versions across and within the LLM families, I attempt to further narrow and re-frame the "refuting a query" as a problem that is inconclusive but satisfiable (SAT) as an abstention. I attempted this by re-designing and modifying a single KB and query to explore the intersection of logical, commonsense and lateral reasoning. Refuting a query under incomplete/confusing knowledge base, does not only require logical reasoning, but also commonsense and lateral reasoning. Lateral reasoning/thinking requires deviating from the common "paths" i.e. "outside the box" Jiang et al. (2023) that relies on uncommon information hidden in the problem/query, while logical reasoning requires step by step reasoning based on facts and rules of logic/inference. For this purpose, I consider a small hand-curated knowledge base using polysemy noun as a keyword for word sense disambiguation. I chose "bat" as the keyword. I designed the KB as follows:

"Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat."

I used different types of answer choices for several variations of the query and by using aforementioned knowledge base as standard part of the prompt input to three different recent versions of three different families of LLMs, I explored proof-of-concept method required to develop a potential new dataset. I manually entered these prompts, given the lack of computational resources as well as limitations on API calls that exhausted quickly. Therefore I re-ran all prompts that I created, by manually running in the Conversational AI applications for each one of the LLMs directly, for consistent, reproducible approach.

### Using Resolution Refutation

Resolution is refutation complete, if a set of clauses, that are in first-order skolemized clausal normal forms, are *unsatisfiable*, since the resolution derives empty clauses and verifies unsatisfiability. But resolution is not decidable if the set of clauses may be satisfiable, resolution refutation (RR) procedure may never terminate. Traditional RR does not have world knowledge unless explicitly included in the

KB and do not consider word sense disambiguation especially for words of same part-of-speech tags since skolemized clausal forms are not meaning preserving Akama and Nantajeewarawat (2021). I included explicit equivalence that preserves meanings between two key features: as "...Wings are not barrels. Barrels are not wings...". More details in 4.

### Types of questions, and variations in the query for including of various types of reasoning

I included broadly, six types of questions based on a query for the given KB:

- **With and without Resolution refutation:** In the first method I add a query to the KB with a query with multiple steps and two questions where one question uses resolution refutation and other does not. The questions ask to refute the query (by answering if true/false/inconclusive). Prompt 1.
- **Without Resolution refutation:** In the second method I add a query to the KB as a question without any steps. Prompt 2.
- **With Resolution Refutation:** In the third method, I used the same query with the resolution refutation alone as a question to answer if the query is true or false or inconclusive. "Query: The brown bat can fly. Using resolution refutation, prove if the query is true or false or inconclusive.". This method focused on resolution refutation as a method to seek inconclusiveness, under world knowledge and commonsense reasoning capabilities alongside lateral and logical thinking. Prompt 3.
- **With Images:** In the fourth method, I supplied two images, one that of a flying brown bat that is a mammal, another that of a brown baseball bat. I added a question "By looking at the KB and two images, answer if the statement "Brown Bat can fly" true or not or inconclusive?". Prompt 4.
- **With valid diversion:** In this method, to avoid the introduction of "can" as the ability to fly which may often trigger birds/mammals that can fly (based on world knowledge), I instead introduce a sentence: "brown bat flew.". I added a different question this time, to attempt to divert away from the commonsense assumption and added query statement "brown bat is a mammal" and asked the LLM to select the correct answer from following answer choices: "For the statement: Brown bat is a mammal, select the Correct Ans from following answers: 1. True 2. False 3. Inconclusive 4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player.". Prompt 5.
- **With confusing choices:** In the sixth method, I added two additional confusing answer choices, which is an attempt to make LLMs fall back to commonsense: "5.Answer 4 and other answers with other reasoning 6.Combination of answer under specific assumptions not included in aforementioned answers". Prompt 6.
- **Without real world context:** In the seventh, eight and ninth methods, as in 4 prompts 7 to 9, I replaced any references to polysemy noun "bat" with wumpus nouns that do not have any real world references. For prompts 8 and 9, to introduce the confusion for inconclusive answer, I introduced syllogisms. Both with and without resolution refutation, prompts 8 and 9 are inconclusive. Prompts 7, 8 & 9.

## Results

### Preliminary results based on Claude-4.5 Gpt-5.1 and Gemini 3.0 pro model families

I supplied different prompts designed to help insufficient information for deducing inconclusiveness discussed in , following are the preliminary observations:

- Each one of the prompts described in 4, were repeated five times on each one of the three models respectively: Claude Sonnet 4.5, Gemini 3.0 Pro and GPT 5.1.
- Both Gemini 3.0 pro and GPT 5.1, commonly always choose True as the correct answer. Both of the models appeal to common sense reasoning of the mammal as a subset of animal family, brown bat as a specific species that can fly, bat has wings. The semantic grounding of the word "can fly" to "ability to fly" as a common reason for selecting the answer choice that any instance of a "Brown Bat" in "Brown bat can fly" would only lead to Brown bat as a mammal. No deviations/creative variations of reasoning were explored by the two LLMs.
- Gemini 3.0 pro interestingly, when running the KB with queries but without resolution refutation, grounded its reasoning with visual images of bat and wings of bat as labeled image for reasoning trace.
- With Gemini-3.0 Pro and GPT-5.1, using resolution refutation (as in prompt 3 under 4), did not lead towards inconclusiveness as the answer. As for Claude-4.5 and GPT-5.1 models in comparison to their respective predecessors i.e Claude-3x, GPT-4x models, did derive Inconclusive or True under resolution refutation due to initial grounding of clausal forms distinguishing the two bats based on world knowledge. The reasoning traces for prompt numbered 5 4 did include reasoning that a baseball ball could also fly not because of the ability to fly by itself.
- For prompt 5 in 4, Claude-4x and Claude-4.5 models, demonstrate similar confusion between True or Inconclusive or both, suggesting commonsense reasoning or logical and lateral reasoning combined or both. In this specific case, based on reasoning produced by the Claude-4.5, it seems likely that logical and lateral reasoning combined with commonsense reasoning may have led them to choosing "True if this and Inconclusive if that" 4, based on the reasoning trace. But none of the reasoning here uses resolution refutation.
- By inferring from the findings from aforementioned two points, the prompts 3 and 5 in prompts sectiabstentionon 4, it seems that combining the resolution refutation by instantiating the constants with commonsense grounding for the polysemy noun forms in-context can help trigger potentially lateral reasoning and may be combined with commonsense and logical reasoning. This combined

method maybe applied for generating a benchmark challenge to provide additional action statements similar to used in prompt 5 4 to help LLM for support, relevance and reasoning in-context, put together.

- Prompts 7 to 9 are based on Wumpus world and do not have any real world context. For prompt 7, all three models result in True.
- For prompt 8, which uses resolution refutation and uses syllogisms without world knowledge references, the results evaluate to Inconclusive by all three models.
- For prompt 9, which does not use resolution refutation but uses syllogisms without world knowledge references, the results evaluate to Inconclusive by all three models. This is when lateral thinking and logical reasoning using resolution refutation agree and result in inconclusive.

Proposed method to generate the benchmark with and without Resolution Refutation which requires to decompose the data generation process into multiple steps: one with commonsense grounding of the constants and predicates and then using Resolution refutation, where *brown_bat* $\neq$ *bat_mammal* is a required condition. This is to say any noun form polysemy word that has word sense disambiguation such as (bat, rose, may, mean) is suitable. I note that this method is different from that of the examples such as "She sells seashells at the seashore". The method here only notes a knowledge base with satisfiable by refutation queries but are inconclusive.

### Framework, Example Prompts and Example dataset

I propose a framework for generating a potential benchmark dataset for evaluation of LLM at the intersection of lateral reasoning and logical reasoning with and without references to real world. Steps as follows:

- Select a polysemy noun. Each one of the keywords in (bat, head, nail, paper) have two different meanings in the noun forms and with real world references. Example: bat.
- Select one fact each for each one of the two meanings for the each one of the keywords. Example: "Bat is a mammal. Bat is used in baseball." where this information is valid in real world.
- Select two distinct features of two different meanings of a given keyword and generate two new facts using "have" or "contain" depending on reference. Example: "Bats have a barrel. Bat flies with wings".
- Generate two equivalence clauses for two distinct features selected in previous step, for the two meanings of a given keyword. (**for ensuring meaning preservation by common sense, not refutation**). Example: "Wings are not barrels. Barrels are not wings.".
- Introduce a repeated re-assertion of a fact generated in step 3 with "can" or other relevant forms.
- Select a random color. Prepend the color to new instance of keyword and generate a query using this new colored instance to construct an introduction. For the query, use one of the two facts to replace the keyword with colored keyword instance *color* $\in$ *black, white, brown, red, yellow, blue, green, pink*. Example: "There is a Brown bat" or "There is a pink nail" and so on.
- Add an action or negated action sentence based on step 2, replacing the noun with colored keyword instance (Ex: brown bat). The idea is to add confusion such that the action/property that could distinguish two meanings of the polysemy noun is absent, leading to inconclusiveness. Examples: "There is a brown bat. Brown bat flew." or "There is a pink nail. Pink nail did not fasten the object."
- Using the facts in step 1 such as containment or $\exists$ assertions from facts, generate a query as a statement. Example: "Query: Brown bat is a mammal", or "Query: Pink nail is part of the finger". The resulting query "Is the statement "Brown bat is a mammal" true or false or inconclusive?".
- Generate a question with multiple choices can be framed as: "For the statement: Brown bat is a mammal, select the Correct Ans from following choices: 1. True, 2. False, 3. Inconclusive, 4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player.".
- Following these steps, I generated four KBs 4.
- Finally, for each one of the KBs generated replace each reference to the keyword in the KB with only one of the non-real-world keywords from (Umpus, Wumpus, Vumpus, Zumpus).
- Modify resulting KB in previous step to include syllogism and containment for creating context, similar to prompts 8 and 9.

I suggest the method that for real world context KBs: providing an answer choice as "True if this and Inconclusive if that" format similar to prompt 5 in 4. The example prompts and example dataset generation script are available here: 4

## Conclusion

Recent studied showed that LLMs fail in recognizing the uncertainty in the language for abstention. Recent studies suggest LLMs have room for improvement in lateral thinking such as riddles and puzzles. LLMs face a saturation of benchmarks within and across different model families and recent emergent misalignment and variation from rewards hacks, which hints towards what future versions of LLMs may seem like. This work underscores evaluation of what type of proof-of-concept method is required, how to attempt to take a calculated measure for assessing what type of method may benefit for adding an evaluative value. By using confusing references to polysemy nouns that are constants/predicates within the Knowledge base, may tap into world knowledge, commonsense reasoning, lateral reasoning paired with logical reasoning for detecting inconclusiveness. I proposed proof-of-concept method to create a KB with query that is satisfiable but inconclusive irrespective of whether grounded in World knowledge or not. The proposed method introduces steps to attempt to cover logical,

lateral and commonsense reasoning for deducing inconclusiveness. A future direction to this work involves generating the dataset with multiple choice questions (MCQs) (prompt 5) and with Wumpus world context (prompts 8 and 9) where correct answer is either inconclusive or combination of true and inconclusive.

## Acknowledgments

I thank a member that provided guidance and feedback relevant to class project briefly during later part of the Spring 2025 semester. My acknowledgment for their guidance. Additional acknowledgments for using chatbot icon/wumpus/-bats and using the free-tier Conversational AI access for three LLMs: https://tinyurl.com/inconclusive-sub2

## References

Akama, K.; and Nantajeewarawat, E. 2021. Skolemization that preserves logical meanings. *International Journal of Innovative Computing, Information and Control*, 17(1): 1–13.

Anwar, U.; Saparov, A.; Rando, J.; Paleka, D.; Turpin, M.; Hase, P.; Lubana, E. S.; Jenner, E.; Casper, S.; Sourbut, O.; et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Buss, S. 1998. An introduction to proof theory.

Chandu, K. R.; Li, L.; Awadalla, A.; Lu, X.; Park, J. S.; Hessel, J.; Wang, L.; and Choi, Y. 2024. Certainly Uncertain: A Benchmark and Metric for Multimodal Epistemic and Aleatoric Awareness. *arXiv preprint arXiv:2407.01942*.

Deveci, İ. E.; and Ataman, D. 2025. The Ouroboros of Benchmarking: Reasoning Evaluation in an Era of Saturation. *arXiv preprint arXiv:2511.01365*.

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:2402.00367*.

Harrison, J. 2009. *Handbook of practical logic and automated reasoning*. Cambridge University Press.

Jiang, Y.; Ilievski, F.; Ma, K.; and Sourati, Z. 2023. BRAINTEASER: Lateral thinking puzzles for large language models. *arXiv preprint arXiv:2310.05057*.

Kirichenko, P.; Ibrahim, M.; Chaudhuri, K.; and Bell, S. J. 2025. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. *arXiv preprint arXiv:2506.09038*.

MacDiarmid, M.; Wright, B.; Uesato, J.; Benton, J.; Kutasov, J.; Price, S.; Bouscal, N.; Bowman, S.; Bricken, T.; Cloud, A.; et al. 2025. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397*.

Saadat, A.; Sogir, T. B.; Chowdhury, M. T. A.; and Aziz, S. 2024. When Not to Answer: Evaluating Prompts on GPT Models for Effective Abstention in Unanswerable Math Word Problems. *arXiv preprint arXiv:2410.13029*.

Saparov, A.; and He, H. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Saparov, A.; Pang, R. Y.; Padmakumar, V.; Joshi, N.; Kazemi, M.; Kim, N.; and He, H. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36: 3083–3105.

Saparov, A.; Pawar, S.; Pimpalgaonkar, S.; Joshi, N.; Pang, R. Y.; Padmakumar, V.; Kazemi, S. M.; Kim, N.; and He, H. 2024. Transformers Struggle to Learn to Search. *arXiv preprint arXiv:2412.04703*.

Shukor, M.; Rame, A.; Dancette, C.; and Cord, M. 2024. Beyond Task Performance: Evaluating and Reducing the Flaws of Large Multimodal Models with In-Context Learning. arXiv:2310.00647.

Smolensky, P.; Fernandez, R.; Zhou, Z.; Opper, M.; Davies, A.; and Gao, J. 2025. Mechanisms of symbol processing for in-context learning in transformer networks. *Journal of Artificial Intelligence Research*, 84.

Tan, D.; Woodruff, A.; Warncke, N.; Jose, A.; Riché, M.; Africa, D. D.; and Taylor, M. 2025. Inoculation Prompting: Eliciting traits from LLMs during training can suppress them at test-time. *arXiv preprint arXiv:2510.04340*.

Tomani, C.; Chaudhuri, K.; Evtimov, I.; Cremers, D.; and Ibrahim, M. 2024. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*.

Wen, B.; Yao, J.; Feng, S.; Xu, C.; Tsvetkov, Y.; Howe, B.; and Wang, L. L. 2024. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*.

## Reproducibility Checklist

---

**Instructions for Authors:**

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

**Example:** If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

### 1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) partial, more on github

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) Yes, to the best of my knowledge.

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) Yes, to the best of my knowledge.

### 2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) No, to the best of my knowledge.

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

### 3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) No. I am attempting to create my own dataset, potentially a benchmark and have explored and had to reconduct the experiments since LLMs versions changed leading to different variations. Therefore I re- designed a new method that adapts my prior work and attempted to enhance and created a required ground- work for the same. My dataset does not follow a the existing dataset format or line of idea. My dataset is at- tempting to address inconclusivess as a subset of ab- stention that is likely answered using lateral, logical and with or without commonsense reasoning.

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) Type your response here

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) Type your response here

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) Type your response here

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) Type your response here

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) Type your response here

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) Type your response here

### 4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) Answer: No. I also do not have access to computational resources required to finetune and also no access to exhaustive use of all three LLMs. Developing a proof-of-concept methodology by using over-the-counter access to LLMs features would restrict access to configuration/hyper parameter settings esp. with recent versions of LLMs. Therefore my experiments are manual, as in served as prompts to the open access APIs under free-tier. I focused on developing methodology in an attempt to create a method for generating a dataset that may challenge LLMs a little further into the future and distancing a potential benchmark saturation, based on Deveci and Ataman (2025), Betley et al.(2025) Mac-

Diarmid et al. (2025) and Tan et al. (2025).

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Type your response here

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) Type your response here

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) Type your response here

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Type your response here

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Type your response here

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) Type your response here

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Type your response here

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Type your response here

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) Type your response here

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) Type your response here

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) Type your response here

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Type your response here

## Ethics Statement

I have written this paper by myself without using any AI tools. This is my own work, I confirm this does not include anyone else's ideas. I worked on this project as part of a class project but without much LLM-specific guidance. The credits of this earlier work can be revealed if it reaches such a stage. However, the current work included here is significantly different from class project and uses different methods due to newer versions of LLMs that respond differently and recent research works led to change in how this paper is written. Please contact me in case you need more details.

## Appendix

### Using Resolution Refutation

Resolution is refutation complete, if a set of clauses, that are in first-order skolemized clausal normal forms, are unsatisfiable, since the resolution derives empty clauses and verifies unsatisfiability. But resolution is not decidable if the set of clauses may be satisfiable, i.e. if there is a query that is a logical consequence of the set of clauses, then the Resolution Refutation (RR) procedure may never terminate. If clause C1 is a first-order clausal normal form that is unsatisfiable, then C $\vdash$ since resolution combined with skolemization and classification is a refutation complete calculus for First Order Logic Harrison (2009). By lifting lemma and herbrand's theorem, proof for first-order resolution's refutation completeness can be proved. If a formula is satisfiable, proof search can terminate or go on forever, as there are infinitely many clauses that can be generated by resolution. Therefore, resolution is not a decision procedure for First-order logic and is undedicable if the program never terminated so far the query is satisfied. There may be other forms of resolution Buss (1998) but in this work, only conventional skolemization decomposition is used for this specific knowledge base. The conventional skolemization decomposition are not meaning-preserving clausal forms, over first-order logical clauses using existential quantifiers Akama and Nantajeewarawat (2021) (refer example in Akama and Nantajeewarawat (2021)). Traditional Resolution Refutation does not have world knowledge and do not consider word sense disambiguation especially for words of same part-of-speech tags since skolemized clausal forms are not meaning preserving Akama and Nantajeewarawat (2021). I included implicit equivalence that preserves meanings between two key features: as "...Wings are not barrels. Barrels are not wings...".

### Prompts

Due to changes in release versions of LLMs, the prompts tested on earlier version of the chosen LLMs GPT-x, Claude-x are different, there are two sections. First section Claude 4x, Gemini 3x, GPT 5x are latest. Older version of prompts within same LLM families: Claude 3x, GPT 4x are in next section that were supplied with different prompts.

## Prompts that were run on Claude Sonnet 4.5, Gemini 3.0 Pro, GPT 5.1

- **Prompt 1 with and without Resolution Refutation:** KB: Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Query: The brown bat can fly.

  1. Generate First Order Logical form over knowledge base and query. 2. Then convert to conjunctive normal forms. 3. Then convert to skolemized clausal forms by using skolemization, and give entire knowledge base in skolemized clausal normal forms including the query and negated query and then prove by resolution refutation if the query is true or false or inconclusive. 4. Without using resolution refutation, without first order logical forms and without clausal forms, answer using KB alone if the query is true or false or inconclusive.

- **Prompt 2 without Resolution Refutation:** Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Can the brown bat fly?

- **Prompt 3 with Resolution Refutation:** Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Query: The brown bat can fly. Using resolution refutation, prove if the query is true or false or inconclusive.

- **Prompt 4 with two images of brown bat and brown baseball bat:** I included two images: one for a mammal that is a brown bat and another for a brown baseball bat. Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. By looking at the KB and two images, answer if the statement "Brown Bat can fly" true or not or inconclusive?

- **Prompt 5 with multiple choices involving common sense answer, pure logical answer, combined common sense and logical answer:** Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. The brown bat flew. For the statement: Brown bat is a mammal, select the Correct Ans from following answers:

  1. True

  2. False

  3. Inconclusive

  4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player.

- **Prompt 6 with multiple choices including confusing choices:** Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. The brown bat flew. For the statement: Brown bat is a mammal, select the correct answer from following answers: 1. True 2. False 3. Inconclusive 4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player 5.Answer 4 and other answers with other reasoning 6.Combination of answer under specific assumptions not included in aforementioned answers

- **Prompt 7 without world knowledge, using resolution refutation:** Wumpus is a mammal. Wumpus is used in baseball. Wumpuses have a barrel. Wumpus flies with wings. Wings are not barrels. Barrels are not wings. Wumpus can fly. There is a brown Wumpus. Query: The brown Wumpus can fly. Using resolution refutation, prove if query is true or false or inconclusive ?

- **Prompt 8 without world knowledge, using resolution refutation with syllogisms:** Some Wumpuses are mammals. Some Wumpuses are used in baseball. Some Wumpuses have barrels and are used in baseball. Wumpus that are mammals fly with wings. Wings are not barrels. Barrels are not wings. Some Wumpuses can fly. There is a brown Wumpus. Query: the brown Wumpus can fly. Use resolution refutation to answer from one of the following : 1. True 2. False 3. Inconclusive.

- **Prompt 9 without RR, without any world knowledge references, with syllogisms:** Some Wumpuses are mammals. Some Wumpuses are used in baseball. Some Wumpuses have barrels and are used in baseball. Wumpus that are mammals fly with wings. Wings are not barrels. Barrels are not wings. Some Wumpuses can fly. There is a brown Wumpus. Can the brown Wumpus can fly ? Answer from one of the following : 1. True 2. False 3. Inconclusive.

## Prompts that were run on Claude 3.4, Claude 3.7, GPT 4o with "Thinking"

- **With Resolution Refutation (RR) Prompt 1 using RR:** Knowledge base is as follows: Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Query: The brown bat can fly Steps: Generate First Order Logical form over knowledge base and query. Then convert to conjunctive normal forms. Then convert to skolemized clausal forms by using skolemization, and give entire knowledge in skolemized clausal normal forms including the query and negated query and then prove by resolution refutation if this statement is true or not or inconclusive?

- **Without Resolution Refutation (RR): Prompt 3**

**without RR:** This is the knowledge base: Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Is the statement "Brown Bat can fly" true or false or inconclusive?

## Results

Each prompt numbered through 1 to 9 listed in Prompts section of Appendix. For each LLM, same prompt was provided five times and in new session. Due to free-tier and recent versions of LLMs (Nov 2025 release dates), default configuration were used and prompts were manually supplied to each respective Conversational AI chat applications. Only those prompts that were consistent, non-random, produced same reasoning traces, without much deviation in number of tokens generated were marked as final and included. There were several non-trivial prompts that were generated surrounding same KB for lateral and logical reasoning but are not included due to various factors: to maintain the context, consistency and requirement. Each prompt was repeated five times.

**Claude 4.5, Gemini 3.0 Pro and GPT 5.1** Tables 1 1, Table 2 2 and Table 3 3.

| Prompt ID | Claude | Correct Ans |
|---|---|---|
| 1 (+-RR) | T and T | T and Incl |
| 2 (no RR) | T | Incl |
| 3 (RR) | T (3), Incl(2) | T |
| 4 (images + no RR) | T | Incl |
| 5 (MCQ) | T(1),Incon(1), both(3) | both |
| 6 (MCQ) | T(5) | both |
| 7 (RR & no WK) | T | True |
| 8 (RR & no WK) | T | Incl |
| 9 (no RR & no WK) | T | Incl |

Table 1: Results for Claude Sonnet 4.5, RR: Resolution refutation, and True, False, Inconclusive, both (refers to True if this and Inconclusive if that) . No WK ref - No World knowledge reference

| Prompt ID | Gemini 3 | Correct Ans |
|---|---|---|
| 1 (+-RR) | T, T | T, Incl |
| 2 (without RR) | T | Incl |
| 3 (with RR) | T | T |
| 4 (images + no RR) | T | Incl |
| 5 (MCQ) | T(5) | both |
| 6 (MCQ) | T(5) | both |
| 7 (RR & no WK) | T | T |
| 8 (RR & no WK) | T | Incl |
| 9 (no RR & no WK) | T | Incl |

Table 2: Results for Gemini 3.0 Pro, RR: Resolution refutation, and True, False, Inconclusive . No WK ref - No World knowledge reference .

| Prompt ID | GPT 5.1 | Correct Ans |
|---|---|---|
| 1 (+- RR) | T and T | T and Incl |
| 2 (no RR) | T | Incl |
| 3 ( RR) | T | True |
| 4 (images + no RR) | T | Incl |
| 5 (MCQ) | T | both |
| 6 (MCQ) | T | both |
| 7 (RR & no WK) | T | True |
| 8 (RR & no WK) | T | Incl |
| 9 (no RR & no WK) | T | Incl |

Table 3: Results for GPT 5.1, with Resolution Refutation (RR), and True (T), False (F), Inconclusive (Incl), MCQ (multiple choice question). No WK ref - No World knowledge reference.

**Claude 3.4, Claude 3.7, GPT 4o with "Thinking"** Tables 1 4, Table 2 5 and Table 3 6.

| Prompt ID | with RR | LLM | Correct Ans |
|---|---|---|---|
| 1 | Y | T | T |
| 3 | N | T | Incl |

Table 4: Results for Claude 3.4, RR: Resolution refutation, and True, False, Inconclusive - each prompt repeated three times in new context window

| Prompt ID | with RR | LLM | Correct Ans |
|---|---|---|---|
| 1 | Y | T | True |
| 3 | N | Incl | Incl |

Table 5: Results for Claude 3.7, RR: Resolution refutation, True, False, Inconclusive - each prompt repeated three times in new context window

## Manually Generated KBs

The polysemy nouns are: (bat, head, nail, pupil) and the colors are: (black, white, brown, red, yellow, blue, green, pink)

- Step 1: The facts are as follows for each one of polysemy nouns:
  - Bat is a mammal. Bat is used in baseball.
  - Pupils attend school. Pupils are part of the eye.
  - Nail is the tip of the finger. Nails are hammered into the wood.
  - Head is at the top of the human body. Head runs the company.
- Step 2: The "can" pairs of sentences for each meaning are as follows for each one of polysemy nouns:

| Prompt ID | with RR | LLM | Correct Ans |
|---|---|---|---|
| 1 | Y | T | T |
| 3 | N | Incl | Incl |

Table 6: Results for GPT 4o with "Thinking" , RR: Resolution refutation, True, False, Inconclusive - each prompt repeated three times in new context window

| Prompt ID | with RR | System Ans | Correct Ans |
|---|---|---|---|
| 1 | Y | SAT,terminates | True |
| 3 | N | SAT,terminates | Incl |

Table 7: Results for RR program with Clausal normal forms manually solved, RR: Resolution refutation, T-True, False-F, Inconclusive - IC, SAT

  - Bats have a barrel. Bat flies with wings.
  - Pupils study at school. Pupils can dilate in the eye.
  - Nail is the tip of the finger. Nails are hammered into the wood.
  - Head contains 22 bones. Head makes the decisions.
- Step 3 & 4: Selecting characteristics/containment entities/actions and generating equivalences are as follows for each one of polysemy nouns:
  - Wings are not barrels. Barrels are not wings.
  - Eyes are not schools. Schools are not eyes.
  - Nails protect phalanges. Phalange is not a metal.
  - Bones are not decisions. Decisions are not bones.
- Step 5: Repeated assertions of the fact from step 3/4 are as follows for each one of polysemy nouns:
  - Bats can fly.
  - Pupils can dilate.
  - Nails can fasten the objects.
  - Head can decide.
- Step 6 & 7: Repeated assertions, add an action in past tense of the fact from step 3/4 are as follows for each one of polysemy nouns:
  - There is a *brown bat*. The brown bat flew. Query: Brown bat is a mammal.
  - There is a *green pupil*. The green pupil did not dilate. Query: Green pupil is from the eye.
  - There is a *pink nail*. The Pink nail did not fasten the objects. Query: Pink nail is part of the finger.
  - There is a *blue head*. The blue head did not decide. Query: Blue head is part of the body.

### Resulting KBs are as follows:

- Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. *Bats can fly.* There is a *brown bat*. Brown bat flew. Query: Brown bat is a mammal.
- Pupils attend school. Pupils are part of the eye. Pupils study at school. Pupils can dilate in the eye. Eyes are not schools. Schools are not eyes. *Pupils can dilate.* There is a *green pupil*. The green pupil did not dilate. Query: Green pupil is from the eye.
- Nail is the tip of the finger. Nails are hammered into the wood. Nails are made of metal to fasten objects. Nails protect phalanges. Phalange is not a metal. Metal is not a Phalange. *Nails can fasten the objects.* There is a *pink nail*. The Pink nail did not fasten the objects. Query: Pink nail is part of the finger
- Head is at the top of the human body. Head runs the company. Head contains 22 bones. Head makes the decisions. Bones are not decisions. Decisions are not bones. *Head can decide*. There is a *blue head*. The blue head did not decide. Query: Blue head is part of the body.

### Prompts updated March 2026

**Example prompt** —
https://github.com/sushmaanandakoju/logically-inconclusive-kbs?tab=readme-ov-file#prompt

**Example data generation script** —
https://github.com/sushmaanandakoju/logically-inconclusive-kbs

1. Select a polysemy noun. Each one of the polysemy noun keywords in - bat, head, nail, paper- that have two different meanings in the noun forms and with real world references. Example: bat. Add this to json key: "KB".

2. Select one fact each for each one of the two meanings for the each one of the keywords. Example: "Bat is a mammal. Bat is used in baseball." where this information is valid in real world. Add this to json key: "KB".

3. Select two distinct features of two different meanings of a given keyword and generate two new facts using "have" or "contain" depending on reference. Example: "Bats have a barrel. Bat flies with wings". Add this to json key: "KB".

4. Generate two equivalence clauses for two distinct features selected in previous step, for the two meanings of a given keyword. **(for ensuring meaning preservation by common sense, not refutation)**. Example: "Wings are not barrels. Barrels are not wings.". Add this to json key: "KB".

5. Introduce a repeated re-assertion of a fact generated in step 3 with "can" or other relevant forms. Add this to json key: "KB".

6. Select a random color. Prepend the color to new instance of keyword and generate a query us-

ing this new colored instance to construct an introduction. For the query, use one of the two facts to replace the keyword with colored keyword instance *color* $\in$ *black, white, brown, red, yellow, blue, green, pink*. Example: "There is a Brown bat" or "There is a pink nail" and so on. Add this to json key: "KB".

7. Add an action or negated action sentence based on step 2, replacing the noun with colored keyword instance (Ex: brown bat). The idea is to add confusion such that the action/property that could distinguish two meanings of the polysemy noun is absent, leading to inconclusiveness. Examples: "There is a brown bat. Brown bat flew." or "There is a pink nail. Pink nail did not fasten the object." Add this to json key: "KB".

8. Using the facts in step 1 such as containment or $\exists$ assertions from facts, generate a query as a statement. Example: "Query: Brown bat is a mammal", or "Query: Pink nail is part of the finger". The resulting query "Is the statement "Brown bat is a mammal" true or false or inconclusive?".

9. Generate a question with multiple choices can be framed as: "For the statement: Brown bat is a mammal, select the Correct Ans from following choices: 1. True, 2. False, 3. Inconclusive, 4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player.".

10. Place results from steps 8 and 9 into a json list with key: "Inconclusive".

11. Generate reasoning why this KB is "Inconclusive".

12. Now combine the resulting KB in this format: "key": "Bat", "kb": "Bat is a mammal. Bat is used in baseball. Bats have a barrel. Bat flies with wings. Wings are not barrels. Barrels are not wings. Bats can fly. There is a brown bat. Brown bat flew." "inconclusive": ["query": "Brown bat is a mammal.", "mcq": "For the statement: Brown bat is a mammal, select the Correct Ans from following choices: 1. True, 2. False, 3. Inconclusive, 4. True if brown bat is a mammal and Inconclusive i.e. undecidable if brown bat could have accidentally thrown by a baseball player.", "Reasoning": "Brown bat can be a mammal or baseball bat when a baseball player may have thrown the brown bat."]

13. For each generated KB, generate a conclusive query, mcq, reasoning format and add a "conclusive" list with query, mcq, reasoning items and add it to the array.

14. Repeat all the steps for remaining polysemy noun keywords listed in step 1.

15. Generate other polysemy noun keywords. And repeat steps 1 to 10 for each new polysemy noun.

16. Place all generated results in a json array format.

## Responses from Three Models

### Claude Sonnet 4.5

https://anonymous.4open.science/r/proof-of-concept-method-for-inconclusiveness-17B3/three-llms-results.md

### Gemini 3.0 Pro

https://anonymous.4open.science/r/proof-of-concept-method-for-inconclusiveness-17B3/three-llms-results.md

### GPT 5.1

https://anonymous.4open.science/r/proof-of-concept-method-for-inconclusiveness-17B3/three-llms-results.md

## Preliminary results based on Claude-3x Gpt-4x model families

Based on the preliminary results from using only resolution refutation method-based prompts for prompts, results vary over Claude 3.7. But for all other LLMs claude 3.5, GPT-4o with Think before responding - the results are consistent though remain consistent with commonsense avoiding any form of acknowledgment to disambiguation in meaning. The common aspect from consistent responses from three different LLMs from three different model families are:

- The generated clausal forms and first order logic forms, differentiate bat as a mammal and bat in baseball and proactively add constants and terms accordingly from non-skolemized forms. The LLMs can vary and represent FOL clauses and signatures in FOL forms which can render different form of verification using sympy.

- I implemented two scripts for each one of the two FOL forms and skolemized clausal forms and applied resolution refutation implemented using Sympy derived symbol and clause grounding.

- LLMs verified on this single prompt used similar substitution methods and clauses to that of implemented algorithmic software solution on the same clausal forms without running into undecidable paths or without running into haling problem.

- LLMs ground early on to the world knowledge and common sense reasoning for grounding the constants and predicates within context for this specific question. This is particularly observed by Claude-3x family of models.

- LLMs align "can fly" synonymous to the ability to fly and ground it to mammal, as a subset of animal category and are able to provide more details of a brown bat as a species.

**For earlier version of prompts over previous versions of LLM families:** In the earlier design of this problem, where the designed KB, query prompts were evaluated over three LLMs families, where LLMs generated clausal normal forms and verified by using Resolution Refutation proof. I supplied the resulting logical forms to a separate implementation to formally verify if generating clausal forms really work. For the two prompts, I have created a ground truth with manually solved solutions. KBs were designed such that query is satisfied. I used Claude 3.5, 3.7, GPT 4o with Thinking. For each LLM, I gave each one of the prompts listed in Appendix section here. For randomness of the results, I repeated the process thrice for each one of the same prompts within same ICL. The six prompts and the results for each LLM and the two KBs are provided in the Appendix sec-

tion. https://anonymous.4open.science/r/proof-of-concept-method-for-inconclusiveness-17B3/four-kb-examples.md.

## Symbol Processing

Recent research studies Smolensky et al. (2025) examines pure symbol processing with cognitive science configurations in an In-Context Learning (ICL) for understanding "..capabilities of transformer networks to perform pure symbol manipulation tasks in which symbol meanings are irrelevant or non-existent". The authors design a "swap" task and utilize various definitions and studies for ICL to evaluate over various LLMs such as GPT-4o, even more recent nano GPT, LLaMa, and others. This work elaborately studied how transformers perform symbol manipulation in ICL. The swap design encodes english sentences with symbols and patterns such as "program, translation and compiler" format (for Turing-complete evaluations over Query, Key, Value Machine (QKVM)) by using deterministic embeddings and other inherent configurations) and such patterns maybe represented by special programs with Production System Language that are mechanistically interpretable. This works suggests more improvements and future works. The pattern of symbols provided is part of QKVM system. This work provides a detailed and comprehensive analysis about potential improvements for symbols processing in transformer models by bringing comprehensive cognitive science perspectives and configurations.