



# Leveraging Information Retrieval to Enhance Spoken Language Understanding Prompts in Few-Shot Learning

Pierre Lepagnol<sup>1,2</sup>, Sahar Ghannay<sup>1</sup>, Thomas Gerald<sup>1</sup>, Christophe Servan<sup>1</sup>, Sophie Rosset<sup>1</sup>

<sup>1</sup>Université Paris-Saclay, CNRS, LISN, France

<sup>2</sup>SCIAM, France

firstname.lastname@lisn.fr

## Abstract

Understanding user queries is fundamental in many applications, such as home assistants, booking systems, or recommendations. Accordingly, it is crucial to develop accurate Spoken Language Understanding (SLU) approaches to ensure the reliability of the considered system. Current State-of-the-Art SLU techniques rely on large amounts of training data; however, only limited annotated examples are available for specific tasks or languages.

In the meantime, instruction-tuned large language models (LLMs) have shown exceptional performance on unseen tasks in a few-shot setting when provided with adequate prompts. In this work, we propose to explore example selection by leveraging Information retrieval (IR) approaches to build an enhanced prompt that is applied to an SLU task. We evaluate the effectiveness of the proposed method on several SLU benchmarks. Experimental results show that lexical IR methods significantly enhance performance without increasing prompt length.

**Index Terms:** SLU, Information Retrieval, Prompt Engineering, Few-Shot Learning

## 1. Introduction

Spoken Language Understanding (SLU) is a critical component of task-oriented dialogue systems [1]. It consists of extracting information from user utterances and providing faithful information to a more extensive system. The extracted information is crucial to the dialogue systems for different sub-tasks, from querying a database (e.g., retrieving items from a database) to answering the user's request.

The SLU task can fall into three sub-tasks: domain classification, i.e. retrieve the domain of an utterance (music recommendation, hotel booking ...); Intent Classification (IC), which identifies the user's intent (e.g. hotel search, book a hotel, play music, ...); Slot-Filling (SF) [1] where it aims to extract semantics concepts (e.g., date, location, ...). In this study, we are interested in the Slot-Filling task that can also be considered as a concept detection task [2]. To illustrate the tasks, let's assume the following example: **“What is the weather like in Abu Dhabi tomorrow?”**. From this sentence, the system could extract an intent label (e.g., `weather_information`) and the different semantic information, such as `location: Abu Dhabi` and `date: tomorrow`.

In recent years, SLU benchmarks and datasets have emerged to assess the relevance of developed approaches. Some are annotated on textual interactions between humans (with a user playing the role of the system) like ATIS [3] while others are based on automatic transcription, such as MEDIA [4] or SLURP[5].

Since the mid-2000s, machine-learning supervised approaches have been widely used to perform SLU tasks, specifically SF tasks [6, 7]. Then, deep-learning approaches have enabled to reach a new step in SLU quality, for instance, mixing LSTM, CNNs and CRF approach [8]. These last years, the fine-tuning of pre-trained transformer-based models (BERT) models [9] was a quality game changer [10]. Today's approaches mainly rely on transformer pre-trained models, tackle either IC and SF independently [11] or jointly [12, 13]. While reaching the best performances, fine-tuned BERT-based approaches suffer from the number of annotated data needed to adapt the model [14] to a specific scenario. Recent approaches have focused on exploiting LLM internal knowledge to perform adaptation with limited number of examples (*few-shot* setting) [15, 16] or no examples (*zero-shot* setting) [17, 18] by prepending the prompts with slots definition. In the meantime, generative LLM is particularly efficient in mimicking the example scheme given in a prompt. In a few-shot setting, state-of-the-art approaches focus on example selection based on the user's intent [15, 18]. While such approaches are promising, they rely on randomness to select prompt examples. However, we hypothesize that random prompt examples are not always relevant for a specific utterance [19].

This work aims to study different example selection methods and build an enhanced prompt using only a limited number of relevant examples. The proposed methodology includes the following:

- Exploration of Information Retrieval approaches for SLU tasks to address the challenge of prompt efficiency. This will involve a comparison of different retrieval mechanisms such as BM25 [20] (Best Matching 25), which is a widely-used lexical ranking function and a state-of-the-art contextual embedding approach such as ColBERT model [21, 22].
- Evaluation of our example selection method across several SLU benchmarks (ATIS, SNIPS, SLURP, MEDIA).

In order to demonstrate the relevance of the present contributions, the paper is organised as follows: in section 2, the proposed approach is explained, in section 3 we describe the experimental protocol with the different models, corpora, and configurations; the results of experiments are analyzed in section 4; and finally, we conclude in section 5 and discuss limitations of the approach in section 6.

## 2. Proposed Approach

While selecting the right examples is a key challenge in order to enhance the quality of the prompt [19], to the best of our knowledge, the current state of the art concerning SLU tasks using prompt engineering for slot-filling (SF) does not select examples based on utterance similarity. In contrast, related ap-

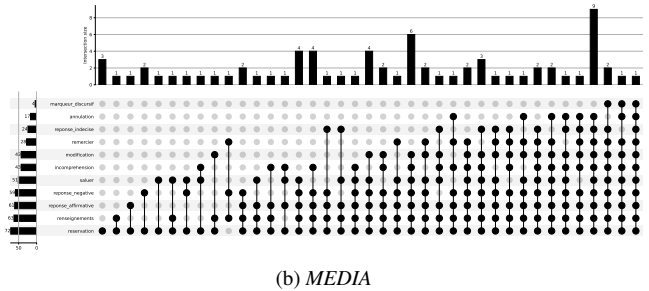
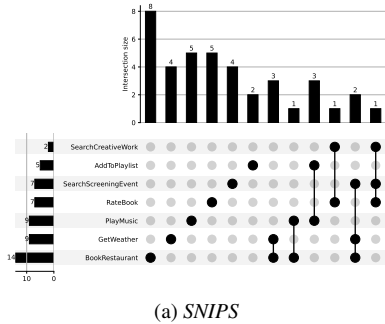


Figure 1: Upset plot illustrating the intersections of slots across different intents. In the SNIPS dataset, only a few slots are shared between intents, indicating minimal overlap. In contrast, the MEDIA dataset shows significant overlap, with many slots being shared among multiple intents.

proaches select examples randomly, using only the intention as an anchor [15, 18]. However, an analysis of the SF discrepancy linked to their intent reveals that the slots are not exclusive to intent. In some cases, the slots overlap significantly across several intents, as illustrated in Figure 1. Restricting examples in prompt to those with identical intent may lead to the exclusion of utterances with the potential to improve performance. Consequently, it would be advantageous to select examples using other features such as utterance text (or/and semantic) similarity.

To better select examples for the SLU prompting, our approach uses information retrieval (IR) methods. The proposed pipeline consists of three main steps, detailed as follows:

**Querying & Retrieving Examples** For a given IR method, we use only the utterance to be analyzed to select a set of top- $K$  closest examples. This selection is made by computing a similarity score between each utterance example from the train set  $U_t$  and the utterance to be analyzed  $U_a$ .

**Prompt Formatting** The prompt is formatted by including the retrieved examples  $U_t$  and their intent to extract slot values using a template (see Figure 2).

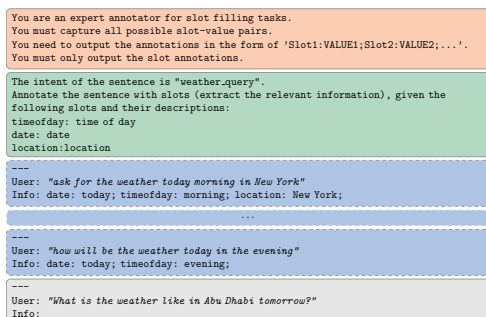


Figure 2: Prompt Template: In red, the context. In green are the instructions. In blue, examples with labeled slots. In gray is the current utterance we want to label.

**Generating & Parsing** The formatted prompt is fed into the LLM after applying the proper prompt template to generate 256 tokens using greedy decoding. The response is finally parsed using regular expressions that catch each slot type/value pair.

### 3. Experimental Setup

This section outlines the experimental methodology for evaluating our approach. We present the datasets used, describe

the chosen language model, detail the retrieval mechanisms employed, and specify the baseline and comparison methods.

#### 3.1. Datasets & Metrics

Our study focuses on four well-known SLU benchmarks (see details in table 1), evaluated using the F1-score:

	ATIS		SLURP		SNIPS		MEDIA	
	Train	Test	Train	Test	Train	Test	Train	Test
Nb. utterances	4978	893	11514	2974	13084	700	13712	3767
Nb. intents	17	16	91	77	7	7	11	11
Nb. slots	80	70	55	53	39	39	74	72

Table 1: Key characteristics of the datasets: number of utterances, the number of intents, and the number of slots in both the training and test sets.

**ATIS:** is an English dataset with queries about airline travel information [3].

**SNIPS:** is an English dataset from the SNIPS personal assistant [23].

**SLURP:** is an English dataset simulating single-turn interactions between users and a voice-controlled assistant [5].

**MEDIA:** is a French dataset about hotel reservations and information [4, 24]. MEDIA corpus is available in a *full* or a *relax scoring* version. In this study, we used the 2022 version [25] of MEDIA with the relaxed scoring version, in which attributes are simplified by excluding the specifiers. This version was enriched with intent labels by [24].

#### 3.2. Language Model

We selected an instruction/chat fine-tuned version of Llama 3.1 8B [26] model called Hermes-3-Llama-3.1-8B [27] for the following reasons: 1) Its long context window allowing to consider long sequences of tokens as input (numerous examples); 2) Its ability in generating structured outputs, making it an appropriate choice for our highly structured task.

#### 3.3. Example selection approaches

To evaluate our example selection approach within the whole training dataset, we propose to compare the two following retrieval methods: the classical lexical **BM25** [20] and the state-of-the-art contextual embedding approach **ColBERT** [21, 22] (see Appendix in supplementary materials for more details). We add the **Random** setup, which consists of taking a random selection of examples from the whole example set, and the **Intent-based** selection takes random examples from the examples set

within the same intent. This last method is largely used in previous work [17, 16]; We also filter our selected examples based on the intent, before retrieval, named **Intent**→**BM25** and the reverse: **BM25** before intent selection (noted **BM25**→**Intent**).

Following [16], we compare our method (prompt template + example selection method) within the few-shot in-context learning setting using 10 examples. As their experience is partially reproducible (only a manual sub-sample of the training set was considered), we use their prompt template and their selection examples to ensure a fair comparison denoted as *Hermes-3* [Recomputed]<sup>1</sup>.

We also reported F1-scores from fully fine-tuned models. Namely HAN (Higher-order Attention Network) [28], FlauBERT-oral-ft [24] and HERMIT[29] a Bi-LSTM with CRF model. They represent the current SOTA for the selected SLU Benchmarks.

## 4. Results & analysis

In this section, we report and discuss the results of the different experiments. It includes a comparison to state-of-the-art SLU methods, the evaluation of the different example selection methods, the impact of retrieved utterances on performances, and the impact of the number of selected examples.

### 4.1. Comparative analysis with State-of-the-Art Methods

Table 2 presents a comparison between the proposed approaches and state-of-the-art. The results, measured in F1-scores, showcase three main categories of models: fully fine-tuned, few-shot prompting with 10 examples from [16] and the proposed approaches using *Hermes-3-Llama-3.1-8B* model.

Fully fine-tuned models, particularly HAN, FlauBERT-oral-ft, and HERMIT, are the topline and demonstrate superior performance on all datasets, achieving the highest F1-scores on ATIS (97.23), MEDIA (87.75), SNIPS (98.26) and SLURP (78.19) [24, 28, 29, 5]. Our baseline uses the experiments from [16] using the *Hermes-3-Llama-3.1-8B* model, which reaches 83.98 of F1-scores. This approach uses intent selection from a manually selected pool of examples from the SNIPS corpus. Since it is not possible to reproduce this manual selection for every corpus, we propose to select the utterances following the approaches we presented in section 3.3. The results of *flan-t5-xxl* and *WizardLM-13B-V1.1* come from [16].

Random selection, serving as baseline, achieves only 38.43%F1, and does not ensure that the retrieved examples align with the specific query. In contrast, intent-based selection, which is the closest to [16] approach, improves performance to 55.23 points in F1 by ensuring that examples share the same high-level intent. Yet, it overlooks fine-grained lexical or slot-level similarities, crucial for this task.

According to the F1 scores, BM25-based methods consistently outperform other selection methods. Notably, the BM25 method is enhanced by incorporating intent information. The “BM25 → Intent” approach performed similarly to the “Intent → BM25” method. However, on ATIS and MEDIA it averaged about 0.794 points lower, while on SLURP and SNIPS the gap increased to roughly 1.5 points.

Finally, the ColBERT approach lies behind BM25 with 55.73 points and 67.61 points, respectively. It indicates that semantic information is less effective than lexical information

<sup>1</sup>Note that their best model is no longer available, which led us to use the *Hermes-3-Llama-3.1-8B* model instead

in selecting examples for prompting. We can conclude that the lexical alignment between utterances and slots is the most valuable information to pass on to the LLM.

Model & Selection	ATIS	MEDIA	SLURP	SNIPS	Mean
Fully Fine-tuned models (need the full training set)					
FlauBERT-oral-ft [24]	-	<b>87.75</b>	-	-	N/A
HAN [28]	<b>97.23</b>	-	<u>55.50</u>	<b>98.26</b>	N/A
HERMIT[29, 5]	-	-	<b>78.19</b>	-	N/A
Few-shot Prompting [16] using 10 examples					
flan-t5-xxl	-	-	47.30	64.70	N/A
WizardLM-13B-V1.1	-	-	47.40	68.50	N/A
<i>Hermes-3</i> [Recomputed]	-	-	-	83.98	N/A
This work: Few-shot Prompting using 10 examples					
<i>Hermes-3</i> + Random	68.24	26.55	15.20	43.73	38.43
<i>Hermes-3</i> + Intent-Based	72.37	34.78	35.72	78.03	55.23
<i>Hermes-3</i> + ColBERT	83.57	45.23	24.02	70.11	55.73
<i>Hermes-3</i> + BM25	86.63	55.56	43.87	84.38	67.61
<i>Hermes-3</i> + BM25→Intent	85.11	60.78	46.14	85.21	69.31
<i>Hermes-3</i> + Intent→BM25	86.38	61.10	44.01	84.32	68.95

Table 2: Results (F1-scores) in Slot Filling task. In **bold** the best score, underline the second best score.

### 4.2. Impact of Retrieved utterances on Performances

In this section, we examine the impact of different selection methods to identify the usefulness of the retrieved examples. We hypothesized that the retrieved examples add additional information for the model to use to generate the correct slots.

dataset	method	Slot Presence	Intents Presence
ATIS	BM25	<b>89.75% ± 7.47%</b>	94.27% ± 5.32%
	ColBERT	80.91% ± 15.06%	84.95% ± 12.41%
	Intent-Based Selection	75.59% ± 14.06%	<b>99.34% ± 0.34%</b>
	Random Selection	67.14% ± 21.49%	76.70% ± 12.46%
MEDIA	BM25	<b>90.53% ± 7.76%</b>	93.42% ± 5.38%
	ColBERT	65.75% ± 22.74%	72.49% ± 22.37%
	Intent-Based Selection	75.61% ± 16.87%	<b>98.18% ± 2.02%</b>
	Random Selection	65.00% ± 23.23%	68.05% ± 26.71%
SLURP	BM25	<b>89.36% ± 8.83%</b>	87.49% ± 9.95%
	ColBERT	59.53% ± 15.15%	44.21% ± 25.73%
	Intent-Based Selection	83.37% ± 14.15%	<b>99.97% ± 0.00%</b>
	Random Selection	57.76% ± 21.41%	28.87% ± 32.80%
SNIPS	BM25	<b>89.81% ± 11.49%</b>	96.41% ± 5.14%
	ColBERT	72.77% ± 21.28%	88.73% ± 10.92%
	Intent-Based Selection	83.52% ± 17.86%	<b>100.00% ± 0.00%</b>
	Random Selection	52.18% ± 35.10%	58.08% ± 35.67%

Table 3: Comparison of the presence of expected slots and intents between each selection method. For example, in ATIS for BM25, 89.75% of the prompt have the right and expected slots into the prompt. In **bold** the best score, underline the second best score.

Table 3 presents the percentage of examples where each method retrieves correct intents and slots (on average over the number of examples). We observe that, BM25 selects more frequently examples containing the necessary slots compared to Intent-based Selection or Random selection, hence it achieved the best score F1-score. BM25 misses the correct slots in only 10% of the test set against about 20-25% for Intent-based methods. On the MEDIA dataset, 90.53% of the prompts include the appropriate slots versus 75.61% for Intent-Based Selection. Furthermore, BM25 selects examples based on the words in the utterance, meaning the selected examples are very close to the current utterance we want to analyze.

The semantic retrieval method with ColBERT underperforms BM25 method and often Intent-Based Selection (on



Figure 3: Performance of the different methods for each number of examples in the prompt for ATIS, SLURP, SNIPS, and MEDIA

SNIPS and SLURP). In the same way, ColBERT miss more often the appropriate slots compared to Intent-Based Selection.

In summary, the results demonstrate that retrieval methods integrating BM25, especially when combined with intent information (either before or after retrieval), substantially outperform random retrieval approaches. The optimal configuration appears to be dataset-dependent: BM25 excels in datasets like ATIS, where slot types are less varied, while intent-guided BM25 methods perform better in more complex datasets like MEDIA and SLURP. This could suggest that tailoring the retrieval strategy to the specific characteristics of the dataset can lead to significant improvements in Slot Filling performance or some redundancies in the ATIS dataset make it easier. Nevertheless, as ATIS was published in the 1990s, it is possible that it was subsequently leaked onto the internet and incorporated into the training set of the *Hermes-3-Llama-3.1-8B* or *Llama 3.1* models.

### 4.3. Varying the number of examples

We study the impact of the number of examples in prompts across the selection approaches and the benchmarks used. This aims to determine what could be the right amount of data needed to perform SLU with the *Hermes-3-Llama-3.1-8B*.

Figure 3 presents the performance of each selection method across the different datasets by varying the number of examples in the prompt. Using BM25 methods (BM25, BM25 → intent and, intent → BM25) consistently outperforms the others methods. Particularly, considering only five examples using BM25 always outperforms random methods considering 100 examples. Consequently, retrieval-based prompts improve performances while benefiting from a lower computational cost (limiting the input sequence length and thus lowering the cost of self-attention processing). While Intent-Based selection significantly improves at higher shot counts, it never surpasses BM25-based methods. These findings suggest that leveraging BM25 as a foundation for slot-filling provides a significant advantage in scenarios with limited labelled examples.

## 5. Conclusion

This study investigates the use of information retrieval (IR) methods to enhance prompt construction in spoken language understanding (SLU) task.

By integrating IR methods, specifically BM25, into the example selection process, we have addressed the challenges of overlapping intents and slots in complex datasets. Our com-

prehensive evaluation across multiple benchmarks showed that BM25 consistently improved F1 performance scores over traditional intent-based selection methods. Moreover, these improvements were achieved without increasing prompt length. We investigated the impact of the number of selected examples and compared both lexical and semantic IR methods. BM25-based methods showed consistent efficiency in different SLU Benchmarks. The effectiveness of the approach is highly dependent on the quality of the retrieval mechanism; suboptimal retrieval could adversely affect overall performance, as demonstrated by the results of ColBERT, which failed to retrieve relevant examples. While the method maintains similar prompt lengths, the retrieval process can introduce additional computational overhead, with larger datasets, which can be significant in real-time applications where latency is critical.

In conclusion, BM25 represents a step forward in the field of prompting for SLU tasks. By effectively balancing specificity and length, and improving performance without additional computational cost, our method offers a practical solution to improve the accuracy and applicability of SLU systems across diverse linguistic and industrial contexts.

To ensure reproducibility, the code and resources used in this study are available at <https://gitlab.lisn.upsaclay.fr/phd-pierre-lepagnol/ir-for-slu>.

## 6. Limitations

Despite promising advances, this study on language models for SLU tasks may have several limitations. Firstly, the reliance on access to a training dataset, for example retrieval, poses challenges, particularly in low-resource settings where such data may be scarce or unavailable, thus limiting the applicability of the method. In addition, there is a possibility that some models performed well because they were trained on all or a part of our considered datasets, namely ATIS and SNIPS. The study also used a single instruction prompt without exploring variations of the task formulation, potentially limiting the understanding of how different prompt designs affect performance. Furthermore, the study focused on one model, potentially overlooking other competitive models. Finally, the robustness of the method in dealing with rare or highly ambiguous slots has not been thoroughly explored, and inaccuracies in initial intent classification could lead to the selection of inappropriate examples, thereby degrading the performance of the system. Future work will focus on addressing these limitations.

## 7. Acknowledgments

This work is supported by the ANRT (Association nationale de la recherche et de la technologie) with a CIFRE fellowship granted to SCIAM<sup>2</sup> (CIFRE N°2022/1608).

This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014242).

## 8. References

- [1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, 2011.
- [2] H. Bonneau-Maynard, C. Ayache, F. Béchet, A. Denis, A. Kuhn, F. Lefevre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, and J. Villaneau, “Results of the French Evalda-Media evaluation campaign for literal understanding,” in *LREC 2006*, Genes, Italy, Mai 2006.
- [3] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The ATIS spoken language systems pilot corpus,” in *Speech and Natural Language: Proc. of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990. [Online]. Available: <https://www.aclweb.org/anthology/H90-1021>
- [4] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, “Semantic annotation of the French media dialog corpus,” in *Proc. Interspeech 2005*, 2005, pp. 3457–3460.
- [5] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, “SLURP: A spoken language understanding resource package,” in *EMNLP, Pennsylvania, June 24-27, 1990*, 1990. Association for Computational Linguistics, Nov. 2020, pp. 7252–7262.
- [6] C. Servan, C. Raymond, F. Béchet, and P. Nocera, “Conceptual decoding from word lattices: application to the spoken dialogue corpus MEDIA,” in *Interspeech 2006 - ICSLP*, Pittsburgh, United States, Sep. 2006.
- [7] S. Hahn, M. Dinarelli, C. Raymond, F. Lefevre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, “Comparing stochastic approaches to spoken language understanding in multiple languages,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1569–1583, 2011.
- [8] X. Ma and E. Hovy, “End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF,” in *ACL 2016*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016.
- [9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. of NAACL*, 2019. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [10] S. Ghannay, C. Servan, and S. Rosset, “Neural Networks approaches focused on French Spoken Language Understanding: application to the MEDIA Evaluation Task,” in *COLING’2020, Barcelona (online)*, Spain, Dec. 2020.
- [11] G. Castellucci, V. Bellomaria, A. Favalli, and R. Romagnoli, “Multi-lingual intent detection and slot filling in a joint bert-based model,” *arXiv preprint arXiv:1907.02884*, 2019.
- [12] Q. Chen, Z. Zhuo, and W. Wang, “Bert for joint intent classification and slot filling,” *arXiv preprint arXiv:1902.10909*, 2019.
- [13] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, “A joint learning framework with bert for spoken language understanding,” *IEEE Access*, 2019.
- [14] H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, “A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–38, Aug. 2023.
- [15] M. He and P. N. Garner, “Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding,” *Tech. Rep.*, Aug. 2023, arXiv:2305.13512 [cs, eess] type: article.
- [16] P. Mirza, V. Sudhi, S. Sahoo, and S. R. Bhat, “Illuminer: Instruction-tuned large language models as few-shot intent classifier and slot filler,” in *LREC-COLING*, Torino, Italy, May 2024.
- [17] Z. Zhu, X. Cheng, H. An, Z. Wang, D. Chen, and Z. Huang, “Zero-shot spoken language understanding via large language models: A preliminary study,” in *LREC-COLING 2024*, Torino, Italia, May 2024, pp. 17 877–17 883.
- [18] L. Qin, F. Wei, Q. Chen, J. Zhou, S. Huang, J. Si, W. Lu, and W. Che, “Croprompt: Cross-task interactive prompting for zero-shot spoken language understanding,” 2024.
- [19] N. Nashid, M. Sintaha, and A. Mesbah, “Retrieval-based prompt selection for code-related few-shot learning,” in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2023, pp. 2450–2462.
- [20] S. Robertson and H. Zaragoza, “The probabilistic relevance framework: Bm25 and beyond,” vol. 3, no. 4, pp. 333–389, apr 2009.
- [21] O. Khattab and M. Zaharia, “Colbert: Efficient and effective passage search via contextualized late interaction over BERT,” in *ACM SIGIR 2020, Virtual*. ACM, 2020, pp. 39–48.
- [22] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, “Colbertv2: Effective and efficient retrieval via lightweight late interaction,” 2022.
- [23] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril *et al.*, “Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces,” *arXiv preprint arXiv:1805.10190*, 2018.
- [24] N. Alavoine, G. Laperriere, C. Servan, S. Ghannay, and S. Rosset, “New Semantic Task for the French Spoken Language Understanding MEDIA Benchmark,” in *LREC-COLING 2024*, Torino, Italy, May 2024.
- [25] G. Laperrière, V. Pelloin, A. Caubrière, S. Mdhaftar, N. Camelin, S. Ghannay, B. Jabaian, and Y. Estève, “The spoken language understanding MEDIA benchmark dataset in the era of deep learning: data updates, training and evaluation tools,” in *LREC*, Marseille, France, Jun. 2022.
- [26] A. Dubey *et al.*, “The llama 3 herd of models,” 2024.
- [27] R. Teknium, J. Quesnelle, and C. Guang, “Hermes 3 technical report,” 2024.
- [28] D. Chen, Z. Huang, X. Wu, S. Ge, and Y. Zou, “Towards joint intent detection and slot filling via higher-order attention,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 7 2022, pp. 4072–4078.
- [29] A. Vanzo, E. Bastianelli, and O. Lemon, “Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU,” in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, S. Nakamura, M. Gasic, I. Zukerman, G. Skantze, M. Nakano, A. Papangelis, S. Ultes, and K. Yoshino, Eds. Stockholm, Sweden: Association for Computational Linguistics, Sep. 2019, pp. 254–263.

<sup>2</sup><https://www.sciam.fr/>