ShZZaM: An LLM+ATP Natural Language to Logic Translator

First1 Last1, First2 Last2
Affiliation
City, Country

Abstract

This paper describes the ShZZaM tool that uses Large Language Models (LLMs) and Automated Theorem Proving (ATP) tools to translate natural language to typed first-order logic in the TFF syntax of the TPTP World. MORE

Large Language Models (LLMs) (Peykani et al. 2025) have shown themselves to be useful in a broad range of applications (Sajjadi Mohammadabadi et al. 2025). However, it is well known that LLMs make mistakes (Huang et al. 2025), and this is acknowledged on LLMs' web interfaces, e.g., ChatGPT admits "ChatGPT can make mistakes. Check important info". In the face of such unreliability, the results from LLMs in mission-critical applications require verification. One approach is to translate the LLM input and output to a logical form that can be checked using Automated Theorem Proving (ATP) tools, e.g., (Yang et al. 2025; Cheng et al. 2025).[1] A key step in this verification pipeline is the faithful translation of the natural language to an appropriate logical form. This task is difficult due to the ambiguous nature of natural language statements, especially informally expressed statements. Work in this area includes LINC (Olausson et al. 2023), FOLIO (Han et al. 2024), and LINA (Li et al. 2024). This paper makes another contribution in this area, taking a new interactive approach to the translation process, zigzagging (hence the 'ZZ' in the tool name) between natural language and logic until convergence is achieved. A key feature of ShZZaM is its use of LLMs and Automated Theorem Proving (ATP) tools, which complement each other in the translation steps.

Figure 1 shows the overall process implemented of ShZZaM. Starting with the natural language, a combination of LLMs and ATP tools make a first translation (step 1 - a "Zig") to the typed first-order logic in the TFF syntax (Sutcliffe et al. 2012; Blanchette and Paskevich 2013) of the TPTP World (Sutcliffe 2024). An LLM is then used to translate the logic back to natural language (step 2 - a "Zag"). An LLM is then used to judge (step 3) whether or not the new and previous natural language statements have the same meaning, at a given level of required similarity. If they do, the logic is accepted as the translation. If they do not, another zigzag is performed, continuing until the natural language pairs converge to the required level of similarity (or a limit is reached).

LLMs know TFF - translation works. LLMs know language - similarity works.

Automated Theorem Proving (ATP) deals with the task of proving theorems from axioms – the derivation of conclusions that follow inevitably from known facts (Robinson and Voronkov 2001). The converse task of disproving conjectures is another facet of interest (Claessen and Sörensson 2003; Blanchette and Nipkow 2010). The axioms and conjectures are written in an appropriately expressive logic, and the solutions (proofs and models) are often similarly written in logic (Sutcliffe 2023). The TPTP World (Sutcliffe 2017) is the well established infrastructure that supports research, development, and deployment of ATP systems. The TPTP World includes the TPTP problem library (Sutcliffe 2017), the TSTP solution library (Sutcliffe 2010), standards for writing ATP problems and reporting ATP solutions (Sutcliffe et al. 2006; Sutcliffe 2008), tools and services for processing ATP problems and solutions (Sutcliffe 2010), and it supports the CADE ATP System Competition (CASC) (Sutcliffe 2016). The web page `tptp.org` provides access to all components.

The TPTP language (Sutcliffe 2023) is one of the keys to the success of the TPTP World. The TPTP language is used for writing both problems and solutions, which enables convenient communication between ATP systems and tools. Problems and solutions are built from *annotated formulae* of the form:

language(*name*, *role*, *formula*, *source*, *useful_info*)

The *language*s supported are `cnf` (clause normal form), `fof` (first-order form), `tff` (typed first-order form), and `thf` (typed higher-order form). The *role*, e.g., `axiom`, `lemma`, `conjecture`, defines the use of the formula. In a *formula*, terms and atoms follow Prolog conventions – functions and predicates start with a lowercase letter and variables start with an uppercase letter. The language also supports interpreted symbols, e.g., the truth constants `$true` and `$false`. The main logical connectives in the TPTP language are `!`, `?`,

[1] For a more comprehensive survey, just ask your favourite LLM to "show some research on how LLMs make mistakes, and the need for symbolic checking of LLM output".

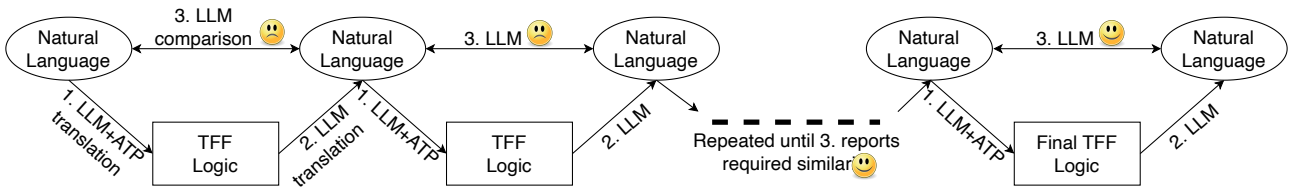

Figure 1: ShZZaM process

~, |, &, =>, <=, <=>, and <~>, for the mathematical connectives $\forall$, $\exists$, $\neg$, $\vee$, $\wedge$, $\Rightarrow$, $\Leftarrow$, $\Leftrightarrow$, and $\oplus$ respectively. Equality and inequality are expressed as the infix operators = and !=. The *source* and *useful_info* are optional.

1 Example Solutions and their Visualizations

2 Conclusion

This paper describes the derivation and interpretation viewers in the TPTP World: the Interactive Derivation Viewer (IDV), the Interactive Tableau Viewer (ITV), the Interactive Interpretation Viewer (IIV), and the Interactive Kripke Viewer (IKV). Users and developers of ATP systems are able to examine their ATP solutions in an interactive graphical environment, providing insights into features of the solutions. The viewers are freely accessible through SystemOnTPTP.

References

Blanchette, J., and Nipkow, T. 2010. Nitpick: A Counterexample Generator for Higher-Order Logic Based on a Relational Model Finder. In Kaufmann, M., and Paulson, L., eds., *Proceedings of the 1st International Conference on Interactive Theorem Proving*, number 6172 in Lecture Notes in Computer Science, 131–146. Springer-Verlag.

Blanchette, J., and Paskevich, A. 2013. TFF1: The TPTP Typed First-order Form with Rank-1 Polymorphism. In Bonacina, M., ed., *Proceedings of the 24th International Conference on Automated Deduction*, number 7898 in Lecture Notes in Artificial Intelligence, 414–420. Springer-Verlag.

Cheng, F.; Li, H.; Liu, F.; Van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. In Kwok, J., ed., *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 10400–10408. AAAI Press.

Claessen, K., and Sörensson, N. 2003. New Techniques that Improve MACE-style Finite Model Finding. In Baumgartner, P., and Fermueller, C., eds., *Proceedings of the CADE-19 Workshop: Model Computation - Principles, Algorithms, Applications*.

Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; Zhou, W.; Coady, J.; Peng, D.; Qiao, Y.; Benson, L.; Sun, L.; Wardle-Solano, A. Szabó, H.; Zubova, E.; Burtell, M.; Fan, J.; Liu, Y.; Wong, B.; Sailor, M.; Ni, A.; Nan, L.; Kasai, J.; Yu, T.; Zhang, R.; Fabbri, A.; Kryscinski, W.; Yavuz, S.; Liu, Y.; Lin, X.; Joty, S.; Zhou, Y.; Xiong, C.; Ying, R.; Cohan, A.; and Radev, D. 2024. FOLIO: Natural Language Reasoning with First-Order Logic. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22017–22031. Association for Computational Linguistics.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 42:1–55.

Li, Q.; Li, J.; Liu, T.; Zeng, Y.; Cheng, M.; Huang, W.; and Liu, Q. 2024. Leveraging LLMs for Hypothetical Deduction in Logical Inference: A Neuro-Symbolic Approach.

Olausson, T.; Gu, A.; Lipkin, B.; Zhang, C.; Solar-Lezama, A.; Tenenbaum, J.; and Levy, R. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Languag Models with First-Order Logic Provers. In Pino, J., and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5153–5176. Association for Computational Linguistics.

Peykani, P.; Ramezanlou, F.; Tanasescu, C.; and Ghanidel, S. 2025. Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions. *Applied Sciences* 15(14):8103.

Robinson, A., and Voronkov, A. 2001. *Handbook of Automated Reasoning*. Elsevier Science.

Sajjadi Mohammadabadi, S.; Kara, B.; Eyupoglu, C.; Uzay, C.; Tosun, M.; and Karakuş, O. 2025. A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications. *Electronics* 14(18):3580.

Sutcliffe, G.; Schulz, S.; Claessen, K.; and Van Gelder, A. 2006. Using the TPTP Language for Writing Derivations and Finite Interpretations. In Furbach, U., and Shankar, N., eds., *Proceedings of the 3rd International Joint Conference on Automated Reasoning*, number 4130 in Lecture Notes in Artificial Intelligence, 67–81. Springer.

Sutcliffe, G.; Schulz, S.; Claessen, K.; and Baumgartner, P. 2012. The TPTP Typed First-order Form with Arithmetic. In Bjørner, N., and Voronkov, A., eds., *Proceedings of the 18th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, number 7180 in Lecture Notes in Artificial Intelligence, 406–419. Springer-Verlag.

Sutcliffe, G. 2008. The SZS Ontologies for Automated Reasoning Software. In Sutcliffe, G.; Rudnicki, P.; Schmidt, R.;

Konev, B.; and Schulz, S., eds., *Proceedings of the LPAR Workshops: Knowledge Exchange: Automated Provers and Proof Assistants, and the 7th International Workshop on the Implementation of Logics*, number 418 in CEUR Workshop Proceedings, 38–49.

Sutcliffe, G. 2010. The TPTP World - Infrastructure for Automated Reasoning. In Clarke, E., and Voronkov, A., eds., *Proceedings of the 16th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, number 6355 in Lecture Notes in Artificial Intelligence, 1–12. Springer-Verlag.

Sutcliffe, G. 2016. The CADE ATP System Competition - CASC. *AI Magazine* 37(2):99–101.

Sutcliffe, G. 2017. The TPTP Problem Library and Associated Infrastructure. From CNF to TH0, TPTP v6.4.0. *Journal of Automated Reasoning* 59(4):483–502.

Sutcliffe, G. 2023. The Logic Languages of the TPTP World. *Logic Journal of the IGPL* 31(6):1153–1169.

Sutcliffe, G. 2024. Stepping Stones in the TPTP World. In Benzmüller, C.; Heule, M.; and Schmidt, R., eds., *Proceedings of the 12th International Joint Conference on Automated Reasoning*, number 14739 in Lecture Notes in Artificial Intelligence, 30–50.

Yang, X.-W.; J-J., S.; Guo, L.-Z.; Zhang, B.-W.; Zhou, Z.; Jia, L.-H.; Dai, W.-Z.; and Li, Y.-F. 2025. Neuro-Symbolic Artificial Intelligence: Towards Improving the Reasoning Abilities of Large Language Models. In Kwok, J., ed., *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 10770–10778. AAAI Press.