Prompt-based Language Generation for Complex Conversational Coaching Tasks across Languages

Alain Vázquez and Maria Inés Torres
University of the Basque Country (UPV/EHU)
alain.vazquez@ehu.eus and manes.torres@ehu.eus

Abstract

We investigate the role of prompt-based demonstrators in improving natural language generation for coaching-oriented dialogue systems in different languages. These systems present significant challenges due to their need for semantically accurate, goal-driven responses across diverse dialogue act taxonomies and languages. We define three types of prompt demonstrators, i.e., pairs of meaning representation-utterance, that include different degrees of specification in such meaning representation. We then fine-tune pretrained language models separately for four very different languages and evaluate how the specificity of these demonstrators affects the quality of the generated sentences. Our experiments show that more specific prompts lead to more coherent and accurate outputs, particularly for low-resource languages and small models. Additionally, we observe promising zero-shot performance with larger models, showing a complementary value of prompts. These results demonstrate that simple prompting strategies, combined with fine-tuning, can significantly improve output quality in complex dialogue generation tasks across languages.

1 Introduction

Dialogue systems for coaching and healthcare applications must handle complex, user-centered interactions that go beyond giving or requesting information. These systems aim to support behavioral change through goal-oriented conversations, a challenge that requires high control and contextual awareness (Zhou et al., 2024; Carchiolo and Malgeri, 2024). Traditional rule-based systems fall short in flexibility, while Large Language Models (LLM)-based agents offer better adaptability but raise concerns about control and factual consistency, especially in sensitive domains like health (Carchiolo and Malgeri, 2024; Galland et al., 2024).

Table 1: EMPATHIC corpus characteristics per language.

Language	Corpus size	Running words	Vocabulary
English	1.2k	13.2k	1.7k
Spanish	1.1k	11.7k	2.0k
French	1.2k	13.0k	1.9K
Norwegian	1.1k	12.7k	1.7k

Recent work emphasizes the importance of pretrained language models and prompt-based learning for improving generation in such scenarios (Algherairy and Ahmed, 2024). These methods allow models to generalize from a few examples using in-context learning, which is especially valuable when annotated data is limited or multilingual coverage is required (Zhou et al., 2024; Liu et al., 2023).

For goal-setting dialogues, even small variations in the input meaning representation (MR) can significantly affect the response's quality and intent realization (Ramirez et al., 2023; Reed et al., 2022). Despite the wide research in prompt-based learning, the effects of some structural variations in the demonstrators are underexplored. In this work, we expand our research on how the similarity between MR inputs and demonstrators affects natural language generation in different domains (Vázquez and Torres, 2025) by introducing a cross-lingual perspective in a specific use case: complex dialogue systems in coaching contexts.

2 EMPATHIC task

The EMPATHIC project (Torres et al., 2019; Olaso et al., 2021; Vázquez et al., 2023) aimed to build a virtual coach that can assist users in changing their unhealthy habits, if any, through conversations. This virtual coach does not generally give advice or information. On the contrary, it follows the GROW model (Leach, 2020), a behavioral model that promotes users' reflection with questions. Based on this model, the virtual coach first encourages the

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 601–608
Aug 25–27, 2025.

Table 2: Examples of the EMPATHIC corpus in English.

MR	Sentence
GSQ_what_obj (action = change)	Would you like to change something in your habits?
RQ_curr_sit (numbers = five, food = fruits, food = vegetables)	Do you eat five fruits and vegetables a day?
Gen_Hello (user_name = Françoise)	Hello again, Françoise.
Gen_Agreement ( )	That's true.

users to look for a **G**oal, making them aware of the possible weakness of their routines, in particular nutritional routines, by assessing their **R**eality. Next, the users explore their **O**ptions and **W**ill to act, defining a plan for goals such as reducing the sugar/salt intake or a regular schedule for meals. Consequently, the EMPATHIC task implies higher complexity for a conversational system than traditional information providers (Wen et al., 2015; Novikova et al., 2017).

In this context, WoZ experiments (Justo et al., 2020) were carried out in Spanish, French, or Norwegian, the target languages of the project. In these experiments, a real user and a wizard, a person playing the role of the virtual coach, simulated a coaching session. These coaching sessions were transcribed and translated into English as well as into the three target languages. Therefore, the resulting EMPATHIC corpus[1] is a dataset with equivalent versions in four languages. Note that one of them is a low-resourced one: Norwegian. Table 1 shows a similar number of samples for the four languages. However, intrinsic differences between languages lead to different numbers of total words (running words) and vocabulary size.

Table 2 shows examples of coaching turns along with the MRs employed to annotate them. Questions promoted by the GROW model, such as "Would you like to change something in your habits?" or "Do you eat five fruits and vegetables a day?" resulted in domain-specific Dialog Acts (DAs) like `GSQ_what_obj` and `RQ_curr_sit`, respectively. `GSQ` and `RQ`, which stand for Goal Set Question and Reality Question, are two of the seven types of questions employed in the GROW model. Additionally, the EMPATHIC corpus also contains open-domain DAs like `Gen_Hello` for greetings or `Gen_Agreement` for agreement expressions. In total, 78 ad-hoc DAs were defined for the EMPATHIC corpus, a high number in comparison with other datasets such as E2E (Novikova et al., 2017) and ViGGO (Juraska et al., 2019) that employ one and nine DAs for 51k and 6.9k samples, respectively. Finally, the MRs include attributes with their corresponding values found in each turn, such as `food=fruits` or `action=change`.

3 Meaning representations

Prior works (Vazquez et al., 2024; Ramirez et al., 2023; Reed et al., 2022) have shown that MRs, including an example of the task to perform in the input, i.e., a task demonstrator, generate more accurate and natural sentences. In this work, we aim to confirm the effectiveness of these kinds of representations for complex tasks, already demonstrated in English (Vázquez and Torres, 2025), across languages. Specifically, we consider three MRs with a demonstrator in the input, the Prompt representations. Then, we compare them with the original MR of the EMPATHIC corpus, the Baseline representation.

Table 3 shows a schema of the four representations. The baseline is just the Input MR found in the EMPATHIC corpus, while the prompts add a task demonstrator to this input. In consequence, the models generate sentences from the same MR in all the representations. For the prompts, the demonstrators are MR-sentence pairs consisting of a Demonstrator MR and a Demonstrator sentence, which are extracted from the corpus. And although these proposed inputs present the same format, they differ in their specificity, i.e., the similarity between the Demonstrator MR and the MR Input. The different levels of alignment between demonstrators and MR are indicated in Table 3 with colors, the letters n and m, and apostrophes. Same colors and letters indicate necessarily equal, while different colors and letters and apostrophes mean not necessarily equal. Therefore, the Demonstrator MR must have only the same DA as the Input MR for Prompt 1, the same DA and number of attributes for Prompt 2, and the same DA and attributes for Prompt 3.

Table 4 presents a real example of an MR of the English EMPATHIC corpus and its corresponding task demonstrator for each prompt. The DA of the three Demonstrator MRs and the Input MR is `RQ_curr_sit`, highlighted in blue in the table. The

[1]Available at low cost on https://catalog.elra.info/en-us/repository/browse/ELRA-S0414/

Table 3: Meaning representations overview. Baseline includes only the Input MR (common for the four representations), whereas Prompt representations add a task demonstrator to the input. Same colour (bleu for DAs and red for attributes) in Task demonstrator and Input MR means necessarily identical. Similarly, letters n and m for the number of attributes are used to specify when this feature must necessarily be the same (or not). Finally, apostrophes in values, attributes, and demonstrator sentences (s, s' and s") mean not necessarily equal.

Meaning representation	Task demonstrator	Input MR
Baseline	-	**DA** ($attr_1 = val_1, \dots, attr_n = val_n$)
Prompt 1	**DA** ($attr'_1 = val'_1, \dots, attr'_m = val'_m$) + s	**DA** ($attr_1 = val_1, \dots, attr_n = val_n$)
Prompt 2	**DA** ($attr''_1 = val''_1, \dots, attr''_n = val''_n$) + s'	**DA** ($attr_1 = val_1, \dots, attr_n = val_n$)
Prompt 3	**DA** ($attr_1 = val'''_1, \dots, attr_n = val'''_n$) + s"	**DA** ($attr_1 = val_1, \dots, attr_n = val_n$)

Table 4: Task demonstrators of a real Input MR from the English EMPATHIC corpus. Similarly to Table 3, the DA in blue and attributes (in this case only 'action') in red highlight that they are identical in the Task demonstrator and the MR input. Contrastingly, demonstrator attributes in violet highlight those that differ from the input attributes.

Input MR	**RQ_curr_sit** (**action** = eat)
Prompt 1 demonstrator	**RQ_curr_sit** (**action** = tell ; **freq** = daily) + Can you tell me about your daily eating habits?
Prompt 2 demonstrator	**RQ_curr_sit** (**food** = water) + And do you think you drink enough water?
Prompt 3 demonstrator	**RQ_curr_sit** (**action** = change) + Would you like to change your eating habits?

demonstrators of Prompt 2 and Prompt 3 match the number of attributes with the baseline input, 1. In addition, the attribute of Prompt 3 is the same as the input MR, i.e., `action`. Therefore, the table depicts the different levels of specificity of the three proposed representations.

4 Experiments

The cross-lingual experiments that we present in this section follow the methodologies proposed in (Vázquez and Torres, 2025) by using the same setup and evaluation metrics but adding new language models for non-English languages.

4.1 Experimental framework

For the validation of the proposed representations, we fine-tuned GPT-2 models (Radford et al., 2019) for each pair of representation and language. In particular, we chose GPT-2 Medium for English and three versions of the GPT-2 Small for the other three languages: MarIA (Gutiérrez-Fandiño et al., 2022) for Spanish, BelGPT-2 (Louis, 2020) for French and Norwegian GPT-2 social[2] for Norwegian.

These fine-tuning experiments included 5-fold cross-validation. The byte-pair tokenizer was the same as in the pre-training (Sennrich et al., 2016). We performed over five training epochs using a learning rate scheduler with a linear warm-up that starts 5e-5, an Adam optimizer with weight decay (Kingma and Ba, 2014), and a batch size of 8. In the generation phase, we produce five sentences per MR, constrained to a maximum length of 80 tokens. We set the temperature to 1.0 in order to obtain more variable outputs.

4.2 Metrics

Table 5 describes the five evaluation metrics employed in this work. This evaluation covers different aspects of the generations. First, it includes widely-used metrics such as (BLEU) (Papineni et al., 2002) and Slot Accuracy (Li et al., 2020). Next, BERT models (Devlin et al., 2019) are employed to evaluate the semantics with different approaches for BLEURT (Sellam et al., 2020) and LaBSE (Feng et al., 2022). Finally, we include Dialogue Act Accuracy (DAC), an underexplored metric that evaluates the accuracy of classifiers to predict the source DA of a generated sentence. Thus, it measures the coherence of the generations with the DA intent given in the MR input.

4.3 Automatic evaluation

Table 6a and Figure 2 of the Appendix A present the results for the fine-tuned models and the metric evolutions, respectively. Both show that the models fine-tuned with the prompts perform better than Baseline models according to the five metrics. In addition, the more specific the task demonstrator is, the higher the benefit is: Prompt 3 is the best Prompt representation. In this regard, there are some exceptions for English and French models for DAC, where Prompt 1 outperforms the others. In these cases, we guess that even if generation can be worse, the source DA could be more easily identifiable. In fact, these particular models obtain the lowest Slot Accuracy values within the same language, which indicates more omissions in these

[2]https://hf.rst.im/pere/norwegian-gpt2-social

Table 5: Metrics description.

Metric	Evaluated aspect	Reference	Methodology to obtain the score
BLEU	Lexic	Corpus sentences for same Input MR	BLEU-4 with a smoothing function (Chen and Cherry, 2014)
BLEURT	Semantic	Corpus sentences with the same Input MR	BERT model trained with human ratings for similarity of pair of sentences
LaBSE	Semantic	Corpus sentences with the same Input MR	Cosine similarity between BERT embeddings of generations and references
Slot Accuracy	Semantic	MR Input (focused on the values)	Percentage of values of the MR Input in the generated sentence
DAC	Communicative intent	MR Input (focused on the DA)	F1-score with DistilBERT models (Sanh, 2019) as DA classifiers

Table 6: Metric scores for each combination of metric (*Metr*), language (*Lang*) and representation. In the languages, *en*, *es*, *fr* and *no* stand for English, Spanish, French and Norwegian. For the representations, *Base* is the Baseline representation, P_i refers to Prompt i representations and *Val* is the Validation score for DAC. Evolution of the metric scores epoch by epoch can be found in Figure 2 of Appendix A.

(a) After fine-tuning

Metr	BLEU				BLEURT				LaBSE				Slot Accuracy				Dialog Act Accuracy				
Lang	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Val**	**Base**	**P1**	**P2**	**P3**
en	0.18	0.19	0.20	**0.26**	0.54	0.56	0.57	**0.60**	0.66	0.67	0.68	**0.70**	0.88	0.86	0.87	**0.90**	0.68	0.65	**0.71**	0.69	0.69
es	0.11	0.13	0.15	**0.22**	0.41	0.45	0.47	**0.53**	0.59	0.63	0.64	**0.67**	0.70	0.69	0.70	**0.78**	0.66	0.53	0.68	0.65	**0.68**
fr	0.12	0.12	0.14	**0.21**	0.34	0.36	0.37	**0.44**	0.64	0.67	0.67	**0.70**	0.71	0.67	0.67	**0.73**	0.68	0.54	**0.67**	0.62	0.65
no	0.10	0.12	0.13	**0.20**	0.42	0.46	0.47	**0.51**	0.64	0.67	0.67	**0.69**	0.66	0.69	0.69	**0.75**	0.66	0.52	0.61	0.61	**0.64**

(b) Zero shot-settings

Metr	BLEU				BLEURT				LaBSE				Slot Accuracy				Dialog Act Accuracy			
Lang	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**	**Base**	**P1**	**P2**	**P3**
en	0.00	0.01	0.01	**0.01**	0.22	0.27	0.27	**0.28**	0.20	0.26	0.26	**0.27**	0.38	0.50	0.52	**0.61**	0.03	**0.21**	0.20	0.19
es	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	**0.01**	0.09	**0.13**	0.12	0.12	**0.22**	0.21	0.20	0.19	0.04	0.03	**0.04**	0.04
fr	0.00	0.00	0.00	**0.00**	**0.00**	0.00	0.00	0.00	0.15	0.17	0.17	**0.17**	0.32	0.32	0.33	**0.35**	0.03	0.04	0.03	**0.04**
no	0.00	0.01	0.01	**0.01**	0.19	0.20	**0.20**	0.20	**0.19**	0.18	0.18	0.18	0.08	0.09	**0.10**	0.09	**0.03**	0.03	0.03	0.03

generations. Finally, the improvement with the enriched inputs is more evident and similar between them in all the non-English models. Then, the demonstrators seem to be more relevant in small models.

The zero-shot capabilities of the prompt representations may also be affected by the model. Table 6b and Figure 2 show that these MRs improve Baseline results before training in English for all the metrics except for BLEU. However, this behavior does not appear in the non-English target languages. We can assume that the difference with English resides in the GPT-2 models: Medium for English and Small versions for the other languages. A bigger size and more adequate pre-training for dialogue tasks in the Medium version can result in better generations in zero-shot settings.

The metric analysis validates some of our findings about them from (Vázquez and Torres, 2025). Comparing the very low BLEU results (principally after fine-tuning) with all the other metrics, we confirm its low reliability in evaluating the generation quality. At the semantic level, BLEURT presents bigger differences in its values across languages after fine-tuning than LaBSE. BLEURT, trained with human ratings, tends to detect omissions/hallucinations that LaBSE, based on sentence embeddings, does not. This assumption is supported by slot accuracy, a metric to evaluate the percentage of omission. The behavior of this metric follows the same tendency as BLEURT. Finally, DAC, with results very close to the validation values, evidences the robustness of the models to generate coherent sentences with the expected communicative intention.

4.4 Qualitative evaluation

Table 7 shows examples of generations for four different MRs. For each example, the table provides the MR, the language, the references, the training setup (fine-tuned or zero-shot), and up to three generations for each representation. These generations are selected and ordered in the table according to the average value over all the metrics except the DAC[3]. We confirm that the specificity of the prompts is more relevant with complex inputs (first example of the Table) and in non-English languages (third example). In these cases, Prompt 3 is the only one that always produces at least one adequate sentence. For simpler inputs (second example), the most basic prompts are effective in producing good generation, even for MR inputs in which the Baseline is not.

For zero-shot settings like the last example

[3]The DAC has been excluded because it is a classification metric that only assigns values of 1 and 0 to each generation.

Table 7: Generation examples. Each example contains the MR, the language, the references from the original corpus, the training setup (fine-tuned or zero-shot), and up to three generations for each representation (Baseline, Prompt 1, Prompt 2, and Prompt 3). These generations are selected and ordered according to the average value over all the metrics except DAC, since it only assigns 1 and 0 scores as a classification metric. <no_sense_seq> stands for sequences with no sense.

MR Input	Int_travel_quest (action = do ; freq = each time)	**Language**	English
References	- Do you have something you like to do each time you go on vacation? - Is there something you like to do each time you go on vacation?	**Training setup**	Fine-tuned
Baseline	- And do you do it to do it each time? - For each trip, do you do it differently? - Do you do it each time?		
Prompt 1	- Can you tell me about the to-do for each time? - To do each time? Each time you do it, what's the objective?		
Prompt 2	- Can you tell me about your to-do for a specific date or every time? -Do you do it each time? - Can you tell me about your to-do?		
Prompt 3	- Do you have something you like to do each time you go on vacation? - And do you have something you like to do each time you go on vacation? - Each time you go on vacation?		

MR Input	Gen_Yes_no_quest (food = fish)	**Language**	English
References	- Do you eat fish?	**Training setup**	Fine-tuned
Baseline	- And fish? - And eat more fish?		
Prompt 1	- What about fish? - So you like fish, right? And fish?		
Prompt 2	- Do you eat fish? - What about fish? - And eat fish?		
Prompt 3	- Do you eat fish? - And do you eat fish? - And fish?		

MR Input	Int_viajes_comentario (acción = imagina ; acción = decir)	**Language**	Spanish
References	- Como puede imaginar, no viajo mucho. Sólo puedo escuchar lo que tienes que decir.	**Training setup**	Fine-tuned
Baseline	- ¿Y crees que tienes un objetivo muy claro en el viaje de qué te ha motivado? - ¿Te gustaría ir a un sitio que te gustaría contar? - ¿Con qué frecuencia sueles hablar con sueles afectarte?		
Prompt 1	- Estoy aprendiendo a decir. Me interesa especialmente la música y los viajes. - Tengo que contar esto.		
Prompt 2	- Imagina ¿y por qué te gusta viajar? - Imagina ¿sueles viajar con otra persona? - Imagina		
Prompt 3	- Como puede imaginar, no viajo mucho. Sólo puedo escuchar lo que tienes que decir. - Como puede imaginar, no viajo mucho.		

MR Input	RQ_curr_sit (food = fruits ; food = vegetables ; freq = a day ; numbers = five)	**Language**	English
References	Do you eat five fruits and vegetables a day?	**Training setup**	Zero-shot
Baseline	- <no_sense_seq> - <no_sense_seq> - <no_sense_seq>		
Prompt 1	- <no_sense_seq> - <no_sense_seq> - <no_sense_seq>		
Prompt 2	- <no_sense_seq> - <no_sense_seq> - <no_sense_seq>		
Prompt 3	- <no_sense_seq> Do you eat five fruits and vegetables a day? <no_sense_seq> - Do you eat five vegetables a day? - <no_sense_seq>		

of the table, most of the generations contain sequences with no sense[4]. Nonetheless, Prompt 3 often presents good generations. Note that these good outcomes are identified by the metrics, as we can observe in the generations for Prompt 3 in the table. However, among these generations, "<no_sense_seq> Do you eat five fruits and vegetables a day? <no_sense_seq>" has been identified as the best sentence instead of "Do you eat five vegetables a day?". This situation can be resolved by adding an LLM perplexity to our metrics to discard all the generations that include these kinds of sequences. All in all, despite these undesired sequences in most generations, we consider the results promising for extending the zero-shot generation approach with prompt inputs to larger LLMs such as GPT-3.5 (Brown et al., 2020) and LLaMA (Touvron et al., 2023), which have shown strong performance in prompt-based tasks.

5 Conclusions

This work explored how prompt-based inputs can enhance natural language generation in cross-lingual, coaching-oriented dialogue systems. These tasks are particularly complex due to their need for controlled, semantically rich outputs that support user reflection, often in low-resource language settings and with diverse dialogue act taxonomies. Our results show that increasing the similarity between the demonstrator and the input meaning representation improves both coherence and accuracy, especially when using smaller pre-trained models. These findings suggest that prompting strategies can meaningfully improve generation quality in highly structured, goal-driven dialogues. Future directions include testing prompt transferability across languages and domains, and integrating such strategies into real-time interactive systems.

[4]We denote these sequences as <no_sense_seq>.

Acknowledgments

This work has been partially supported by the CRYSTAL HORIZON-MSCA-SE grant 101182965, the Spanish MCIU by the BEWORD project grant PID2021-126061OB-C42, and by the University of the Basque Country research group GUI23/016.

References

Atheer Algherairy and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Vincenza Carchiolo and Michele Malgeri. 2024. Quantitative review of conversational agents in e-health using NLP and AI. In *Global Congress on Emerging Technologies (GCET-2024)*, pages 161–168. IEEE.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024. Generating unexpected yet relevant user dialog acts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 192–203.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172.

Raquel Justo, Leila Ben Letaifa, Cristina Palmero, Eduardo Gonzalez-Fraile, Anna Torp Johansen, Alain Vázquez, Gennaro Cordasco, Stephan Schlögl, Begoña Fernández-Ruanova, Micaela Silva, et al. 2020. Analysis of the interaction between elderly people and a simulated virtual coach. *Journal of Ambient Intelligence and Humanized Computing*, 11(12):6125–6140.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sarah Leach. 2020. Behavioural coaching: The GROW model. In *The Coaches' Handbook*, pages 176–186. Routledge.

Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu. 2020. Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 97–106, Online. Association for Computational Linguistics.

Sijie Liu, Yiquan Fang, Hua Cheng, Yiming Pan, Yufei Liu, and Caiting Gao. 2023. Large language models guided generative prompt for dialogue generation. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 10–17. IEEE.

Antoine Louis. 2020. BelGPT-2: a GPT-2 model pre-trained on French corpora. `https://github.com/antoiloui/belgpt2`.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.

Javier M Olaso, Alain Vázquez, Leila Ben Letaifa, Mikel De Velasco, Aymen Mtibaa, Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, Gérard Chollet, César Montenegro, Asier López-Zorrilla, et al. 2021. The EMPATHIC virtual coach: A demo. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 848–851.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Angela Ramirez, Kartik Agarwal, Juraj Juraska, Utkarsh Garg, and Marilyn Walker. 2023. Controllable generation of dialogue acts for dialogue systems via few-shot response generation and ranking. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 355–369.

Lena Reed, Cecilia Li, Angela Ramirez, Liren Wu, and Marilyn Walker. 2022. Jurassic is (almost) all you need: Few-shot meaning-to-text generation for open-domain dialogue. In *Conversational AI for Natural Human-Centric Interaction: 12th International Workshop on Spoken Dialogue System Technology, IWSDS 2021, Singapore*, pages 99–119. Springer.

V Sanh. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

María Inés Torres, Javier Mikel Olaso, César Montenegro, Roberto Santana, Alain Vázquez, Raquel Justo, José Antonio Lozano, Stephan Schlögl, Gérard Chollet, Nazim Dugan, et al. 2019. The EMPATHIC project: mid-term achievements. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 629–638.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alain Vázquez, Asier López Zorrilla, Javier Mikel Olaso, and María Inés Torres. 2023. Dialogue management and language generation for a robust conversational virtual coach: Validation and user study. *Sensors*, 23(3):1423.

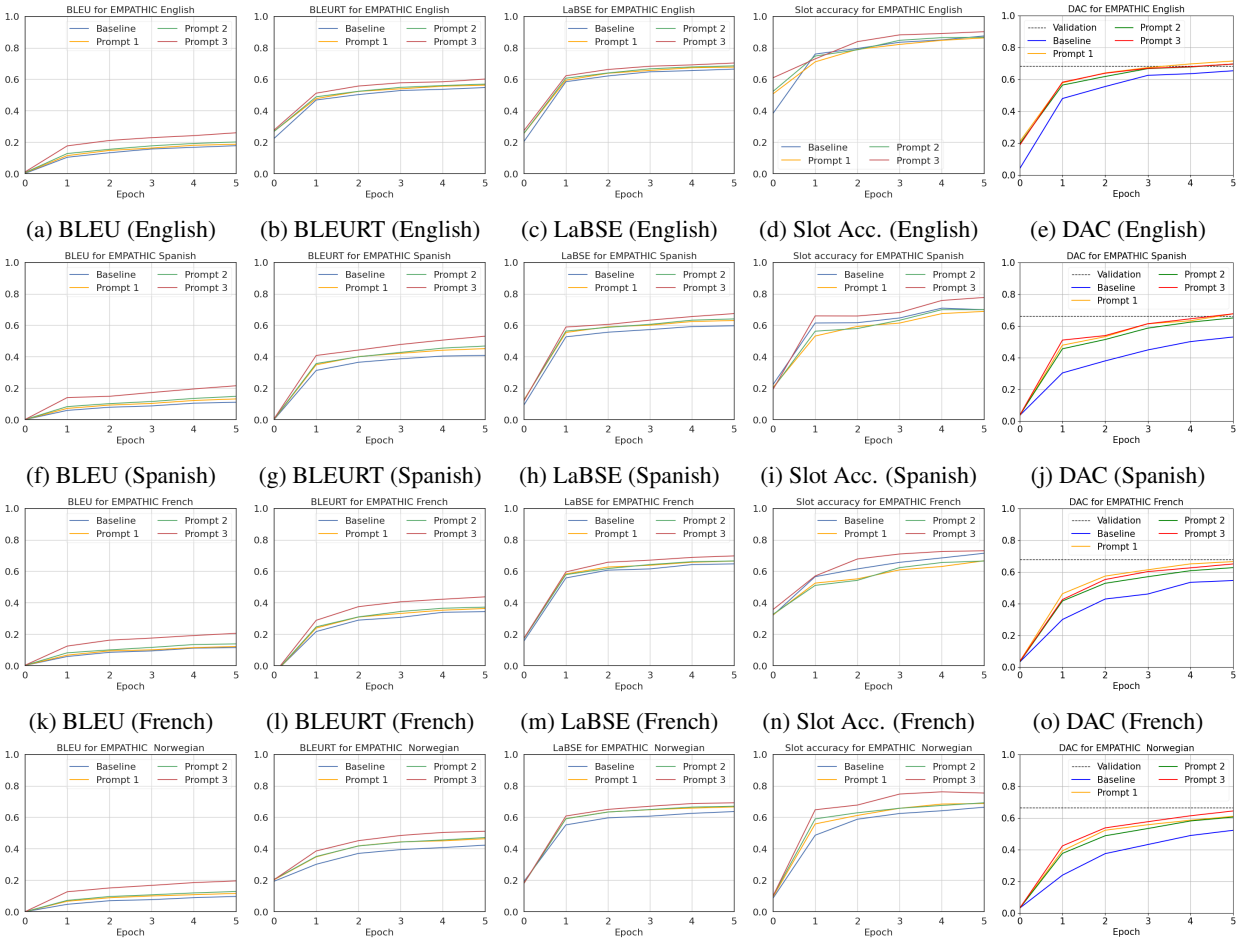
Alain Vazquez, Angela Maria Ramirez, Neha Pullabhotla, Nan Qiang, Ranran Haoran Zhang, Marilyn Walker, and Maria Inés Torres. 2024. Knowledge-grounded dialogue act transfer using prompt-based learning for controllable open-domain NLG. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 78–91.

Alain Vázquez and María Inés Torres. 2025. Impact of prompt-based meaning representations for language generation in dialogue tasks: A comprehensive exploration of the relevance of tasks, corpora and metrics. https://doi.org/10.5281/zenodo.15310835. Preprint.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, and Nikolaos Agadakos. 2024. Modeling low-resource health coaching dialogues via neuro-symbolic goal summarization and text-units-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11498–11509.

A Fine-tuning evolution



(a) BLEU (English) (b) BLEURT (English) (c) LaBSE (English) (d) Slot Acc. (English) (e) DAC (English)

(f) BLEU (Spanish) (g) BLEURT (Spanish) (h) LaBSE (Spanish) (i) Slot Acc. (Spanish) (j) DAC (Spanish)

(k) BLEU (French) (l) BLEURT (French) (m) LaBSE (French) (n) Slot Acc. (French) (o) DAC (French)

(p) BLEU (Norw.) (q) BLEURT (Norw.) (r) LaBSE (Norw.) (s) Slot Acc. (Norw.) (t) DAC (Norw.)

Figure 2: Evolution of the average score of each metric during the fine-tuning with each representation for each language. *Slot Acc.* stands for Slot Accuracy and *Norw.* for Norwegian. In the DAC results, there is a dotted line for Validation results.