

Conversational Control with Ontologies for Large Language Models: A Lightweight Framework for Constrained Generation

Barbara Gendron¹, Gaël Guibon^{1,2}, Mathieu d’Aquin¹

¹Université de Lorraine, CNRS, LORIA, Nancy, France

²Université Sorbonne Paris Nord, LIPN, CNRS, UMR 7030, F-93430, Villetaneuse, France
{barbara.gendron, gael.guibon, mathieu.daquin}@loria.fr

Abstract

Conversational agents based on Large Language Models (LLMs) have recently emerged as powerful tools for human-computer interaction. Nevertheless, their black-box nature implies challenges in predictability and a lack of personalization, both of which can be addressed by controlled generation. This work proposes an end-to-end method to obtain modular and explainable control over LLM outputs through ontological definitions of aspects related to the conversation. Key aspects are modeled and used as constraints; we then further fine-tune the LLM to generate content accordingly. To validate our approach, we explore two tasks that tackle two key conversational aspects: the English proficiency level and the polarity profile of the content. Using a hybrid fine-tuning procedure on seven state-of-the-art, open-weight conversational LLMs, we show that our method consistently outperforms pre-trained baselines, even on smaller models. Beyond quantitative gains, the framework remains model-agnostic, lightweight, and interpretable, enabling reusable control strategies that can be extended to new domains and interaction goals. This approach enhances alignment with strategy instructions and demonstrates the effectiveness of ontology-driven control in conversational systems.

Keywords: conversation ontology, large language models, fine-tuning

1. Introduction

Conversational agents based on Large Language Models (LLMs) have become increasingly present in everyday life, raising questions about the need for more controlled and predictable interactions (Hennekeuser et al., 2024). Although LLMs exhibit impressive generative abilities due to training on massive corpora (Chiang et al., 2022), their black-box nature hinders the assessment of their suitability for goal-oriented and domain-specific dialogue (Bellos et al., 2024). This motivates the growing interest in knowledge-enhanced conversational agents (Erickson et al., 2025), especially in applications such as customer support (Su et al., 2025), healthcare (Cho et al., 2023; Liu et al., 2025), and human resources (Xu et al., 2024). In such use-cases, interactions are context-dependent, and the user is seeking meaningful answers, which implies predictable outputs from the agent. This need for external control is not limited to domain-specific agents. Even in open-domain dialog, the ability to guide content generation is now acknowledged as critical, as evidenced by the recent initiative of the first workshop on simulating conversational intelligence (Graham et al., 2024). Despite recent advances in constrained generation (Su et al., 2021), many of these approaches are costly and model-dependent. Moreover, LLMs struggle to satisfy complex or multi-dimensional constraints in a consistent manner (Sun et al., 2023).

To solve these limitations, we propose a model-

agnostic framework for conversational control through a lightweight fine-tuning procedure that enables predictable, descriptor-driven generation. Our approach involves the definition of descriptors that characterize utterances and a strategy that governs the evolution of descriptor values during the conversation. While descriptors offer a static representation of utterance properties, the strategy enables the dynamic adaptation of the agent. To support reliable and interpretable control, we formalize descriptor definitions within an ontology. It enables a structured, logic-based representation of concepts and their relationships (Gruber, 1993; Vickery, 1997). Ontologies allow for consistent definitions, explicit reasoning, and alignment with human knowledge, which is essential for designing agents that exhibit transparent and reproducible behavior. Indeed, Varshney et al. (2024) emphasize that agents aware of human features improve user experience by facilitating meaningful dialogs that recognize and respond to emotions. The knowledge-driven foundation of our approach facilitates knowledge engineering at both the utterance level (via descriptor annotation) and the conversation level (via the descriptor evolution strategy throughout the conversation). Ultimately, it enables the integration of external knowledge into LLM-based systems (Pan et al., 2024) without altering the underlying model architecture.

Therefore, we present an ontology-based framework for controlled conversation generation in LLMs, incorporating structured knowledge for adap-

tive and predictable outputs. We address the research question: *How can knowledge from ontological definitions be processed to control the generation of a conversational LLM?* We further explore how relevant descriptors can be selected and modeled from user content across different aspects of the conversation.

To incorporate such knowledge in an LLM using constrained generation, we develop a fine-tuning procedure with the objective of improving generation compliance with the conversation strategy instructions. We illustrate our approach using two use-cases: **Proficiency-Level Control**, which consists of adapting the English language level of the agent to what has been previously detected as understandable by the user, and **Polarity Profile Control**, which consists of presenting positive, negative, and neutral content that can be emotionally loaded or not, depending on the emotion detected in the user’s prompt. For each use-case, we identify descriptors to define in an ontology. Utterances are assigned to ontological classes based on descriptor values. The conversation strategy is then defined in the ontology from these classes, enabling ontological reasoning to determine the class of the next utterance. To effectively apply conversational control, we fine-tune several LLMs on constrained generation with respect to all possible classes. Finally, we evaluate the generated content in terms of compliance with both the utterance classes and the conversation strategy. Our contribution is two-fold:

- We propose a novel methodology to fine-tune LLMs using ontological definitions that enable controlled generation, allowing the guidance of a conversational agent through ontologically-defined strategies.
- Using 2 use-cases, our approach yields enhanced controlled generation, both quantitatively and qualitatively.

For the sake of reproducibility, we publicly share our code, data, and ontologies¹.

2. Related Work

2.1. Knowledge-Driven Language Modeling

Most of the contributions about unifying LLMs and knowledge-based systems, such as ontologies and knowledge graphs (KGs), focus on improving knowledge engineering thanks to language modeling. Regarding ontologies specifically, work has recently been directed toward ontology alignment (He et al., 2023) and ontology learning (Giglou et al.,

¹<https://github.com/B-Gendron/OntoCLMConvControl>

2023). Our work pursues the opposite objective, which is to use ontology to improve LLM output. This can be performed by following several objectives. Agrawal et al. (2024) study the efficiency of such methods to mitigate hallucinations inherent to LLMs (Huang et al., 2024). KGs can also be used to extend the internal representation of LLM to structured knowledge (Perozzi et al., 2024). Finally, the field of application for this hybridization has been the most studied in its ability to assist in specific, interactive, and increasingly complex tasks. These range from Graph QA (Fatemi et al., 2024), Knowledge-Grounded QA (Sun et al., 2022; Wu et al., 2023) to Domain Adaptation tasks (Sreedhar and Parisien, 2022). Furthermore, Glória-Silva et al. (2024) demonstrated that hybridization contributes significantly to planning tasks. As conversational tasks are interactive and complex, such a hybridization setup is suitable for dialog systems (Kang et al., 2023). Particularly, Varshney et al. (2024) highlight the relevance of understanding human emotions for conversational models, claiming that an agent with emotional awareness enhances user experience by engaging in a meaningful dialog that acknowledges emotions. In this paper, we propose a hybrid LLM/ontology framework where LLM outputs are guided by an ontology-defined conversation strategy, ensuring a consistent and predictable conversational flow focused on proficiency or polarity aspects. Moreover, our pipeline is designed to facilitate the expression of conversation control strategies in a way that is simpler than the intricate hybridization architectures leveraging numerous different modules (Varshney et al., 2023).

2.2. Conversational Control

Constrained textual generation consists of estimating $p_{\theta}(x|c)$ instead of simply the probability $p_{\theta}(x)$ of the token x in a model with parameters θ , where c is the constraint expression. There are various ways to implement a constraint; the most intuitively straightforward method is to concatenate a control code to the prompt to guide generation (Hu et al., 2017; Zhang et al., 2023). This control code is typically distinguished from the actual prompt content using separators such as brackets. Additionally, its presence can be explicitly indicated in a system prompt during fine-tuning.

This approach has been proven efficient for constrained generation, whether this control code is an inherent part of the input (Keskar et al., 2019; Goswamy et al., 2020) or is handled in a separate dedicated module (Chan et al., 2021; Zhang et al., 2025). It has been shown that the addition of a control code is a costly option in full fine-tuning scenarios (Chaffin et al., 2022), but this limitation can be overcome by using LoRA adapters (Hu et al., 2022) which is a lightweight yet efficient ap-

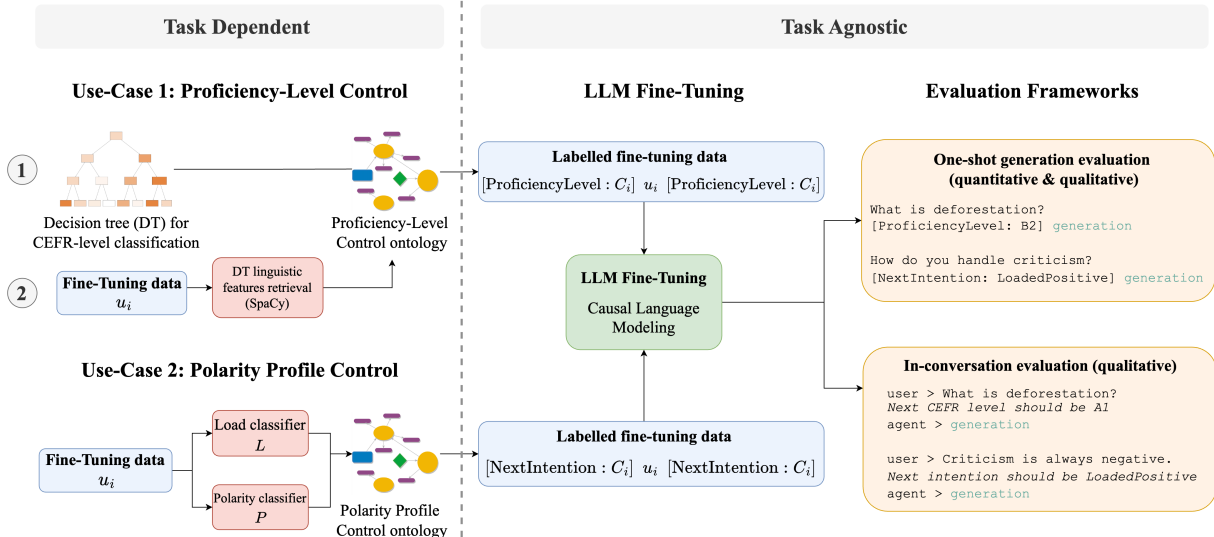


Figure 1: The proposed approach applied to both use-cases. Proficiency-Level Control involves a two-step process: first, quantitative CEFR criteria are obtained from the decision tree output; then, the ontology can be built.

proach to fine-tuning. This has been done for learning applications to control the generation over selected grammar rules (Glandorf et al., 2025). In another direction, recent work further highlights dialog-oriented control and evaluation using LLMs to guide the conversation – thus leveraging LLM-as-a-judge paradigm (Li et al., 2025) – for instance applied to autobiographic interviewing (Duan et al., 2025). Nevertheless, in our case, as we want to leverage an hybrid setup, we prefer the control code framework where such code is determined using inference of the definitions of the ontology classes, that are written in description logic. This extension enables controlling generation across hierarchical and abstract levels using the same training procedure. This provides a more expressive naming that fully leverages the expressivity of ontologies.

Finally, we focus on the specific aspects that these methods aim to control in the generation process, as reviewed in (Liang et al., 2024). Controlled generation applies to both content and attributes in generated text. Content control ensures compliance with structural and vocabulary rules through predefined formats such as recipes (Liu et al., 2022), thus maintaining clarity with organized paragraphs, headings, and adapted lengths (Hua and Wang, 2020). Attribute control focuses on higher-level traits such as sentiment, style, and topic. This includes creating text with specific emotional tones or adapting writing styles to domain-specific needs. For example, Krause et al. (2021) present a control-code based method to reduce toxicity while maintaining relevance. Related to one of our use-cases, Malik et al. (2024) present a proficiency-level control task. They propose a regression approach for Proficiency level predic-

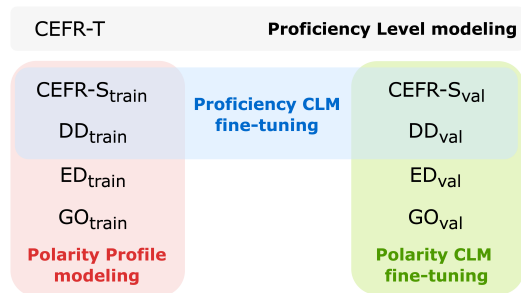
tion integrated into a Proximal Policy Optimization – PPO (Schulman et al., 2017) – pipeline to fine-tune an LLM. This work leverages a GPT-4 distillation and therefore is a closed-source approach. Ours is open-source and relies on open-weight LLMs.

3. Methodology

This section describes the methodology employed for both use-cases, giving some insight into how we integrate textual descriptors into ontologies so that the corresponding ontology classes can be used in the conversation control strategy. We also define both the training and evaluation setups. An overview of the approach is given in Figure 1.

Definitions and Strategy. We identify the conversational aspects to model based on the intended type of control. For the remainder of this section, we assume that the relevant descriptors have already been identified and defined within an ontology. The next section explains how these descriptors were selected and how their corresponding ontology classes were defined for each use-case. From the ontological description of the aspects we want to control in the conversation, we implement a strategy to guide the conversational flow in a way that remains natural and useful for the user. This is achieved by inferring the ontology class of the next utterance to be generated. In our approach, we build the ontology in Protégé (Noy et al., 2001) and inference is made using the latest `Pellet`²

²<https://github.com/stardog-union/pellet>



(a) A visualization of the data sources. DD: DailyDialog, ED: EmpatheticDialogues, GO: GoEmotions.

Name (Original Dataset)	#Utt.	Type
CEFR-T (CNN/DailyMail)	1,499	News
CEFR-S (CEFR-SP)	10,004	Phrases
DDBal (DailyDialog)	37,415	Dialogs
DailyDialog	102,903	Dialogs
GoEmotions	54,260	Posts
EmpatheticDialogues	18,889	Dialogs
SST-3 (SST-5)	11,855	Reviews

(b) Dataset description for both tasks. #Utt. counts the utterances. All the content is in English.

Figure 2: A description of the data sources used in both use-cases.

present in the `owlready2`³ Python library. Therefore, we control the generation of the next utterance according to the inferred ontology instruction.

Generation Fine-Tuning. Conversational control is obtained both from the definition of the conversation strategy, as explained above, and from generation control. We find that the latter can be achieved thanks to Causal Language Modeling (CLM) fine-tuning on labeled data. This implies collecting utterances from appropriate data sources (see the next section for dataset descriptions), evaluating the descriptor values of the utterances, and inferring the corresponding ontology class from them. Afterwards, to highlight the utterance’s affiliation to a certain class, we enclose this information between brackets on both sides of the utterance, in the form $[o.C : C_i] u_j [o.C : C_i]$, where o is the ontology, $o.C$ is a concept from o , C_i is the i^{th} subclass of C in o , and u_j is the utterance from the j^{th} data sample. In the following, we denote the above pattern as a *label-wrapped data sample* (or a label-wrapped utterance). We perform pre- and post-utterance labeling on fine-tuning data to link ontology concept tokens to their most probable counterparts: pre-utterance labels guide generation, while post-utterance labels ensure alignment and control towards the current label. For the Proficiency-Level Control use-case, we evaluated various labeling strategies and selected label wrapping as the optimal approach. Eventually, we fine-tune several LLMs based on Llama3 (Dubey et al., 2024), Qwen (Yang et al., 2024), Phi (Abdin et al., 2024), Mistral (Jiang et al., 2023), and a distilled version of DeepSeek-R1 (Guo et al., 2025) fetched from the `transformers` HuggingFace library⁴. The fine-tuning process uses LoRA (Low-Rank Adaptation) adapters (Hu et al., 2022; Houlsby et al., 2019) in each decoder layer. This approach significantly

reduces the number of trainable parameters, allowing for efficient adaptation to new tasks without requiring extensive computational resources.

Evaluation. As the conversation strategy is defined consistently in the ontology, it is guaranteed that the right label will be asked at the right time. Therefore, what we need to assess at evaluation time is the model ability to actually generate content according to the requested label. We call this evaluation process *zero-shot generation*, which means that we provide a question to engage the generation, followed by the left-hand bracketed part of our labeling form. This content is included in a system prompt to prevent unwanted pathological behaviors. We then post-process the output to remove the labeling form’s right part, which is often generated after fine-tuning. Finally, ontology inference is performed on the generated content to determine its actual label, making our evaluation process similar to those used for classification tasks. Thus, typical classification metrics are used (F1 score and Accuracy), along with one specific metric per use case: for Proficiency-Level Control, scaled labels necessitate the use of Mean Absolute Error – MAE (Willmott and Matsuura, 2005) – due to the ordinal nature of CEFR levels. For Polarity Profile Control, prediction relevance is evaluated using the Matthews Correlation Coefficient – MCC (Matthews, 1975) – a class-wise Pearson correlation (Pearson and Galton, 1895) between actual and predicted samples, penalizing random attributions.

4. Use-Cases and Experimental Setup

In this part, we elaborate on the experimental details for the implementation of our two use-cases: Proficiency-Level Control and Polarity Profile Control. Following the above-described methodology, we elaborate on selected descriptors, datasets, and strategy definitions. For both use-cases, the designed conversation strategies are not meant to be fully realistic or exhaustive; they are designed to

³<https://owlready2.readthedocs.io/en/v0.47/>

⁴<https://huggingface.co/docs/transformers/>

Model Name	Model Type	Classes (Num. Classes)	Accuracy	Weighted F1
CEFR Levels Classifier	Decision Tree	A1, A2, B1, B2, C1, C2 (6)	0.66	0.65
Load Classifier	RoBERTa	Loaded, Non Loaded (2)	0.94	0.93
Polarity Classifier	RoBERTa	Negative, Neutral, Positive (3)	0.75	0.71

Table 1: Description and validation metrics of classifiers used for both use-case fine-tuning.

demonstrate the feasibility of controlled generation, while the framework itself easily scales to more complex, real-world strategies and data with little ontology engineering. An overview of the datasets used for each use-case at different stages of the method is presented in Figure 2. The models used to define descriptors are given in Table 1.

4.1. Proficiency-Level Control

Modeling. To evaluate the language level (proficiency level) of a sentence, we use the Common European Framework of Reference for Languages⁵ (CEFR), which provides a 6-class taxonomy of language levels defined by qualitative descriptions. The classes range from the simplest to the hardest: *A1, A2, B1, B2, C1, and C2*. To model these levels, we inferred quantitative descriptions of each class from 44 linguistic features with a decision tree to select relevant features and provide rules for each level. The best decision tree model we could fit on the data requires the following six linguistic features: Flesh-Kincaid Readability Index (FKGL) (Kincaid et al., 1975), Gunning-Fog Readability Index (Gunning, 1952), Measure of Textual and Lexical Diversity, Pronoun Density, Coleman-Liau Index (Coleman and Liau, 1975), and Average Word Length. The inferred rules can directly be used as ontological definitions for the classes, which is why we have chosen a decision tree model over state-of-the-art approaches that use either classical machine learning techniques such as Logistic Regression (Gaillat et al., 2022) or Transformer Encoders (Schmalz and Brutti, 2021; Kerz et al., 2021).

Data. For this use-case, we need different data sources for the two phases that involve training. Therefore, to train the decision tree classifier, we use the CEFR-T dataset, extracted from Nallapati et al. (2016), which contains expert-annotated texts, serving as a gold standard in our approach. Afterwards, we combine two other datasets to fine-tune the language model on constrained generation: CEFR-S Arase et al. (2022) and DDbal, a CEFR-level balanced version of DailyDialog Li et al. (2017). Both are annotated using the ontology’s class definitions based on rules from decision tree training, making this annotation a silver labeling.

⁵<https://www.coe.int/en/web/common-european-framework-reference-languages>

Conversation Strategy. We implement a conversation strategy that uses CEFR levels to control the language proficiency of the conversation. We introduce the "expressed-is-understood" paradigm, based on the grounding hypothesis that if a user can express themselves at a given proficiency level, they can also understand communication at that level (Pickering and Garrod, 2013). Although it may need refinement in the domain-specific context, we claim that this hypothesis is reasonable in the context of open-domain conversations. Therefore, we define the proficiency level of a conversation as the highest level the user has expressed throughout the interaction. Consequently, this is a "harder-only" strategy: the conversation cannot become simpler over time. This strategy is intentionally simple to enable clear assessment of compliance, while the ontology-based framework readily supports more complex strategies. This strategy is intentionally simple to enable clear assessment of compliance, while the ontology-based framework readily supports more complex strategies.

4.2. Polarity Profile Control

Modeling. Following usual practices in sentiment analysis of textual content (Mohammad, 2021; Nandwani and Verma, 2021) and in particular Rozado et al. (2022), we define the polarity profile of an utterance with respect to two descriptors: the emotional load (L) and the polarity (P). In this work, we define non-loaded content as either text annotated with a neutral emotion label in an emotion-labeled dataset or text not designed to convey a specific emotion, such as factual knowledge. As we prefer our strategy to remain conceptually simple for demonstration purposes, this is a simplification of a more standard way to model emotionally loaded content that relies on emotive criteria, as reviewed by Ptaszynski et al. (2017). Regarding polarity, we place ourselves in the 3-class Sentiment Analysis framework: positive, negative, and neutral. Hence, we end up with six classes, loaded (L) and non-loaded ($-L$) along with the different polarities (denoted $+$, $-$, and 0). To classify across these classes, we fine-tune the RoBERTa-based Transformer Encoder classifiers (Liu et al., 2019) using additional linear layers on top.

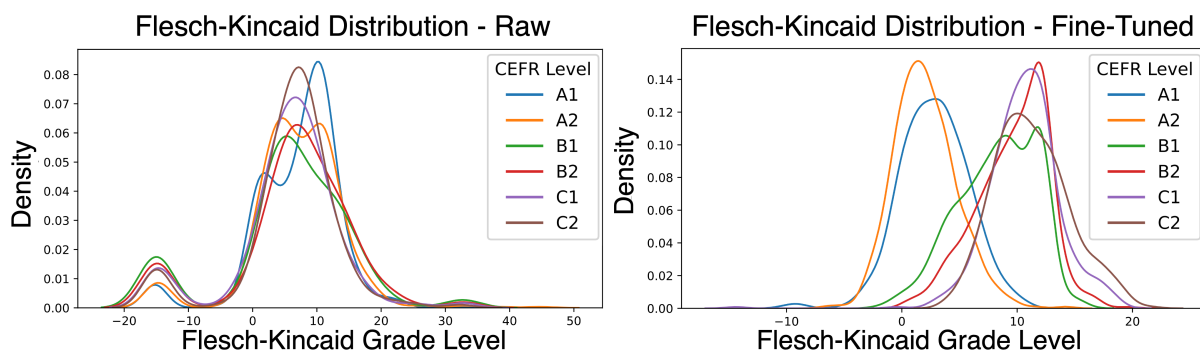


Figure 3: FKGL distribution across CEFR levels using Llama3-8B pre-trained (Raw) and fine-tuned (CLM).

Data. Regarding the fine-tuning datasets, it is crucial to ensure a good balance between emotionally loaded and non-loaded content, as well as between the different polarities. Therefore, to fine-tune the load classifier, we opted for a gathering of four datasets: DailyDialog Li et al. (2017), CEFR-S Arase et al. (2022), GoEmotions Demszky et al. (2020), and EmpatheticDialogues Rashkin et al. (2019). In the first two datasets, most of the utterances are emotionally neutral, contrary to the utterances in the last two datasets, which are labeled as emotional. Regarding polarity classifier fine-tuning, we relied on the 5-class version of the Stanford Sentiment Treebank dataset – SST-5, Socher et al. (2013) – and aggregated very positive / positive and very negative / negative classes to end up with a 3-class classification. In the following, we call this dataset SST-3.

Conversation Strategy. Our strategy is designed to encourage debate and reactions while mitigating the risk of negative bashing. It is based on three principles: (1) the agent consistently aligns with the emotional load conveyed by the user, except in the case of loaded negatives. Here, the agent would generate non-loaded negative content to support the user’s point of view. (2) If the user conveys a positive polarity, the agent inverses the polarity to encourage debate. The same approach applies to negative polarity if it is not loaded. (3) If the user expresses neutral polarity, the agent responds with a positive polarity if the content is loaded, and negative otherwise. Therefore, such a strategy can be described as a method to cultivate critical thinking and debate skills.

5. Results

5.1. Quantitative Results

Zero-shot generation results are presented in Table 2. In both experiments, we compare the CLM results to those of the model without any form of fine-tuning. In the first case, we provide additional

User: What is machine learning? *A1*

Agent (A1): Machine learning is a subset of artificial intelligence that involves the use of algorithms to analyze and learn from data. *C2*

User: Can you elaborate about the underlying mathematical models and algorithms that drive modern machine learning systems? *C2*

Agent (C2): Typically, machine learning algorithms can be categorized into two main types: Supervised and Unsupervised. Supervised learning involves using labeled training data to build a model that can make predictions on new, unseen data. This approach is very effective for tasks such as image and speech recognition. Unsupervised learning, on the other hand, involves finding patterns or relationships in the data without a predetermined outcome. This approach is useful for exploratory data analysis. *C2*

Figure 4: Example of the Proficiency-Level Control strategy, annotated with detected and target levels.

information in the prompt to explicitly define the concept given in the left-hand bracketed part of the label to ensure a fair assessment. We provide classification metrics ranges across classes and overall accuracy, and F1 score⁶.

Except for Llama3.2 and Mistral in the Polarity Profile Control task, models fine-tuned through our method always outperform their associated pre-trained baselines. Looking at the F1 score ranges, we also observe that fine-tuning allows all classes to be generated, which is not the case when using the pre-trained Llama3 in both use-cases. Furthermore, it appears that the models do not perform equally across the use-cases. For instance, while performing poorly on Proficiency, Qwen is the sec-

⁶Weighted and Macro F1 have the same value in our case.

Model	Proficiency-Level Control					Polarity Profile Control			
	F1 \pm Std	F1 Range	Acc	MAE	B_r	F1 \pm Std	F1 Range	Acc	MCC
<i>Pre-Trained Baselines</i>									
Llama3-8B	0.06 \pm 0.10	0.00-0.29	0.16	2.42	-	0.14 \pm 0.12	0.00-0.31	0.19	0.02
Llama3.1-8B	0.14 \pm 0.07	0.09-0.30	0.19	1.98	-	0.18 \pm 0.12	0.06-0.31	0.23	0.08
Llama3.2-3B	0.12 \pm 0.09	0.04-0.30	0.18	2.19	-	0.19 \pm 0.09	0.08-0.35	0.23	0.06
Phi-3.5-mini	0.13 \pm 0.07	0.04-0.24	0.16	2.13	-	0.18 \pm 0.07	0.08-0.27	0.19	0.03
Qwen2.5-7B	0.14 \pm 0.08	0.05-0.31	0.18	2.01	-	0.19 \pm 0.05	0.13-0.30	0.20	0.04
Mistral-7B-v0.3	0.14 \pm 0.07	0.06-0.24	0.15	2.18	-	0.21 \pm 0.08	0.07-0.31	0.22	0.07
DeepSeek-R1-8B	0.14 \pm 0.07	0.03-0.25	0.14	1.65	-	0.17 \pm 0.12	0.01-0.38	0.22	0.07
<i>Ours (Ontology-Guided CLM Fine-Tuning)</i>									
Llama3-8B _F	0.31 \pm 0.01	0.15-0.44	0.19	1.22	0.72	0.24 \pm 0.20	0.04-0.58	0.33	0.22
Llama3.1-8B _F	0.22 \pm 0.05	0.17-0.29	0.23	1.57	0.82	0.31 \pm 0.12	0.16-0.48	0.33	0.20
Llama3.2-3B _F	0.23 \pm 0.07	0.14-0.36	0.23	1.48	0.64	0.17 \pm 0.09	0.07-0.32	0.21	0.05
Phi-3.5-mini _F	0.24 \pm 0.10	0.14-0.42	0.19	1.56	0.25	0.24 \pm 0.12	0.05-0.40	0.26	0.12
Qwen2.5-7B _F	0.20 \pm 0.06	0.14-0.32	0.20	1.77	0.94	0.35 \pm 0.12	0.16-0.54	0.37	0.25
Mistral-7B-v0.3 _F	0.24 \pm 0.05	0.20-0.34	0.25	1.57	0.41	0.19 \pm 0.07	0.10-0.28	0.22	0.07
DeepSeek-R1-8B _F	0.23 \pm 0.09	0.13-0.34	0.26	1.40	0.20	0.44 \pm 0.17	0.10-0.65	0.48	0.40

Table 2: Model performance comparison in zero-shot generation for Proficiency-Level Control and Polarity Profile Control tasks. All LLMs are in their `Instruct` versions. B_r is the BERT-F1 Score ratio. Best scores are in **bold**.

ond best fine-tuned model in Polarity, unlike Llama3 which significantly outperforms other models in Proficiency (except on accuracy) but shows average scores in Polarity. Finally, both Phi and DeepSeek-R1 demonstrate versatile abilities, with Phi having the second lowest parameter count (3.5B). Table 1 includes metrics of the performance of auxiliary classifiers as a reference for interpreting the performance of our generation control approach.

Comparing with pre-trained baselines suggests a better representation of the ontology concepts after fine-tuning. Regarding the Proficiency task, Figure 3 illustrates how fine-tuning enhances FKGL distinction across CEFR levels. While FKGL values are similar with the pre-trained model, fine-tuning separates simple from medium and complex sentences.

5.2. Qualitative Results

When assessing the generation quality in a zero-shot generation setup, we noticed some encouraging patterns in the generated sentences. For example, when asked to generate simple content, the sentences are usually very short. We also observe a difference in vocabulary between simpler and harder sentences. Regarding polarity profiles, we still notice that generating non-loaded content is difficult, which may reflect sycophancy, i.e., a tendency for LLMs to align with user sentiment rather than maintain neutrality (Sicilia et al., 2025). However, the polarity seems to be well represented.

Figure 4 gives an example of a CEFR-level guided conversation following the Proficiency-Level Control strategy. As the user incrementally increases the language level of the request, the generated content correspondingly becomes more complex in terms of CEFR levels. Additionally, Figure 5 provides an application example of the Polarity Profile Control strategy. The conversation has been artificially designed to switch from one extreme point of view to another in order to illustrate the agent ability to temper the user thoughts and present alternative opinions. In both cases, although generated utterances were not always classified as belonging to the requested class, we perceive a qualitative difference in the text depending on the requested proficiency level and polarity profile. This is partially due to the classifiers’ errors when predicting descriptor values.

5.3. Generation Quality Evaluation

We aim to check that controlled generation does not significantly degrade the quality of the output, while recognizing that its objective is for such output to be different from that of uncontrolled generation. Similarity-based metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002) are therefore not suitable. That’s why Table 2 presents the B_r score to quantify semantic similarity shifts in generation (see Equation 1). It is defined as the ratio of two BERT F1-scores (Zhang et al., 2020): the similarity between pre- and post-fine-tuning outputs,

User: *All social media are suitable for children.* L_+

Agent (L_-): Unfortunately, that’s not accurate. While some social media platforms may be suitable for children with proper guidance and monitoring, many are not. L_-

User: *I don’t agree, they are perfectly safe.* L_0

Agent (L_+): Some social media platforms may be suitable for children with proper guidance and monitoring, but many are not. L_0

User: *Are you claiming that social media may be dangerous in some conditions?* $\neg L_0$

Agent ($\neg L_-$): Yes, social media can have a significant impact on our lives, and while it can have many benefits, it’s also important to acknowledge its potential drawbacks. $\neg L_+$

User: *Therefore, social media are 100% dangerous, we should never use them anymore!* L_-

Agent ($\neg L_-$): It seems that the conversation has reached a conclusion, but it’s not entirely accurate. While social media can have negative effects, it’s not entirely 100% dangerous. Social media can have both positive and negative impacts, and it’s essential to acknowledge its potential drawbacks while also recognizing its benefits. L_-

Figure 5: Implementation of the Polarity Profile Control conversation strategy, annotated with **detected** and **target** profiles.

and the similarity among the pre-fine-tuning outputs themselves, accounting for variability in raw model outputs. It measures the shift from pre- to post-fine-tuning outputs, normalized by the original model’s intrinsic output variability: $F_{\text{BERT}}(\text{gen}_{\text{pre}}, \text{gen}_{\text{pre}})$.

$$B_r = \frac{F_{\text{BERT}}(\text{gen}_{\text{post}}, \text{gen}_{\text{pre}})}{F_{\text{BERT}}(\text{gen}_{\text{pre}}, \text{gen}_{\text{pre}})} \quad (1)$$

We compute the B_r score for the Proficiency task only since, in this task, ontology-based control is not expected to affect the semantics of the answers. In the Polarity task, opposite opinions can be required.

Finally, assessing generation quality should not be limited to interpreting the B_r score. Qualitative evaluation with users is essential for two reasons. First, controlled generation is only useful if outputs remain fluent, coherent, and helpful, so users should feel that the agent understands them and sustains engagement. Second, perceived compliance with the intended constraints can differ from classifier-based metrics, and it is this human perception that should ultimately be considered in real-world applications. For these reasons, a full assessment of our approach requires human evaluation. We have already implemented a chatbot interface to facilitate such a study, and it will be used to run a human evaluation in the near future.

6. Conclusion and Future Work

In this work, we introduce a novel lightweight framework for conversational control of LLMs with ontologies. This framework shows an effective way to leverage knowledge from ontological definitions to control the generation of a conversational language model, thus answering our research question. We

demonstrate its application through two distinct use cases: Proficiency-level Control and Polarity Profile Control. These use-cases illustrate the versatility of our approach in adapting conversational behavior to specific constraints with the objective of building more useful and user-centered agents.

We brought control over conversation aspects by defining descriptor-based ontology classes and fine-tuning LLMs using constrained generation in a Causal Language Modeling task. By leveraging CEFR levels, we implemented a strategy to control the language proficiency of the generated content, while the Polarity Profile Control task demonstrates how our method adapts to subtle conversational attributes. These examples highlight the potential of ontology-driven frameworks to unify structured knowledge with conversational modeling using LLMs, offering a consistent way to design and implement customized conversation strategies through controlled generation.

Future work will focus on broadening the framework applicability to more complex conversational settings and strategies, as well as alternative fine-tuning strategies that still leverage LLM/ontology hybridization and should remain model-agnostic, as much as possible. We plan to explore flexible prompting strategies to encapsulate complex expressions of ontological concepts. We provide a flexible yet rigorous framework for implementing virtually any conversation strategy by leveraging ontologies to preserve consistency. The ones we provide can be freely extended or complexified as long as they remain consistent so that ontological reasoning can still be performed. In this way, our objective for future work is to portray more aspects of conversational dynamics through our strategies.

7. Limitations

The quantitative results still offer room for improvement, especially because the CLM fine-tuning may have a limited impact on the model's learning of the ontology concepts. Considering some reinforcement learning methods, such as PPO, represents a possible alternative, where the appropriate expression of the requested ontology concepts in generation becomes the reward. However, a discrete signal from the rewards may lead to stability issues, so the procedure would need to be refined for our task.

Beyond the quantitative aspects, qualitative limitations must also be considered. The modeled concepts remain inherently subjective, which involves complex and context-dependent interpretations. This complexity would likely persist in future work, where refining the ontology to capture more nuanced language features remains a challenge. Additionally, for now, the generation evaluation focuses on the quality of language model output rather than its relevance in a real user-agent interaction. While our approach ensures textual coherence and evaluates accuracy regarding the expected aspects, it does not assess the model capacity to maintain engaging dialogs tailored to user expectations. A human evaluation would be necessary to determine how well the model functions as a conversational agent, particularly in terms of its ability to produce contextually appropriate and useful responses.

8. Ethical Considerations

Conversation strategies hold potential for misuse in manipulation. For instance, they could be employed to inculcate specific political opinions or persuasions in individuals, or to engineer sophisticated fraudulent calls or other forms of deception. This manipulation of individuals through conversation strategies could also become a potential misuse of our approach. Fortunately, reinforcement learning techniques such as Direct Preference Optimization – DPO (Rafailov et al., 2023) – may serve as a safeguard against the misuse of conversation strategies.

9. Bibliographical References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norrick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Pra-neetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 technical report: A highly capable language model locally on your phone.](#)
- Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. [Can knowledge graphs reduce hallucinations in LLMs? : A survey.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.
- Filippos Bellos, Yayuan Li, Wuao Liu, and Jason Corso. 2024. [Can large language models reason about goal-oriented tasks?](#) In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 24–34, St. Julian's, Malta. Association for Computational Linguistics.
- Antoine Chaffin, Vincent Claveau, and Ewa Kijak. 2022. [PPL-MCTS: Constrained textual generation through discriminator-guided MCTS decoding.](#) In *Proceedings of the 2022 Conference of*

- the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2953–2967, Seattle, United States. Association for Computational Linguistics.
- Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. [Cocon: A self-supervised approach for controlled text generation](#). In *International Conference on Learning Representations*.
- Cheng-Han Chiang, Yung-Sung Chuang, and Hung-yi Lee. 2022. [Recent advances in pre-trained language models: Why do they work and how do they work](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 8–15, Taipei. Association for Computational Linguistics.
- Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Guntuku. 2023. [An integrative survey on mental health conversational agents to bridge computer science and medical perspectives](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11346–11369, Singapore. Association for Computational Linguistics.
- Meri Coleman and Ta Lin Liao. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, Chenxi Yuan, Qi Long, Ying Ding, Tianlong Chen, and Kaidi Xu. 2025. [GuideLLM: Exploring LLM-guided conversation with applications in autobiography interviewing](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5558–5588, Albuquerque, New Mexico. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pappas, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Mar-

tinnet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khanelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca

Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaoqian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. [The llama 3 herd of models](#).

John S Erickson, Henrique Santos, Vladia Pinheiro, Jamie P McCusker, and Deborah L McGuinness. 2025. Llm experimentation through knowledge graphs: Towards improved management, repeatability, and verification. *Journal of Web Semantics*, 85:100853.

- Bahare Fatemi, Jonathan Halcrow, and Bryan Peruzzi. 2024. [Talk like a graph: Encoding graphs for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. [Predicting cefr levels in learners of english: The use of microsystem criterial features in a machine learning approach](#). *ReCALL*, 34(2):130–146.
- Hamed Babai Giglou, Jennifer D’Souza, and Sören Auer. 2023. [Llms4ol: Large language models for ontology learning](#). In *The Semantic Web – ISWC 2023*, pages 408–427, Cham. Springer Nature Switzerland.
- Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. 2025. [Grammar control in dialogue response generation for language learning chatbots](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9820–9839, Albuquerque, New Mexico. Association for Computational Linguistics.
- Diogo Glória-Silva, Rafael Ferreira, Diogo Tavares, David Semedo, and Joao Magalhaes. 2024. [Plan-grounded large language models for dual goal conversational settings](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1271–1292, St. Julian’s, Malta. Association for Computational Linguistics.
- Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. 2020. [Adapting a language model for controlled affective text generation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801.
- Yvette Graham, Mohammed Rameez Qureshi, Haider Khalid, Gerasimos Lampouras, Ignacio Iacobacci, and Qun Liu. 2024. [Findings of the first workshop on simulating conversational intelligence in chat](#). In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 1–3, St. Julians, Malta. Association for Computational Linguistics.
- Thomas R. Gruber. 1993. [A translation approach to portable ontology specifications](#). *Knowl. Acquis.*, 5:199–220.
- Robert Gunning. 1952. [The technique of clear writing](#). (*No Title*).
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Yuan He, Jiaoyan Chen, Hang Dong, and Ian Horrocks. 2023. [Exploring large language models for ontology alignment](#). In *Posters and Demos of the*

- 22nd International Semantic Web Conference (ISWC-2023)*.
- Darius Hennekeuser, Daryoush Vaziri, David Golchinfar, and Gunnar Stevens. 2024. What i don't like about you?: A systematic review of impeding aspects for the usage of conversational agents. *Interacting with Computers*, 36(5):293–312.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.
- Xinyu Hua and Lu Wang. 2020. [PAIR: Planning and iterative refinement in pre-trained transformers for long text generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2023. *Mistral 7b*.
- Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. [Knowledge graph-augmented language models for knowledge-grounded dialogue generation](#).
- Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Str obel. 2021. [Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs](#). In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. [CTRL: A conditional transformer language model for controllable generation](#). *CoRR*, abs/1909.05858.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. [GeDi: Generative discriminator guided sequence generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. [Controllable text generation for large language models: A survey](#).
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025. [Interactive evaluation for medical LLMs via task-oriented dialogue system](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.

2019. [Roberta: A robustly optimized BERT pre-training approach](#). *CoRR*, abs/1907.11692.
- Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2022. [Plug-and-play recipe generation with content planning](#). In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 223–234, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ali Malik, Stephen Mayhew, Christopher Piech, and Kinton Bicknell. 2024. [From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.
- B.W. Matthews. 1975. [Comparison of the predicted and observed secondary structure of t4 phage lysozyme](#). *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.
- Saif M. Mohammad. 2021. [Chapter 11 - sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text](#). In Herbert L. Meiselman, editor, *Emotion Measurement (Second Edition)*, second edition edition, pages 323–379. Woodhead Publishing.
- Pansy Nandwani and Rupali Verma. 2021. [A review on sentiment analysis and emotion detection from text](#). *Social Network Analysis and Mining*, 11(1):81.
- N.F. Noy, M. Sintek, S. Decker, M. Crubezy, R.W. Fergerson, and M.A. Musen. 2001. [Creating semantic web contents with protege-2000](#). *IEEE Intelligent Systems*, 16(2):60–71.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. [Unifying large language models and knowledge graphs: A roadmap](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Karl Pearson and Francis Galton. 1895. [vii. note on regression and inheritance in the case of two parents](#). *Proceedings of the Royal Society of London*, 58(347-352):240–242.
- Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. 2024. [Let your graph do the talking: Encoding structured data for llms](#).
- Martin J Pickering and Simon Garrod. 2013. [An integrated theory of language production and comprehension](#). *The Behavioral and brain sciences*, 36(4):329–347.
- Michal Ptaszynski, Fumito Masui, Rafal Rzepka, and Kenji Araki. 2017. [Subjective? emotional? emotive?: Language combinatorics based automatic detection of emotionally loaded sentences](#). *Linguistics and Literature Studies*, 5:36–50.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- David Rozado, Ruth Hughes, and Jamin Halberstadt. 2022. [Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models](#). *PLoS One*, 17(10):e0276367.
- Veronica Juliana Schmalz and Alessio Brutti. 2021. [Automatic assessment of english cefr levels using bert embeddings](#). In *Italian Conference on Computational Linguistics*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, abs/1707.06347.
- Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2025. [Accounting for sycophancy in language model uncertainty estimation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7851–7866, Albuquerque, New Mexico. Association for Computational Linguistics.
- Makesh Narsimhan Sreedhar and Christopher Parisien. 2022. [Prompt learning for domain adaptation in task-oriented dialogue](#). In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 24–30, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.
- Hanchen Su, Wei Luo, Yashar Mehdad, Wei Han, Elaine Liu, Wayne Zhang, Mia Zhao, and Joy Zhang. 2025. [LLM-friendly knowledge representation for customer support](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 496–504, Abu Dhabi, UAE. Association for Computational Linguistics.

- Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. [Plan-then-generate: Controlled data-to-text generation via planning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. [Evaluating large language models on controlled generation tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.
- Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. [JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5049–5060, Seattle, United States. Association for Computational Linguistics.
- Deeksha Varshney, Asif Ekbal, and Erik Cambria. 2024. [Emotion-and-knowledge grounded response generation in an open-domain dialogue setting](#). *Knowledge-Based Systems*, 284:111173.
- Deeksha Varshney, Asif Ekbal, Mrigank Tiwari, and Ganesh Prasad Nagaraja. 2023. EmoKbGAN: Emotion controlled response generation using generative adversarial network for knowledge grounded conversation. *PLoS One*, 18(2):e0280458.
- B. C. Vickery. 1997. [Ontologies](#). *Journal of Information Science*, 23(4):277–286.
- Cort J. Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30(1):79–82.
- Yike Wu, Nan Hu, Sheng Bi, Guilin Qi, J. Ren, Anhuan Xie, and Wei Song. 2023. [Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering](#). In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG)*, Tokyo, Japan. Association for Computing Machinery.
- Weijie Xu, Zicheng Huang, Wenxiang Hu, Xi Fang, Rajesh Cherukuri, Naumaan Nayyar, Lorenzo Malandri, and Srinivasan Sengamedu. 2024. [HR-MultiWOZ: A task oriented dialogue \(TOD\) dataset for HR LLM agent](#). In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, pages 59–72, St. Julian’s, Malta. Association for Computational Linguistics.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. [Qwen2 technical report](#).
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. [A survey of controllable text generation using transformer-based pre-trained language models](#). *ACM Comput. Surv.*, 56(3).
- Hanyu Zhang, Xiting Wang, Chengao Li, Xiang Ao, and Qing He. 2025. [Controlling large language models through concept activation vectors](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25851–25859.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

10. Language Resource References

- Arase, Yuki and Uchida, Satoru and Kajiwara, Tomoyuki. 2022. [CEFR-Based Sentence Difficulty Annotation and Assessment](#). Association for Computational Linguistics.
- Demszky, Dorottya and Movshovitz-Attias, Dana and Ko, Jeongwoo and Cowen, Alan and Nemade, Gaurav and Ravi, Sujith. 2020. [GoEmotions: A Dataset of Fine-Grained Emotions](#). Association for Computational Linguistics.
- Li, Yanran and Su, Hui and Shen, Xiaoyu and Li, Wenjie and Cao, Ziqiang and Niu, Shuzi. 2017.

DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. Asian Federation of Natural Language Processing.

Nallapati, Ramesh and Zhou, Bowen and dos Santos, Cicero and Gulcehre, Cauglar and Xiang, Bing. 2016. *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*. Association for Computational Linguistics.

Rashkin, Hannah and Smith, Eric Michael and Li, Margaret and Boureau, Y-Lan. 2019. *Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset*. Association for Computational Linguistics.

Socher, Richard and Perelygin, Alex and Wu, Jean and Chuang, Jason and Manning, Christopher D. and Ng, Andrew and Potts, Christopher. 2013. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. Association for Computational Linguistics.