# Experiments on the Use of LLMs for the Translation of the Babylonian Talmud

Mafalda Papini[1], Davide Albanesi[1], David Dattilo[2], Emiliano Giovannetti[1], Simone Marchi[1]
[1] Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Italy - name.surname@ilc.cnr.it
[2] Progetto Traduzione Talmud Babilonese S.c.a r.l., Italy - david.dattilo@talmud.it

## ABSTRACT (ENGLISH)
In this paper, we present an experiment on the use of Large Language Models (LLMs) in the translation of the Babylonian Talmud into Italian. The experiment focuses on translation and demonstrates how the combined use of LLMs and Translation Memory can improve the quality of suggested translations in a Computer-Assisted Translation context. The initial results highlight both the positive contribution of this hybrid technique and the limitations posed by the nature of complex texts like the Babylonian Talmud, whose translation can only be interpretative.
**Keywords:** Babylonian Talmud; Computer-Assisted Translation; Large Language Models; Artificial Intelligence

## ABSTRACT (ITALIANO)
*Esperimenti sull'uso dei LLM per la traduzione del Talmud babilonese.* In questo articolo, presentiamo un esperimento sull'uso dei modelli di linguaggio di grandi dimensioni (Large Language Models) nella traduzione del Talmud babilonese in italiano. L'esperimento riguarda la traduzione e dimostra come l'uso combinato di LLM e Translation Memory possa migliorare la qualità delle traduzioni suggerite in un contesto di traduzione assistita dal computer. I primi risultati evidenziano sia il contributo positivo di questa tecnica ibrida, sia i limiti posti dalla natura di testi complessi come il Talmud babilonese, la cui traduzione non può essere che interpretativa.
**Parole chiave:** Talmud Babilonese; Traduzione Assistita dal Computer; Modelli di Linguaggio di Grandi Dimensioni; Intelligenza Artificiale

## 1. INTRODUCTION
The use of Large Language Models (LLMs) has become pervasive. The fields where such models are being applied are constantly expanding, driven by the curiosity to test their effectiveness on specific datasets and the undeniable evidence of their growing reliability. To evaluate the performance of the various models currently available, specific tasks have been defined to test their linguistic, cognitive, and multimodal capabilities[1]. Some of these tasks are central to research focusing on the linguistic competence of the models and their knowledge of the world, such as GPQA (Graduate-Level Google-Proof Q&A) (Rein et al., 2023) and MMLU-Pro (Massive Multitask Language Understanding) (Wang et al., 2024a). These particular tasks offer insights into the applications of the models' access to knowledge (both factual and linguistic) and the performance that can currently be achieved. It is therefore unsurprising that LLMs have been applied to the task discussed in this article, namely computer-assisted translation (CAT), due to their ability to understand and generate natural language.
While the use of LLMs in Machine Translation is not surprising (for a review, see Wang et al., 2024b), CAT supported by LLMs is another hot research frontier. The synergistic relationship between LLMs and Translation Memory (TM), the core of CAT systems, is central to the work presented in (Mu et al., 2023). The results show that the performance of an LLM trained for translation can be significantly enhanced by integrating TM-derived data into prompts. Although there is literature on the application of machine translation to religious texts (Liebeskind et al., 2024; Zaid & Bennoudi, 2023; Beal, 2022), to the best of our knowledge, no experiments have yet been conducted on the integration of TMs and LLMs in computer-assisted translation of religious texts.
This contribution illustrates an application of LLMs in the support of translating the Babylonian Talmud, carried out within the framework of the Babylonian Talmud Translation Project into Italian. As partly emerges from the experiment described later, undertaking the translation of the Talmud presents significant challenges due to the text's intrinsic complexity and the diverse range of topics it covers. The Talmud is primarily written in Biblical Hebrew and Babylonian Aramaic, using an extremely concise and

[1] For an overview of the main tasks, see the page: https://github.com/huggingface/leaderboards/blob/main/docs/source/en/open_llm_leaderboard/about.md (cons. 24/01/2025).

often cryptic style, rendering a literal translation incomprehensible. Additional words have been included in the translation to enhance readability. The text is structured into thematic "blocks" and further divided into "logical units" (e.g., objections, questions). Additional notes cover Halakhah, Nature, Linguistics, and brief biographies. Each volume also includes appendices containing glossaries of various nature.

## 2. METHODOLOGY

The translation of the Babylonian Talmud into Italian, carried out as part of the project of the same name since 2012, is supported at every stage by the Traduco software (Giovannetti et al., 2016), a CAT tool specifically developed for the project. Like any CAT tool, Traduco uses a TM as a database to suggest translations.

A TM is essentially a database containing sentence pairs that automatically records translated segments alongside their original source texts during translation. The primary aim of a TM is to reuse previous translations, thus enhancing the speed and consistency of the translation. TMs are particularly effective when working with texts characterized by a significant number of repeated or partially repeated segments (such as standardized expressions), a feature prominently found in highly formulaic and repetitive texts like the Babylonian Talmud.

After 12 years of translations, the TM contains 291,708 pairs of translated segments and a total of 1,633,544 segment pairs if historical versions are also considered. Thanks to this vast amount of available data, the system is able to suggest an exact translation for 35.94% of the segments, classified within the system as a "five-star" translation (net of any adjustments required by the context). The basic idea for this experiment arose precisely from the large amount of data available in the TM and the possibility of combining these data with the translation capabilities of language models.

Specifically, the following strategy was adopted: if a new segment to be translated appears in the TM with at least three suggestions having a similarity score greater than or equal to 50% (Bellandi et al., 2016), the support of an LLM is utilized.

The adopted methodology was inspired by two well-known techniques: Retrieval Augmented Generation, or RAG (Lewis et al., 2020), and few-shot prompting (Brown et al., 2020).

The idea is to integrate the prompt with a set of segments suggested by the TM for the new segment to be translated, providing the model with partial examples of translations[2]. This approach ensures that the model is not left entirely free to translate solely based on its internal knowledge but is guided by examples to preserve, as much as possible, the translational style already adopted, particularly in the choice of lexicon.

Normally, in a RAG system, semantic similarity measures are used to identify contexts. In this case, as well, the suggestions provided by the TM take this criterion into account, as the Translation Memory System already incorporates an embedding model that adjusts the similarity measure used to rank suggestions based on the semantic distance between words. For more details, see (Bellandi et al., 2016).

Below is the prompt template prepared for this experiment, where {*target_segment*} represents the segment to be translated:

> *You are an expert in translations from Hebrew-Aramaic to Italian.*
> *In the CONTEXT you will find examples of translations in the form of pairs*
> *where SOURCE is the Hebrew-Aramaic text and TARGET is the Italian text.*
> *CONTEXT:*
> *SOURCE="Hebrew segment",TARGET="Italian segment"*
> *...*
> *SOURCE="hebrew segment",TARGET="italian segment"*
> *Write only the translation of {target_segment} without writing anything else.*

To better illustrate the methodology adopted, we first present a "running example". The textual segment we are about to translate is "כֵּיוָן דַּאֲמַר לָהּ חֲצִי פְּרוּטָה פְּסָקָה", found in the first chapter of the tractate Qiddushin. For this segment, we have the reference Italian translation obtained after the translation process: "*Dal momento che le ha detto: 'Mezza perutà' ha interrotto l'azione*".

[2] In the context of the Talmud Translation Project, the experts adopted a customised text segmentation in which each string can range from a single word to short sentences. This choice responds to the peculiar and repetitive nature of the text and maximises the effectiveness of TM usage.

The system's TM is unable to provide any five-star suggestions for this segment. In other words, this sentence has not yet been translated in the exact form in which it appears. However, for this segment, the TM contains five suggestions with similarity scores ranging from 68% to 50%. The first of these suggestions provides the following translation: "*Dal momento che le dice: 'Domani' ha interrotto l'azione*". Before testing the application of suggestions on the LLM (we used Llama 3.1 70b), the language model was first asked to translate the segment directly without any assistance, resulting in the following translation: "*Poiché disse a lei mezza peruta, la interruppe*".
At this point, the prompt template described above was instantiated, and the model was queried again, this time providing the five TM suggestions as "context" input. The resulting translation was: "*Dal momento che le dice: 'Mezza perutà' ha interrotto l'azione*".
At a glance, the resulting translation is more accurate than the one obtained by querying the LLM directly. Even the word "פְּרוּטָה" was correctly translated with its accented form, "*perutà*", following the contextual examples provided and in accordance with the project's editorial standards.
Methodologically, the process began by extracting a random sample of 300 segment pairs (source-target) from the TM among those fulfilling the previously introduced eligibility requirements (i.e., the presence of at least 3 suggestions with a similarity score of at least 50%). For each segment pair, the "source" was translated both by directly invoking the LLM and by using the above prompt supplemented with translation suggestions. The "target" part was set aside as the reference segment. Subsequently, to evaluate the quality of the translations produced, three different measures used in Machine Translation were applied: SacreBLEU (Post, 2018), BERTscore (Zhang et al., 2020), and Meteor (Lavie & Agarwal, 2007). Given the particular nature of the text and its translation, a multidimensional evaluation approach was deemed preferable. This is because SacreBLEU focuses on lexical adherence, BERTscore analyzes deep semantic similarity, and METEOR accounts for synonyms and morphological flexibility. The results of the initial tests have been documented in two separate reports, one describing two examples extensively[3] and another more concisely compiling all the translations performed[4].
An analysis of the data obtained (see Table 1) revealed that for 237 of the segments considered (79% of the total), the combined use of LLM+TM resulted in a performance improvement across all three metrics compared to using the LLM alone.

| | SacreBLEU | BERTscore | Meteor |
|---|---|---|---|
| LLM | 11.0939 | 0.7283 | 0.2097 |
| LLM+TM | 35.7756 | 0.8273 | 0.5082 |
| %increase | 222.48% | 13.60% | 142.37% |

**Table 1. The arithmetic means of the values calculated for the three metrics and the corresponding percentage increases between LLM+TM and LLM.**

To better evaluate the outcome of the experiment on the remaining 21% of the segments (comprising 63 pairs), a domain expert was consulted and asked to classify the translations obtained into one of the following four categories: the translation produced by LLM+TM is the best (16 cases), the translation produced by the LLM alone is the best (4 cases), both translations are acceptable (18 cases), or neither translation is acceptable (25 cases). Excluding these last two classes of results, in which the two techniques respond in agreement, the percentage of cases where LLM+TM produces better translations than LLM stands at 80%, in line with the previous result reported in Table 1.
Additionally, analyzing the remaining cases, it becomes evident that the specific nature of Talmudic text (often requiring an interpretative translation enriched with explanatory additions to improve readability) necessitates considering as reference segments translations that are significantly distant from the literal rendering of the original text. This inevitably results in very low scores on some or all of the metrics considered.
For example, the string "נְדָרִים,", which the model translates as "*voti,*" was translated in the Talmud (and considered as the reference segment in the experiment) with the phrase "*Per quanto riguarda le offerte di*

[3] https://github.com/klab-ilc-cnr/TalmudAI-AIUCD2025/blob/main/Translation.md (cons. 24/01/2025)
[4] https://github.com/klab-ilc-cnr/TalmudAI-AIUCD2025/blob/main/TranslationResults.md (cons. 24/01/2025)

*voto,"*. However, the model with the translation examples provides a better translation ("*nedarìm*"), despite the fact that the measures classify this case as one of the negative ones. This is precisely because this transliteration often appears among the suggested translation examples (see footnote 2 for more details).

## 3. CONCLUSIONS

Although the results presented in this article are preliminary, it is already possible to draw some conclusions regarding the potential contribution that LLMs can offer in supporting the translation of particularly complex texts, such as the Babylonian Talmud.

The combined adoption of LLM and Translation Memory showed an improvement in the quality of the proposed translations, demonstrating that this hybrid approach can help in preserving the translation consistency and style.

Further analyses are necessary, first of all, to investigate in a more analytical manner how much the nature of the reference translation (especially when highly enriched with explanatory additions) and the number, order and nature of the example translations provided to the model from the TM influence the quality of the suggested translations. Furthermore, the information derived from the different styles in which the translations are written will be utilized, such as the use of bold style to represent literal translations.

Among future developments, additional models will be tested, particularly to verify if good performance can also be achieved with smaller-sized models. This would enable reduced hardware investment and easier integration into existing systems, such as Traduco.

Finally, we also plan to experiment with fine-tuning approaches using the extensive collection of parallel segments produced during the translation of the Babylonian Talmud.

## ACKNOWLEDGEMENTS

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the PTTB S.c.a r.l. – Babylonian Talmud Translation Project.

## REFERENCES

Beal, T. (2022). Interface of the Deep: Design Cues for Engaging New Media and Machine Translation with Religious Scriptures. In *The Routledge Handbook of Translation and Religion* (pp. 103–120). Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781315443485-9

Bellandi, A., Benotto, G., Di Segni, G., & Giovannetti, E. (2016). Investigating the application and evaluation of distributional semantics in the translation of humanistic texts: A case study. *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, *28*, 6–11. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NLP4TM_Proceedings.pdf#page=11

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Giovannetti, E., Albanesi, D., Bellandi, A., & Benotto, G. (2016). Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities*, *32*(suppl_1), i47–i62. https://doi.org/10.1093/llc/fqw054

Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In C. Callison-Burch, P. Koehn, C. S. Fordyce, & C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228–231). Association for Computational Linguistics. https://aclanthology.org/W07-0734/

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Liebeskind, C., Liebeskind, S., & Bouhnik, D. (2024). Machine Translation for Historical Research: A Case Study of Aramaic-Ancient Hebrew Translations. *J. Comput. Cult. Herit.*, *17*(2), 20:1-20:23. https://doi.org/10.1145/3627168

Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., & Zhu, J. (2023). Augmenting Large Language Model Translators via Translation Memories. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10287–10299). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.653

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6319

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (No. arXiv:2311.12022). arXiv. https://doi.org/10.48550/arXiv.2311.12022

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., & Chen, W. (2024a). *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark* (No. arXiv:2406.01574). arXiv. https://doi.org/10.48550/arXiv.2406.01574

Wang, Y., Zhang, J., Shi, T., Deng, D., Tian, Y., & Matsumoto, T. (2024b). Recent Advances in Interactive Machine Translation With Large Language Models. *IEEE Access*, *12*, 179353–179382. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3487352

Zaid, A., & Bennoudi, H. (2023). AI vs. Human Translators: Navigating the Complex World of Religious Texts and Cultural Sensitivity. *International Journal of Linguistics, Literature and Translation*, *6*(11), 173–182. https://doi.org/10.32996/ijllt.2023.6.11.21

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019, September 25). *BERTScore: Evaluating Text Generation with BERT*. International Conference on Learning Representations. https://openreview.net/forum?id=SkeHuCVFDr