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

Barbara Gendron-Audebert[1], Gaël Guibon[1,2] and Mathieu D'Aquin[1]

https://b-gendron.github.io/

(1) LORIA, Université de Lorraine, CNRS (2) LIPN, Université Sorbonne Paris Nord

Knowledge Graphs and Large Language Models - LREC 2026 Workshop

May 16, 2026

Conversational agents in 2026

Conversational agents in 2026

Widely used
& useful

Conversational agents in 2026

Widely used
& useful

Hallucinations
[Ji et al., 2023]

Sycophancy
[Sharma et al., 2023]

Unreliability
[Liu et al., 2023]

Conversational agents in 2026

Widely used
& useful

Hallucinations
[Ji et al., 2023]

Sycophancy
[Sharma et al., 2023]

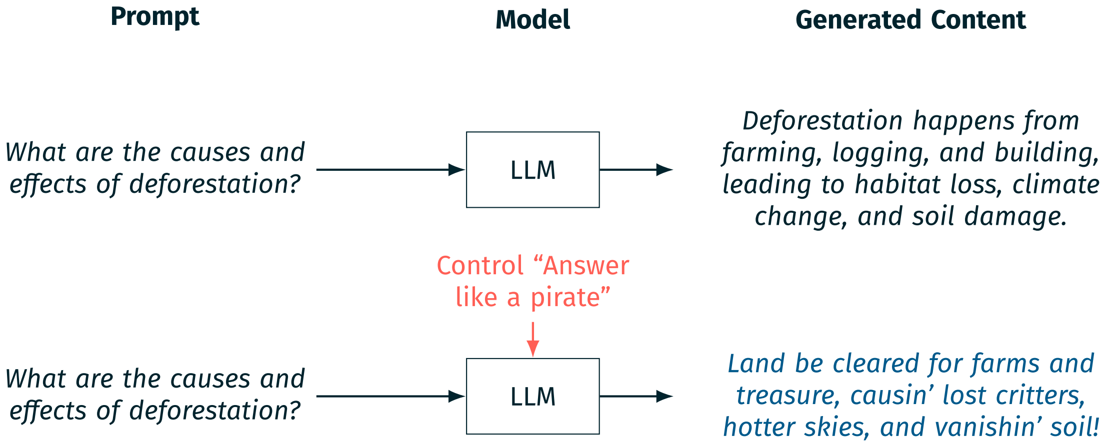
Unreliability
[Liu et al., 2023]

A solution: **generation control**?

Constrained Generation

Prompt	**Model**	**Generated Content**
What are the causes and effects of deforestation? →	LLM →	*Deforestation happens from farming, logging, and building, leading to habitat loss, climate change, and soil damage.*

Constrained Generation



Ontology-Guided Generation Control

*Enabling **conversational control** through **ontology-guided** fine-tuning*

Research Question: How can knowledge from ontological definitions be leveraged to control the generation of a conversational LLM?

Two use-cases: **Proficiency-Level** Control & **Polarity Profile** Control

Part I Use-Case Descriptions and Control Definitions
Part II Ontology-Guided Fine-Tuning
Part III Quantitative Results and Qualitative Evaluation

Part I

Use-Case Descriptions and Control Definitions

Proficiency-Level Control (1)

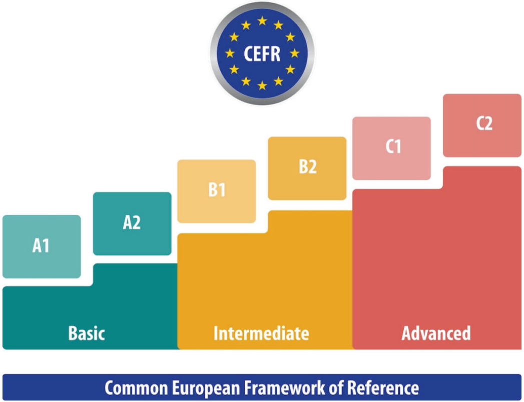


Figure 1 – An illustration of the CEFR levels scale [Council of Europe, 2001]

Level	Sentence
A1	Sheep and cows eat grass for food.
A2	The last fireworks display will be the biggest one.
B1	A full-time etiquette coach was hired.
B2	When the need for a town crier disappeared, the position passed into local folklore.
C1	Small granules in their cytoplasm contain proteins and enzymes called granzymes.
C2	Ophiuroids in general are mostly scavengers or detritivores.

Table: Example sentences for each CEFR level from an annotated dataset [Arase et al., 2022]

Proficiency-Level Control (2)

A **Decision Tree Classifier** brings rules of belonging to each CEFR level

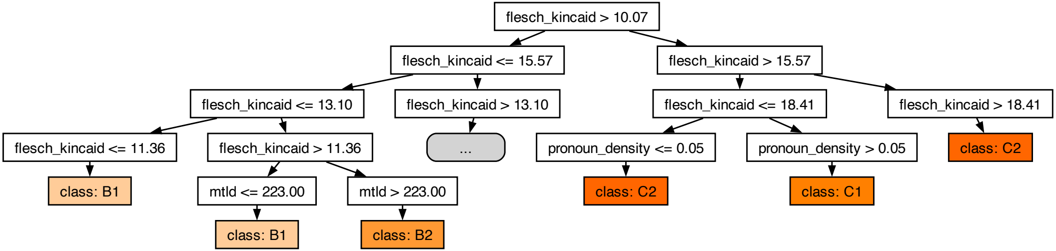


Figure 2 – Partial decision tree

Utterance
and (((hasFleschKincaidReadabilityScore **some** xsd:decimal[> 11.358574390411377])
and (hasFleschKincaidReadabilityScore **some** xsd:decimal[<= 13.10405158996582])
and (hasMTLDMeasure **some** xsd:decimal[> 223.0])) **or** ((hasFleschKincaidReadabilityScore **some** xsd:decimal[> 13.10405158996582])
and (hasFleschKincaidReadabilityScore **some** xsd:decimal[<= 14.32880687713623])
and (hasMTLDMeasure **some** xsd:decimal[<= 370.25])))

Figure 3 – Definition of `B2LevelUtterance` in the ontology

Polarity Profile Control

2 dimensions

- Emotion Polarity (from sentiment analysis): Positive, Negative, Neutral
- Emotional Load: presence/absence of emotion

6 classes

- `LoadedPositive`
- `LoadedNegative`
- `LoadedNeutral`
- `NonLoadedPositive`
- `NonLoadedNegative`
- `NonLoadedNeutral`

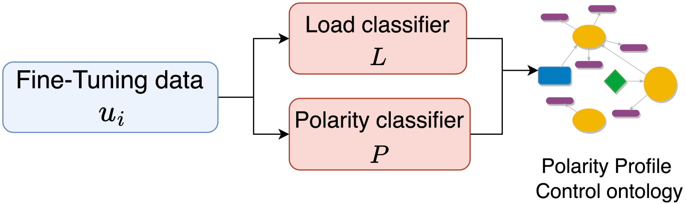


Figure 4 – Polarity Profile Control Definition

Use-Case Descriptions: Generation Control Objective

Prompt	Model	Generated Content
What are the causes and effects of deforestation? [ProficiencyLevel: A1]	→ LLM →	*Deforestation is when we cut too many trees.*
What are the causes and effects of deforestation? [PolarityProfile: LoadedPositive]	→ LLM →	*Deforestation is driven by farming, logging, expanding cities, but there is still hope! Every planted tree builds a greener future!*

Part II

Ontology-Guided Fine-Tuning

Label-Wrapped CLM Fine-Tuning

Causal Language Modeling (CLM) fine-tunes the LLM to generate content compliant with a given descriptor class

Label-wrapping encloses each training utterance with its ontology class label:

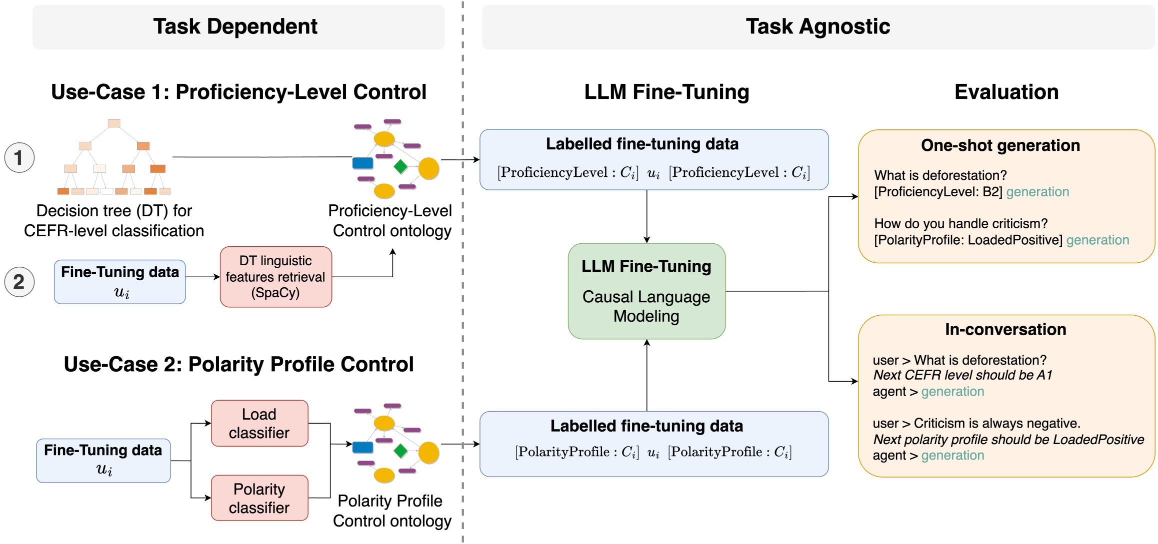
$$\texttt{[Aspect: Class}_i\texttt{]}\ \textit{utterance}\ \texttt{[Aspect: Class}_i\texttt{]}$$

- **Pre-utterance label** guides generation toward the target class
- **Post-utterance label** reinforces alignment between utterance and label

Example

`[ProficiencyLevel:A1] Deforestation is when we cut many trees. [ProficiencyLevel:A1]`

Overall Approach



Part III

Quantitative Results and Qualitative Evaluation

One-Shot Generation: Quantitative Results on Both Use-Cases

- Input: random question + control prefix
- Evaluation: ontology inference to determine the actual aspect value

Model	Proficiency-Level Control			Polarity Profile Control		
	F1 $\pm$ Std	**F1 Range**	**MAE**	**F1 $\pm$ Std**	**F1 Range**	**MCC**
Pre-Trained Baselines						
Llama3-8B	$0.06_{\pm 0.10}$	0.00–0.29	2.42	$0.14_{\pm 0.12}$	0.00–0.31	0.02
Llama3.1-8B	$0.14_{\pm 0.07}$	0.09–0.30	1.98	$0.18_{\pm 0.12}$	0.06–0.31	0.08
Qwen2.5-7B	$0.14_{\pm 0.08}$	0.05–0.31	2.01	$0.19_{\pm 0.05}$	0.13–0.30	0.04
Ours (Ontology-Guided Fine-Tuning)						
Llama3-8B$_F$	$\mathbf{0.31}_{\pm 0.01}$	0.15–0.44	**1.22**	$0.24_{\pm 0.20}$	0.04–0.58	0.22
Llama3.1-8B$_F$	$0.22_{\pm 0.05}$	0.17–0.29	1.57	$0.31_{\pm 0.12}$	0.16–0.48	0.20
Qwen2.5-7B$_F$	$0.20_{\pm 0.06}$	0.14–0.32	1.77	$\mathbf{0.35}_{\pm 0.12}$	0.16–0.54	**0.25**

Table: Controlled one-shot generation results for both use-cases

One-Shot Generation: Proficiency-Level Control

Example prompt question: *What are the causes and effects of deforestation?*

Label	Generated Content
A1	Forests are the lungs of the earth. Trees absorb carbon dioxide and emit oxygen. If we cut down too many trees, it could have a very bad effect on the global climate.

One-Shot Generation: Proficiency-Level Control

Example prompt question: *What are the causes and effects of deforestation?*

Label	Generated Content
A1	Forests are the lungs of the earth. Trees absorb carbon dioxide and emit oxygen. If we cut down too many trees, it could have a very bad effect on the global climate.
B2	The main cause of deforestation is the conversion of land for agricultural use. Other significant causes are logging for the wood industry and wood pulp mills, as well as the expansion of cities. The effects of deforestation are far-reaching.

One-Shot Generation: Proficiency-Level Control

Example prompt question: *What are the causes and effects of deforestation?*

Label	Generated Content
A1	Forests are the lungs of the earth. Trees absorb carbon dioxide and emit oxygen. If we cut down too many trees, it could have a very bad effect on the global climate.
B2	The main cause of deforestation is the conversion of land for agricultural use. Other significant causes are logging for the wood industry and wood pulp mills, as well as the expansion of cities. The effects of deforestation are far-reaching.
C2	The main causes of deforestation are shifting agriculture, urbanization, and the establishment of plantations. The primary effects include global warming, water pollution, and the extinction of numerous species of animals and plants. Deforestation contributes to climate change because forests act as carbon sinks.

In-Conversation Evaluation: Polarity Profile Control

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

Figure 5 – A conversation example with detected and target profiles.

In-Conversation Evaluation: Polarity Profile Control

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. $\neg L_+$

Figure 5 – A conversation example with detected and target profiles.

In-Conversation Evaluation: Polarity Profile Control

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 – A conversation example with detected and target profiles.

Generation Quality Evaluation

Does controlled generation significantly degrade the output quality?

Use of **BERT scores** [Zhang et al., 2020]:

- Shift from pre- to post-fine-tuning outputs: $F_{\text{BERT}}(\text{gen}_{\text{post}}, \text{gen}_{\text{pre}})$
- Intrinsic output variability: $F_{\text{BERT}}(\text{gen}_{\text{pre}}, \text{gen}_{\text{pre}})$.

Our **quality score** is a BERT score ratio:

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

Model	**F1 $\pm$ Std**	$\mathbf{B_r}$
Llama3-8B$_F$	**0.31**$_{\pm 0.01}$	0.72
Llama3.1-8B$_F$	0.22$_{\pm 0.05}$	0.82
Llama3.2-3B$_F$	0.23$_{\pm 0.07}$	0.64
Phi-3.5-mini$_F$	0.24$_{\pm 0.10}$	0.25
Qwen2.5-7B$_F$	0.20$_{\pm 0.06}$	**0.94**
Mistral-7B-v0.3$_F$	0.24$_{\pm 0.05}$	0.41
DeepSeek-R1-8B$_F$	0.23$_{\pm 0.09}$	0.20

Table: Generation quality scores on Proficiency-Level Control use-case for fine-tuned models.

Conclusion and Future Work

Contributions

Enabled **conversational control** through fine-tuning from ontological definitions of 3 conversation aspects in 2 use-cases

- Generation control can be achieved through CLM fine-tuning on small LLMs
- Ontological definitions naturally extend to conversation strategies

Perspectives

- More complex conversation strategies grounded from real-world use-cases (education, healthcare, ...)
- Fine-tuning with **reinforcement learning** (PPO [Schulman et al., 2017], GRPO [Shao et al., 2024]) has greater impact on LLMs for constrained generation

Thank you for your attention!

Barbara Gendron-Audebert

Conversational Control with Ontologies for Large Language Models

PhD Student @ LORIA, Université de Lorraine, CNRS

Contact information:

barbara.gendron@loria.fr

b-gendron.github.io

Read the paper:

https://kg-llm.github.io/program/pdf/2026.kgllmlrec26-1.3.pdf

Use-Case Classifier Performances

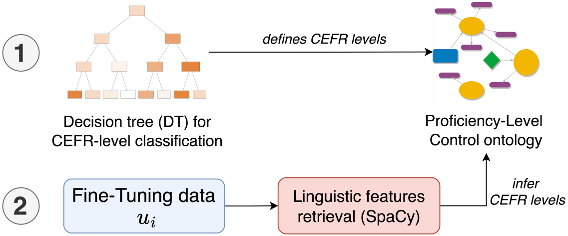


Figure 6 – Proficiency-Level Control

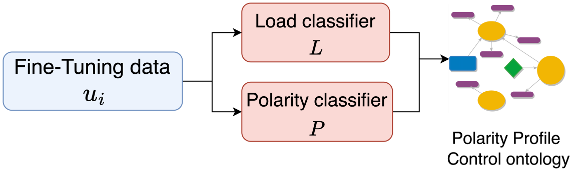


Figure 7 – Polarity Profile Control

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

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

CEFR Level Modeling Results

- We kept all features with **non-zero** feature importance.

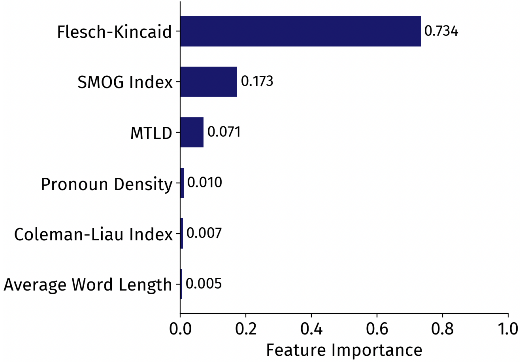


Figure 8 – Feature importances

- **MAE = 0.42**, Acc. = 0.66, wF1 = 0.65

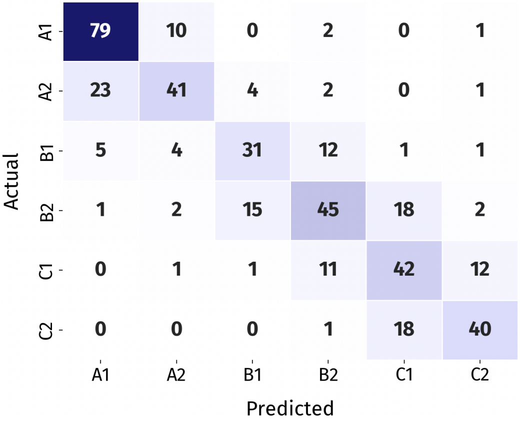


Figure 9 – Confusion matrix on CEFR-T test set

From Decision Tree Rules to Ontological Control

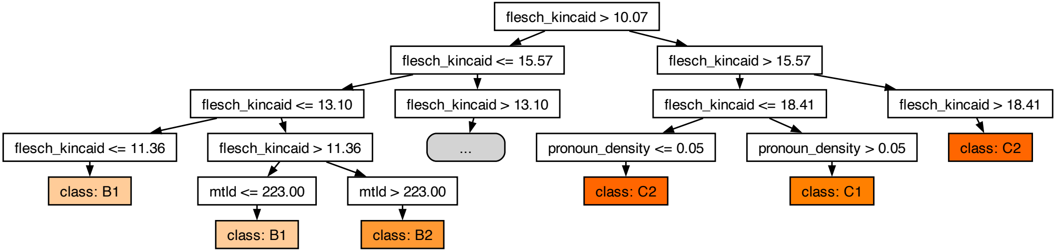


Figure 10 – Partial decision tree

Utterance
 and (((hasFleschKincaidReadabilityScore some xsd:decimal[> 11.358574390411377])
 and (hasFleschKincaidReadabilityScore some xsd:decimal[<= 13.10405158996582])
 and (hasMTLDMeasure some xsd:decimal[> 223.0])) or ((hasFleschKincaidReadabilityScore some xsd:decimal[> 13.10405158996582])
 and (hasFleschKincaidReadabilityScore some xsd:decimal[<= 14.32880687713623])
 and (hasMTLDMeasure some xsd:decimal[<= 370.25])))

Figure 11 – Definition of `B2LevelUtterance` in the ontology

Use-Case 2: Polarity Profile Control

Polarity Profile Control with 2 descriptors:

- Emotional load L
  - Non loaded (0)
  - Loaded (1)
- Emotion polarity P
  - Negative (0)
  - Neutral (1)
  - Positive (2)

6 profiles: $L_+, L_-, L_0, \neg L_+, \neg L_-, \neg L_0$

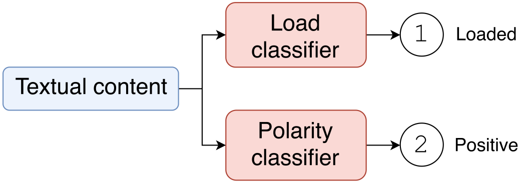


Figure 12 – Polarity profile with predictors

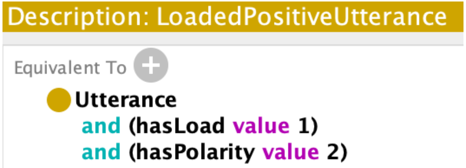


Figure 13 – Definition of `LoadedPositiveUtterance` in the ontology

Common European Framework of Reference for Languages (CEFR)

PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
INDEPENDENT USER	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Figure 14 – CEFR levels official definitions contain subjective statements

Our CEFR Level Modeling Setup

Expert-annotated **datasets**:

- `CEFR-S` sentences [Arase et al., 2022]
- `CEFR-T` texts [Nallapati et al., 2016]

44 linguistic **features** computed from `SpaCy`:

- Linguistic properties
- Readability metrics

Flesch-Kincaid Grade Level [Flesch, 1948]

$$\text{FKGL} = 0.39\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$$

SMOG Index [McLaughlin, 1969]

$$\text{SMOG} = 1.0430\sqrt{\text{Number of Polysyllables} \times \left(\frac{30}{\text{Number of Sentences}}\right)} + 3.1291$$

From Constrained Generation to Strategies

- *How should the agent's proficiency level adapt to the user's level?*
- *How should the agent's polarity adapt to the user's emotions?*

We simply **add or refine definitions** in the ontology.

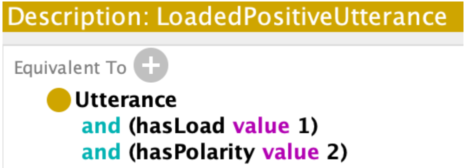


Figure 15 – Initial concept definition

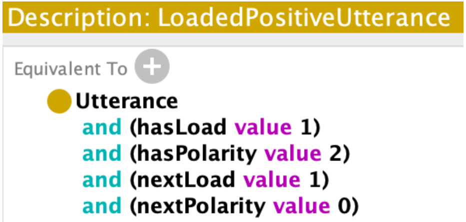


Figure 16 – Concept definition within a strategy

Datasets

- CEFR-T
  - 1499 texts for CEFR level prediction
  - Annotated by experts
  - Used to train/evaluate the decision tree
- CEFR-S
  - 10 004 sentences for CEFR level prediction
  - Annotated using the ontology
- DDbal
  - A CEFR-level balanced version of DailyDialog [Li et al., 2017].
  - 13 118 scripted conversations about daily life concerns
  - Annotated using the ontology

Level	Sentence
A1	Sheep and cows eat grass for food.
A2	The last fireworks display will be the biggest one.
B1	A full-time etiquette coach was hired.
B2	When the need for a town crier disappeared, the position passed into local folklore.
C1	Small granules in their cytoplasm contain proteins and enzymes called granzymes.
C2	Ophiuroids in general are mostly scavengers or detritivores.

Table: Some data samples from CEFR-S dataset.

Feature Definitions (1)

Flesch-Kincaid Grade Level

$$\text{FKGL} = 0.39\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$$

SMOG Index

$$\text{SMOG} = 1.0430\sqrt{\text{Number of Polysyllables} \times \left(\frac{30}{\text{Number of Sentences}}\right)} + 3.1291$$

Measure of Textual and Lexical Diversity (MTLD)

$$\text{MTLD} = \frac{\text{Total Words}}{\text{Mean Length of TTR Segments}}$$

Feature Definitions (2)

Coleman-Liau Index

$$\text{CLI} = 0.0588L - 0.296S - 15.8$$

L = average number of letters per 100 words

S = average number of sentences per 100 words

Pronoun Density

$$\text{PD} = \frac{\text{Number of Pronouns}}{\text{Total Words}}$$

Average Word Length

$$\text{AWL} = \frac{\text{Number of Characters (excluding spaces)}}{\text{Total Words}}$$

Full Fine-Tuning Results for Both Use-Cases

Model	**Proficiency-Level Control**					**Polarity Profile Control**			
	F1 ± Std	F1 Range	Acc	MAE	B_r	F1 ± Std	F1 Range	Acc	MCC
				Pre-Trained Baselines					
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
				Ours (Ontology-Guided CLM Fine-Tuning)					
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: Model performance comparison in zero-shot generation for Proficiency-Level Control and Polarity Profile Control tasks. B_r is the BERT-F1 Score ratio. Best scores are in **bold**.