DSBKTE at ATE-IT: From Token Classification to Zero-Shot Generation: Two Approaches to Italian ATE at EVALITA 2026

Danil Smirnov[1,*], Bita Khashechian[1,*] and Giorgio Maria Di Nunzio[1]

[1]*Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy*

Abstract

This paper presents our participation in Subtask A of the ATE-IT shared task at EVALITA 2026, which targets sentence-level Automatic Terminology Extraction from Italian administrative documents in the waste-management domain. The dataset includes a balanced collection of administrative acts (e.g., municipal regulations, service charters, tenders) and informative texts (e.g., public notices), combining formal institutional language with less rigid informational prose. These heterogeneous documents are characterized by syntactically complex sentences and dense specialized terminology. Given an input sentence, systems must identify all domain-relevant single- and multiword terms and output them in a predefined JSON format, requiring generalization beyond a fixed vocabulary. We investigate and compare two strategies: (i) a supervised neural approach based on an Italian BERT model fine-tuned for token-level term span detection, and (ii) a zero-shot approach using large language models to extract terminology without task-specific training. We report official results and provide an analysis of typical errors to highlight the relative strengths and limitations of supervised versus zero-shot methods for terminology extraction in complex institutional texts.

Keywords

Automatic Terminology Extraction, Information Extraction, BERT, Large Language Models, Italian Administrative Texts

1. Introduction

Automatic Terminology Extraction (ATE) is the task of automatically identifying domain-specific terms from textual data [1]. These terms typically correspond to multiword expressions or noun phrases that denote relevant technical, legal, or institutional concepts within a specific domain. Accurate terminology extraction plays a crucial role in many downstream Natural Language Processing applications, including information retrieval, document classification, ontology construction, and knowledge base population.

The ATE-IT [2] shared task, organized within the EVALITA 2026 evaluation campaign [3], focuses on terminology extraction from Italian administrative documents in the waste-management domain. The dataset is composed of heterogeneous official texts such as municipal regulations, service charters, public tenders, and informational notices issued by local authorities. These documents include both administrative acts and informative texts. Administrative acts typically exhibit formal institutional language and recurring syntactic patterns, whereas informative texts may adopt less rigid stylistic structures. Both document types are characterized by long, syntactically complex sentences and a high density of specialized terminology, making the extraction task particularly challenging.

ATE-IT is structured into two subtasks. Subtask A addresses sentence-level Automatic Term Extraction, where systems are required to identify all relevant domain terms occurring in a given sentence. For each input sentence, the system must output a list of extracted terms following a predefined JSON format. The extracted terms may consist of single words or multiword expressions and are not restricted to a fixed vocabulary, requiring models to generalize beyond surface-level patterns.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT

*Corresponding author.

danil.smirnov@studenti.unipd.it (D. Smirnov); bita.khashechian@studenti.unipd.it (B. Khashechian); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

0000-0001-7116-9338 (G. M. Di Nunzio)

In this work, we focus exclusively on Subtask A. We investigate two approaches to term extraction: a supervised neural model based on BERT [4] fine-tuned for token classification, and a zero-shot approach based on large language models. The supervised approach leverages annotated training data to learn precise term boundaries, while the zero-shot method explores the ability of generative models to identify domain terminology without task-specific training. Through this comparison, we aim to analyze the strengths and limitations of supervised and zero-shot strategies for terminology extraction in complex administrative texts.

The remainder of this paper is organized as follows. Section 2 reviews related work on supervised and zero-shot terminology extraction. Section 3 describes the task and dataset. Section 4 presents the system architectures and experimental setup. Section 5 reports quantitative results. Section 6 discusses error patterns, limitations, and future directions. Section 7 concludes the paper.

2. Related Work

Automatic Terminology Extraction (ATE) has been addressed using statistical, linguistic, and supervised learning approaches [5]. Early methods relied on corpus-based termhood measures such as frequency statistics, TF–IDF variants, and heuristics like C-value to handle multiword terms, but they depend on reliable candidate generation and are sensitive to domain-specific phrasing [6].

Linguistic pipelines typically extract noun phrases using POS tagging or parsing. While grammatical constraints can improve precision, their effectiveness strongly depends on the quality of NLP tools and often degrades in administrative texts with complex formatting and enumerations.

More recent work formulates ATE as a supervised sequence labeling task, closely related to Named Entity Recognition. Neural models based on BiLSTM–CRF and, in particular, Transformer architectures such as BERT achieve accurate boundary detection when fine-tuned with BIO labels, though they require careful subword alignment and post-processing [4].

Large language models have recently enabled zero-shot and few-shot ATE via prompt-based inference. These approaches generalize without task-specific training but often exhibit limited controllability, sensitivity to prompt formulation, and tendencies toward over-generation or boundary inconsistency in structured extraction tasks [7, 8]. Models such as Gemini [9] and LLaMA [10] show strong instruction-following capabilities, yet may still generate annotation-inconsistent spans without additional constraints.

In shared-task evaluations such as EVALITA, systems are compared under standardized formats and strict term-level evaluation. Within this setting, supervised token classification and zero-shot LLM-based extraction represent complementary strategies, trading boundary precision for generalization capability [11].

3. Task and Dataset Description

The ATE-IT shared task [2] evaluates Automatic Terminology Extraction systems on Italian administrative documents in the waste-management domain. The dataset is derived from the ItaIst_GRU corpus [12] and consists of a balanced collection of administrative acts and informative texts issued by local authorities. Administrative acts typically exhibit formal institutional language and recurring syntactic patterns, while informative texts may adopt less rigid stylistic structures. Both document types are characterized by a high density of multiword terminology, making boundary detection particularly challenging.

ATE-IT is organized into two subtasks. In this work, we focus on Subtask A, which addresses sentence-level term extraction. Given an input sentence, systems must identify all domain-specific terms appearing in the sentence and output them as a list of strings. Terms may be single words or multiword expressions and are not restricted to a predefined vocabulary.

The dataset is provided in CSV format and split into training, development, and test sets. Each row corresponds to a sentence–term pair; sentences with multiple annotated terms therefore appear in

multiple rows and must be aggregated at the sentence level during preprocessing.

Annotated terms cover a range of waste-management and administrative concepts (e.g., *servizio di raccolta dei rifiuti urbani*, *utenze domestiche*), many of which are long multiword expressions, making boundary detection a key challenge.

Evaluation follows the official ATE-IT protocol, which computes micro-averaged Precision, Recall, and F_1 score at the term level using exact string matching after normalization [11].

4. Description of the system

4.1. System Overview

We compare two approaches to Automatic Terminology Extraction for ATE-IT Subtask A: a supervised sequence labeling model and a zero-shot prompt-based extraction pipeline. The two methods are implemented as independent pipelines to enable a clear and controlled comparison of their behavior and error patterns.

Both systems operate at the sentence level and share the same input/output protocol, producing a list of extracted terms in the JSON format required by the task. The supervised approach is based on a BERT token classification model trained on annotated data and optimized for precise span detection, while the zero-shot approach relies on a large language model and performs term extraction via prompt-based inference without task-specific training. No ranking or confidence estimation is applied, ensuring full compatibility with the official evaluation procedure.

4.2. Supervised BERT-based Term Extraction

The supervised approach formulates Automatic Terminology Extraction as a sequence labeling problem, where each token in a sentence is assigned a label indicating whether it belongs to a domain-specific term. We adopt the BIO tagging scheme, in which labels `B-TERM` and `I-TERM` denote the beginning and continuation of a term span, respectively, while `O` marks tokens outside any term. This formulation enables explicit modeling of term boundaries and allows the extraction of multiword expressions of arbitrary length.

Preprocessing and label construction. During preprocessing, the training data provided in CSV format is converted into sentence-level instances. Since the dataset represents sentences with multiple annotated terms across multiple rows, all annotations sharing the same document, paragraph, and sentence identifiers are aggregated. Each annotated term is then aligned with its corresponding span in the sentence text. Tokenization is performed at the word level, and BIO labels are assigned by matching normalized word sequences against the gold-standard term spans. Tokens that do not belong to any annotated term are labeled as `O`.

Model architecture. We employ a Transformer-based encoder using a pretrained Italian BERT model, fine-tuned for token classification. The model is implemented using the HuggingFace `AutoModelForTokenClassification` framework [13], which adds a linear classification layer on top of the contextualized token representations produced by BERT [4]. The classifier predicts a probability distribution over the BIO label set for each token in the input sequence.

Subword alignment. As BERT operates on subword units, careful alignment between word-level BIO labels and subword tokens is required. During training, word-level tokens are expanded into subword sequences using the model tokenizer. The BIO label assigned to each word is propagated to its corresponding subwords following standard practices. Special tokens and padding positions are masked during loss computation. At inference time, predicted subword labels are mapped back to word-level labels using the tokenizer alignment information, ensuring consistent reconstruction of term spans.

Training setup. The model is fine-tuned on the ATE-IT Subtask A training split using cross-entropy loss. Training is performed with the HuggingFace `Trainer` API, using the Adam optimizer and a fixed number of epochs. Hyperparameters such as batch size, learning rate, and maximum sequence length are selected to balance training stability and computational efficiency. No additional domain-adaptive pretraining is applied, allowing us to assess the effectiveness of the base Italian BERT model on administrative texts.

Inference and span reconstruction. During inference, the model predicts a BIO label for each subword token. These predictions are aggregated at the word level, and contiguous `B-TERM`/`I-TERM` sequences are merged to form complete term spans. The extracted terms preserve their original surface form as they appear in the input sentence. This reconstruction step is critical for producing outputs that are compatible with the exact string matching used in the official evaluation script.

4.3. Zero-shot LLM-based Term Extraction

We also test a zero-shot terminology extraction pipeline based on large language models (LLMs). In this setting, no task-specific training data is used: the model is prompted to return the domain terms appearing in each input sentence.

Model. Our main system uses Gemini-2.5-Flash [9]. We additionally tested LLaMA 3.1 [10] and GPT-5.1, but Gemini produced the most stable formatting and the best development-set performance among the evaluated LLMs.

Prompting and output format. Sentences were processed in batches, and the model was instructed to extract waste-management and administrative terms and to output one list of terms per sentence (without explanations), matching the Subtask A requirements. We report below the exact prompt template used for all zero-shot experiments. The same instruction was applied uniformly across all batches, without in-context examples.

```
You are an automatic term extraction agent.
You will receive a list of Italian sentences about municipal waste management.
Your role is to extract waste management terms from the sentences.
Output a list of terms for each sentence.

Strictly adhere to the Example Output Format:

Sentence 1: [term1; term2; term3]
Sentence 2: [term4]
Sentence 3: []

Instructions:
* Extract only domain-relevant terms related to waste management.
* Terms can be single- or multi-word.
* Use only lowercase in the output.
* Do not output nested terms (if you output a longer term,
  do not also output its inner substring as a separate term).
* Do not output named entities such as city names or organisation names.
* If a sentence contains no terms, output an empty list [].
* Output exactly N lists of terms, one for each input sentence.
```

Post-processing. We apply lightweight normalization (whitespace and punctuation cleanup) and remove duplicates. No ranking or confidence scoring is used.

4.4. Experimental Setup

All experiments are conducted using the official ATE-IT dataset for Subtask A, which is split into training, development, and test sets. The supervised BERT-based model is trained on the training split and evaluated on the development split. The development data is used for model selection, qualitative analysis, and comparison with the zero-shot approach, while the test split is reserved exclusively for final submission.

The zero-shot LLM-based approach does not rely on annotated training data and is applied directly to the development and test splits using the same extraction pipeline, enabling a direct comparison between supervised and zero-shot methods under identical conditions.

Evaluation metrics. Performance is measured using the official ATE-IT evaluation script for Subtask A, which computes micro-averaged Precision, Recall, and F_1 score at the term level. A prediction is considered correct only if it exactly matches a gold-standard term after normalization; partial matches and overlapping spans are counted as errors, emphasizing strict boundary accuracy.

Implementation details. The supervised model is implemented with the HuggingFace Transformers library and fine-tuned using the `Trainer` API with cross-entropy loss. Input sequences are truncated or padded to a fixed maximum length, and padding tokens are excluded from loss computation. All experiments are run with a fixed random seed to limit variability.

The zero-shot pipeline is implemented as a prompt-based interface to large language models. Each sentence is processed with a single prompt, and outputs are normalized through post-processing. No in-context examples are provided, ensuring a purely zero-shot setting.

Reproducibility. Preprocessing and model configurations are kept fixed across runs, and evaluation is consistently performed with the official script and normalization settings. While exact reproducibility of zero-shot outputs cannot be guaranteed due to model non-determinism, prompt design and post-processing remain unchanged throughout all experiments.

5. Results

We report results for both the development and test sets using the official ATE-IT evaluation script for Subtask A. The script computes term-level micro-averaged Precision, Recall, and F_1 score by comparing system predictions against gold-standard annotations under a strict exact-match criterion. All predictions are generated in the required JSON format using the pipelines described in the previous sections.

Development set. On the development set, the supervised BERT-based model achieves strong performance, reaching a micro-averaged Precision of 0.73, Recall of 0.67, and an F_1 score of 0.70 (Table 1). These results indicate that the model effectively learns term boundaries from annotated data and maintains a good balance between precision and recall across heterogeneous administrative sentences.

The zero-shot LLM-based approach, evaluated under the same conditions, shows a different performance profile. It attains a Precision of 0.48 and Recall of 0.60, resulting in an F_1 score of 0.53. While recall is relatively high, the lower precision reflects a tendency to over-generate candidate terms or produce imprecise spans, which are penalized by the exact-match evaluation. All zero-shot results on the development set are obtained using the Gemini `2.5-flash` model [9].

Test set. Following the release of the official test set, we submitted both systems using the same configurations adopted for development experiments. Table 2 reports the official test-set results, together

Approach	Precision	Recall	F1-score
BERT (supervised)	0.73	0.67	0.70
Zero-shot LLM (Gemini)	0.48	0.60	0.53

Table 1
Term-level micro-averaged Precision, Recall, and F_1 score on the ATE-IT Subtask A development set.

Approach	Micro P	Micro R	Micro F1	Type P	Type R	Type F1
Baseline	0.497	0.559	0.526	0.435	0.508	0.469
BERT (supervised)	0.617	0.578	0.597	0.576	0.460	0.512
Zero-shot LLM (Gemini)	0.471	0.514	0.492	0.425	0.489	0.455

Table 2
Official test-set results for ATE-IT Subtask A. Micro-level metrics evaluate term occurrences, while type-level metrics evaluate distinct term types.

with the baseline provided by the ATE-IT organizers. In addition to term-level (micro) metrics, we also report type-level Precision, Recall, and F_1, which evaluate coverage of distinct terminology types.

On the test set, the supervised BERT-based model consistently outperforms the baseline across all metrics, improving Micro F_1 from 0.526 to 0.597. This gain is primarily driven by higher precision, indicating more accurate boundary detection and fewer spurious extractions. Type-level results show a similar trend, with the BERT model achieving higher Type F_1 than the baseline, suggesting better coverage of distinct domain terms.

The zero-shot LLM-based approach exhibits stable recall but lower precision, leading to Micro and Type F_1 scores below the baseline. This behavior mirrors the development-set results and confirms that, under strict exact-match evaluation, zero-shot extraction tends to favor coverage over boundary accuracy.

Overall, results on both development and test sets consistently show that supervised sequence labeling with BERT provides the most reliable performance when annotated data is available. Zero-shot LLM-based extraction remains competitive in terms of recall and demonstrates robustness without task-specific training, but its lower precision limits performance under the evaluation criteria adopted in ATE-IT.

6. Discussion

The experimental results confirm that the supervised BERT-based approach provides the most reliable boundary detection under the strict exact-match evaluation adopted in ATE-IT. Leveraging annotated training data, the model achieves strong precision and stable performance across heterogeneous administrative sentence structures, particularly for well-formed multiword terms frequently observed in the corpus. However, its reliance on supervision constrains generalization: rare expressions, highly variable syntactic constructions, and implicitly expressed terminology remain challenging, reflecting the inherent limitations of purely supervised learning in specialized domains.

6.1. Error Analysis

We perform a qualitative error analysis on the development set to identify recurring failure patterns of the supervised BERT-based model and the zero-shot LLM-based extractor. Manual inspection reveals several systematic error categories.

Boundary errors. Both approaches suffer from boundary inaccuracies. The supervised model often truncates long multiword terms with coordination or prepositional attachments, extracting only the core noun phrase. Conversely, the zero-shot approach tends to generate overly long spans by merging

adjacent expressions or including punctuation, especially in enumerations. Although semantically plausible, such errors are penalized by exact-match evaluation.

Over-generation. The zero-shot system frequently over-generates candidate terms in long administrative sentences, including generic or discourse-level nouns not annotated as domain terms. This behavior improves recall but significantly reduces precision, reflecting a preference for semantic salience over annotation consistency.

Missed terms. Both systems occasionally miss relevant single-word or implicit terms, particularly in complex syntactic contexts. For the supervised model, these errors are more common for rare expressions or terms underrepresented in the training data, indicating sensitivity to data sparsity.

Formatting issues. Casing and punctuation inconsistencies affect both approaches and often result in incorrect matches under the strict evaluation protocol, despite limited semantic impact.

Conservative vs. exhaustive behavior. The supervised model may produce empty outputs for difficult sentences, while the zero-shot model almost always returns at least one candidate term, confirming its more exhaustive but less precise behavior.

Overall, the supervised approach provides more consistent and precise boundaries, whereas the zero-shot method offers broader coverage at the cost of increased noise. These complementary error patterns suggest that hybrid strategies could further improve robustness.

6.2. Future Work

The experimental results highlight different performance profiles of supervised and zero-shot approaches for Automatic Terminology Extraction in Italian administrative texts. The supervised BERT-based model achieves higher precision and overall F_1 score, indicating that sequence labeling with annotated data remains the most reliable strategy when accurate term boundaries are required. Its strength lies in consistent extraction of well-formed multiword terms that follow recurring syntactic patterns.

The zero-shot LLM-based approach, while less precise, exhibits higher recall and robustness to previously unseen expressions. This suggests that large language models encode useful domain and linguistic knowledge that can be leveraged without task-specific training. Such behavior is particularly relevant in low-resource scenarios or domains where annotated data is limited or rapidly outdated.

Error analysis shows that many failures are driven by structural properties of the data. Long sentences, complex nominal constructions, and enumerations challenge both approaches in different ways. The supervised model tends to be conservative and may miss relevant terms, whereas the zero-shot model frequently over-generates candidates or produces imprecise boundaries. These complementary error profiles indicate that no single paradigm fully addresses the complexity of the task.

These observations motivate the exploration of hybrid strategies that combine supervised and zero-shot extraction. For example, zero-shot predictions could be used as a fallback when the supervised model produces empty outputs, or as a candidate generation step followed by supervised or rule-based filtering.

Future work may also consider domain-adaptive pretraining of BERT on administrative or legal corpora, span-based extraction models to reduce boundary fragmentation, and the integration of syntactic information to better constrain term spans. For LLM-based methods, further research is needed on improving controllability and boundary consistency through prompt optimization or lightweight adaptation. Overall, combining supervised learning with large language models appears to be a promising direction for improving terminology extraction in complex administrative domains.

6.3. Limitations and Ethical Considerations

The proposed approaches present several limitations. The supervised BERT-based model depends on the availability and quality of annotated data. Although the ATE-IT dataset provides reliable annotations, its limited size and domain scope may restrict generalization to administrative documents with different structure, terminology, or legal context. Rare terms and implicitly expressed domain concepts are particularly prone to being missed.

A further limitation concerns boundary sensitivity. Subtask A evaluation is based on exact string matching, which penalizes even minor boundary deviations such as trailing punctuation, casing mismatches, or partial span truncation. While this criterion enforces precision, it amplifies the impact of superficial errors and highlights the need for stronger normalization or alternative evaluation schemes.

The zero-shot LLM-based approach introduces additional challenges related to controllability and reproducibility. Due to inherent non-determinism, identical prompts may yield different outputs across runs. Moreover, zero-shot extraction tends to over-generate terms or produce annotation-inconsistent spans, leading to reduced precision and limiting its reliability without post-processing.

From an ethical standpoint, terminology extraction from administrative documents may affect downstream applications in the public sector, such as document retrieval or decision-support systems. Extraction errors could propagate to such applications if not properly validated. In addition, the opacity of large language models raises concerns regarding transparency and accountability, particularly in settings where explainability is required.

These considerations underline the importance of human oversight and careful validation when applying automated term extraction in administrative domains. Future work should emphasize robustness, interpretability, and responsible deployment of language technologies in public-sector contexts.

6.4. Observed behavior

Compared to the supervised model, the zero-shot system tends to favor coverage over boundary precision: it often returns semantically plausible candidates but may over-generate and produce spans that are too long, especially in enumerations and long administrative sentences. This behavior aligns with the lower precision observed in evaluation.

7. Conclusion

This work addresses Subtask A of the ATE-IT shared task at EVALITA, focusing on automatic terminology extraction from Italian administrative documents in the waste-management domain. We compare a supervised BERT-based sequence labeling approach with a zero-shot large language model-based extraction pipeline.

Results show that the supervised BERT model achieves the highest term-level F_1 score, benefiting from explicit supervision and accurate boundary detection. The zero-shot approach, while less precise, demonstrates higher recall and the ability to identify relevant terminology without task-specific training.

Error analysis highlights challenges inherent to administrative texts, including long sentences, complex nominal structures, and strict exact-match evaluation. These factors affect both approaches in different ways and explain the observed performance differences.

Overall, this study provides a controlled comparison of supervised and zero-shot methods under the ATE-IT evaluation framework and clarifies their respective trade-offs in terms of precision, recall, and generalization for terminology extraction in administrative domains.

8. Declaration on Generative AI

During the preparation of this work, **Grammarly** was used for grammar and spelling checking. **ChatGPT** and **Gemini** were used for rephrasing and sentence polishing to improve clarity and readability. **ChatGPT** were also used to convert parts of the manuscript into LaTeX code and to adapt the document

to the CEUR-WS one-column CEUR-ART style (formatting assistance). In addition, **Google Translate** was used for text translation and subsequent language polishing. All content was reviewed and edited by the authors, who take full responsibility for the final manuscript.

References

[1] K. Kageura, B. Umino, Methods of automatic term recognition: A review, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 3 (1996) 259–289. doi:10.1075/term.3.2.03kag.

[2] N. Cirillo, G. M. Di Nunzio, F. Vezzani, Ate-it at evalita 2026: Overview of the automatic term extraction italian testbed task, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[3] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL-HLT (2019).

[5] G. M. Di Nunzio, S. Marchesin, G. Silvello, A systematic review of Automatic Term Extraction: What happened in 2022?, Digital Scholarship in the Humanities 38 (2023) i41–i47. doi:10.1093/llc/fqad030.

[6] K. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms:. the C-value/NC-value method, International Journal on Digital Libraries 3 (2000) 115–130. doi:10.1007/s007999900023.

[7] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: International Conference on Learning Representations, 2020.

[8] L. Ouyang, J. Wu, X. Jiang, et al., Training language models to follow instructions with human feedback, in: Advances in Neural Information Processing Systems, 2022.

[9] Google DeepMind, Gemini: A family of highly capable multimodal models, 2024. URL: https://deepmind.google/technologies/gemini/.

[10] Meta AI, The llama 3.1 language model, 2024. URL: https://ai.meta.com/llama/.

[11] N. Cirillo, et al., Ate-it: Automatic terminology extraction for italian, in: Proceedings of EVALITA, 2025. URL: https://nicolacirillo.github.io/ate-it/, shared Task description.

[12] N. Cirillo, D. Vellutino, D. Nicoletti, E. Sabarese, B. Rubino, Itaist_gru, 2025. URL: https://doi.org/10.5281/zenodo.15173712. doi:10.5281/zenodo.15173712.

[13] HuggingFace, Huggingface transformers, 2024. URL: https://huggingface.co/transformers/.