

PROJECT PROPOSAL PRESENTATION

Large Language Models for Software Engineering

02/11/2026

Politecnico di Torino

Reasoning and Inference

1. RETECO: REASONING-ORIENTED RETRIEVAL

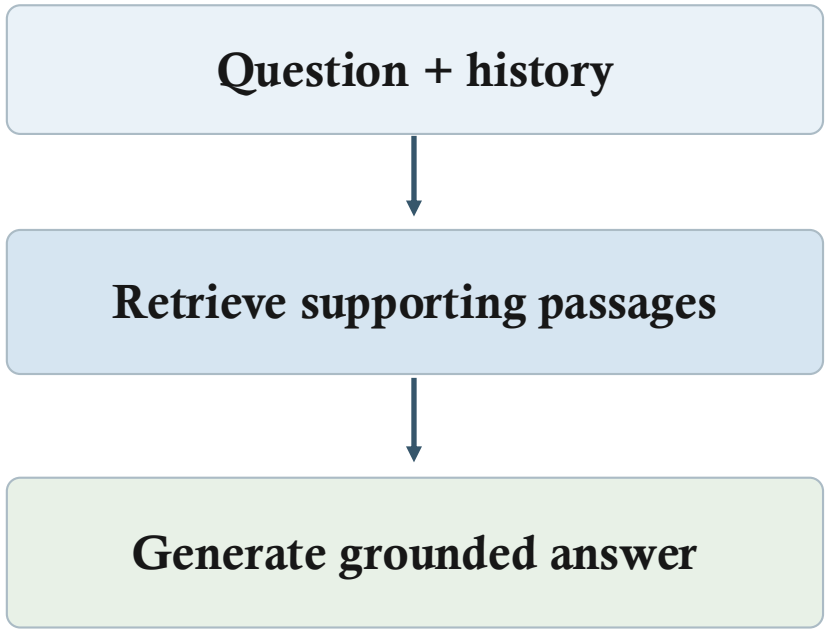
Evaluate how well retrieval systems find evidence when a question requires temporal reasoning or understanding a conversation.

Track 1 - Temporal Grounded Retrieval

Retrieve an ordered list of documents supporting a time-dependent question, either directly or step by step.

Track 2 - Conversational Retrieval

Use previous turns of a conversation to retrieve relevant passages, generate answers from reference passages, or run a full end-to-end RAG system.



Temporal evidence + conversational context

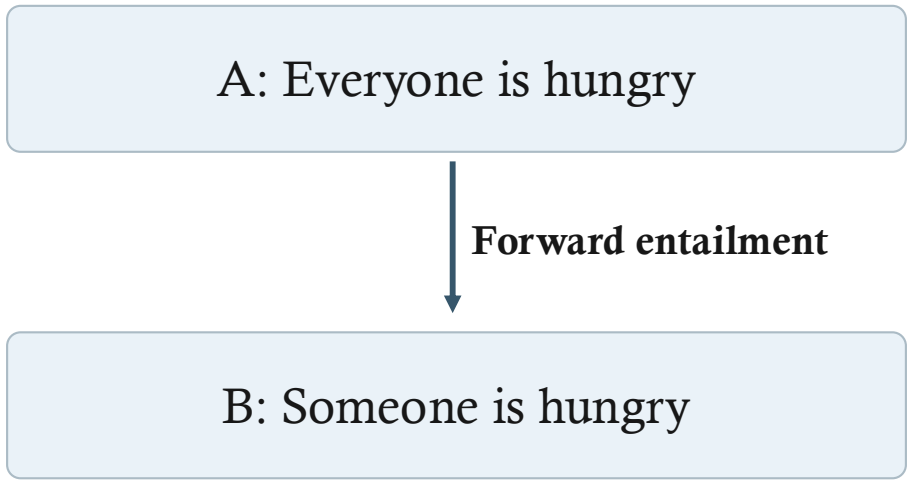
Retrieval + Text Generation

2. DiCo-NLI: DIRECTIONAL-CONSISTENT INFERENCE

Evaluate logical relationships between two phrases, including what happens when the order of the phrases is reversed.

Possible relationships:

- Equivalence: both phrases imply each other.
- Forward entailment: A implies B.
- Backward entailment: B implies A.
- Negative / Other: no entailment relationship.



A → B is valid; B → A is not.

Multi-class Classification

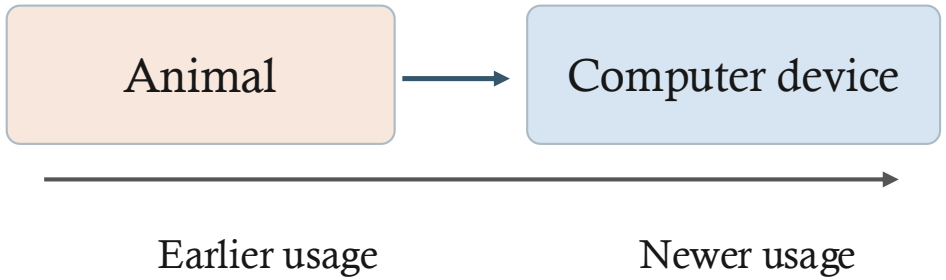
Politecnico di Torino

Lexical and Compositional Semantics

3. SEMANTIC CHANGE ACROSS MULTIPLE TIME PERIODS

Study how the meanings of words evolve across historical periods and how frequently each meaning is used.

Example: the word "mouse"



Subtask 1 - Word Sense Induction

Group occurrences of a word by meaning, then follow the distribution of those meanings through time.

Subtask 2 - Hypothesis-Driven Change

Given a particular meaning, identify which occurrences express that sense.

The same word can acquire new senses.

Usage	Senses
"The mouse ran across the floor."	[2]
"The laboratory mouse..."	[2]
"The mouse was on the kitchen table."	[1,2]
"Click the mouse button."	[1]

10 time periods, 6 languages (including Italian).

Clustering + Classification

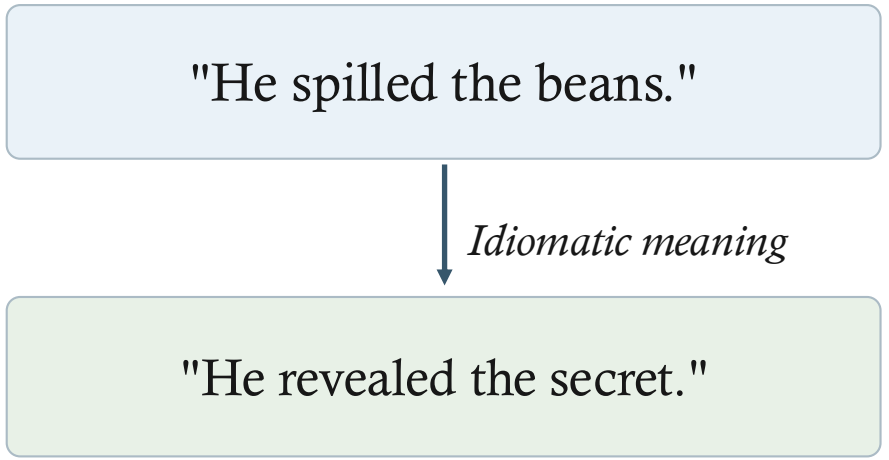
Politecnico di Torino

4. Par-PARSEME: MULTILINGUAL IDIOM PARAPHRASING

Identify idiomatic expressions and rewrite sentences while preserving their intended meaning.

Task:

1. Locate an idiom in a sentence.
2. Understand its meaning in context.
3. Generate a paraphrase without the idiom.



Not word-for-word translation

The benchmark covers 15 languages. Paraphrases are evaluated for quality and diversity (there will be also a manual evaluation!).

Text Generation

5. ComPartMENT: COMPOSITIONALITY IN CONTEXT

Predict how much the meaning of a multiword expression can be derived from the meanings of its individual words.

Focus:
- Noun compounds (e.g., "glass ceiling", "coffee cup").
- Particle verbs (e.g., "take off").

"He took off his coat."

More literal

Task:
Given an expression in context, predict its degree of compositionality (a numerical score).

"Her career took off."

More figurative



Compositionality spectrum (from 0 to 5)

English and German; contemporary and historical texts.

Regression

Politecnico di Torino

Cultural and Cross-Lingual Understanding

6. CULTURALLY GROUNDED VISUAL QUESTION ANSWERING

Answer culture-specific questions about an image, rather than only identifying visible objects.

Task 1 - Spoken Visual QA

Image + spoken question (audio)
→ short textual answer.

Task 2 - Textual Visual QA

Image + written question
→ short textual answer.



IMAGE

Written question

Spoken question

Short answer

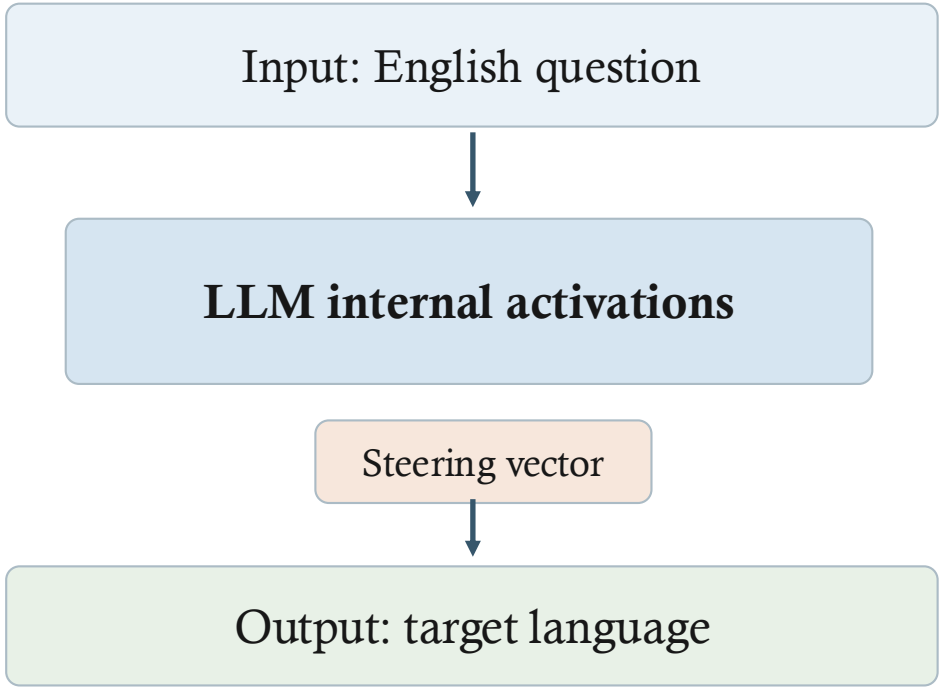
Visual Question Answering (Generation)

7. CLaS: CROSS-LINGUAL ALIGNMENT AND STEERING

Control the output language of an LLM by intervening in its internal representations, while maintaining answer quality.

Four tracks:

1. High-resource language steering.
2. Low-resource language steering.
3. Full multilingual steering.
4. Bonus: multilingual refusals.



Important: simply instructing the model to answer in another language is not allowed. Steering must modify internal computations.

Representation Steering

12. ArtEmotion-28: CROSS-CULTURAL EMOTION ANALYSIS

Evaluate how emotion and sentiment attributed to the same artwork vary across languages and cultures.

Subtask 1 - Emotion Prediction

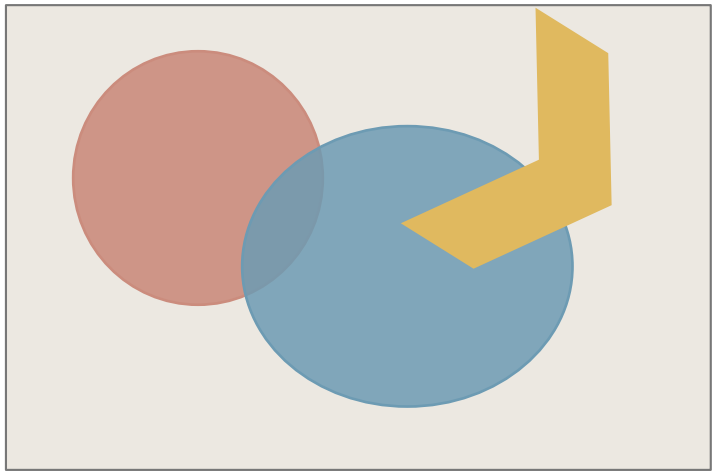
Artwork + native-speaker caption
→ one of 8 emotions. Compare text-only and multimodal runs.

Subtask 2 - Sentiment Divergence

From a source-language sentiment distribution, predict a (different) target-language sentiment distribution for an artwork.

Subtask 3 - Image-Only Prediction

Artwork + target language
→ proportion of positive responses.



Same artwork

Culture A: positive

Culture B: negative

Multi-class Classification + Regression

Politecnico di Torino

Stereotypes, Framing, and Sentiment Analysis

8. StereoQueerEval: STEREOTYPES AND HATE SPEECH

Detect stereotypes and explicit or implicit hate speech targeting LGBTQIA+ people in online comments.

Subtask A - Stereotype Detection

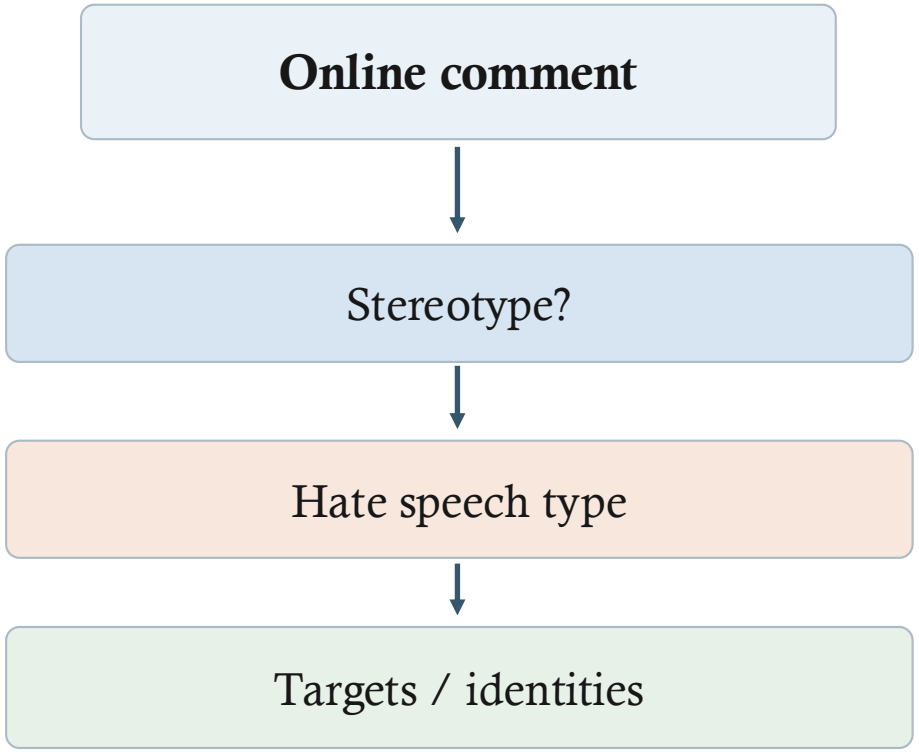
Binary classification: is a stereotype present?

Subtask B - Hate Speech Classification

Explicit hate speech, implicit hate speech, or non-hateful/supportive content.

Subtask C - Target Identification

Identify who is targeted and which identities are mentioned (multilabel and multiclass task).



Binary + Multi-class/Multi-label Classification

9. ANALYSIS OF MULTIMODAL FRAMING IN THE NEWS

Identify the perspectives or frames through which news articles and accompanying images present an event.

Subtask 1 - Text Framing

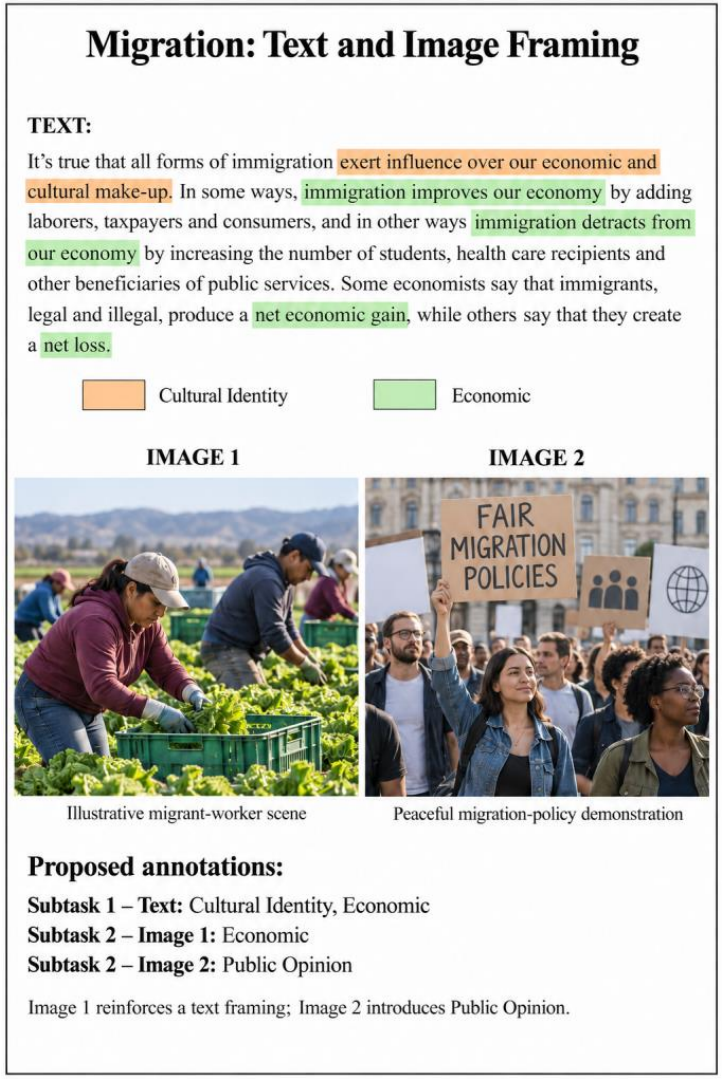
Classify each paragraph into zero or more of 14 framing categories (multi-label classification).

Subtask 2 - Image Framing

Identify frames conveyed by news images, with or without gold text-framing labels.

Subtask 1 can be completed using text-only models.

Multi-label Classification



Politecnico di Torino

Emerging Topics: Agents and AI Tools

10. AgentRisk: RISK IN AGENTIC INTERACTIONS

Evaluate risks in multi-turn interactions between an AI agent, users, and external tools.

Subtask 1 - Risk Level Estimation

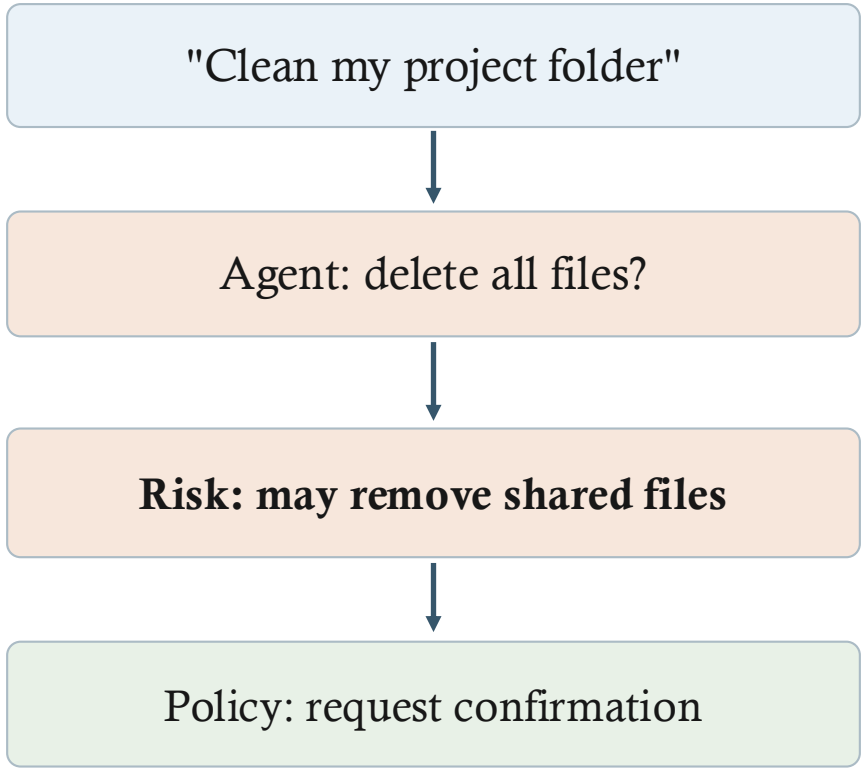
Rate the risk from 1 (Minimal) to 5 (Critical).

Subtask 2 - Risk Characterization

Identify the potential harm domain and the reason the action is risky (multilabel classification).

Subtask 3 - Policy-Guided Intervention

Choose an appropriate response under a given risk policy.



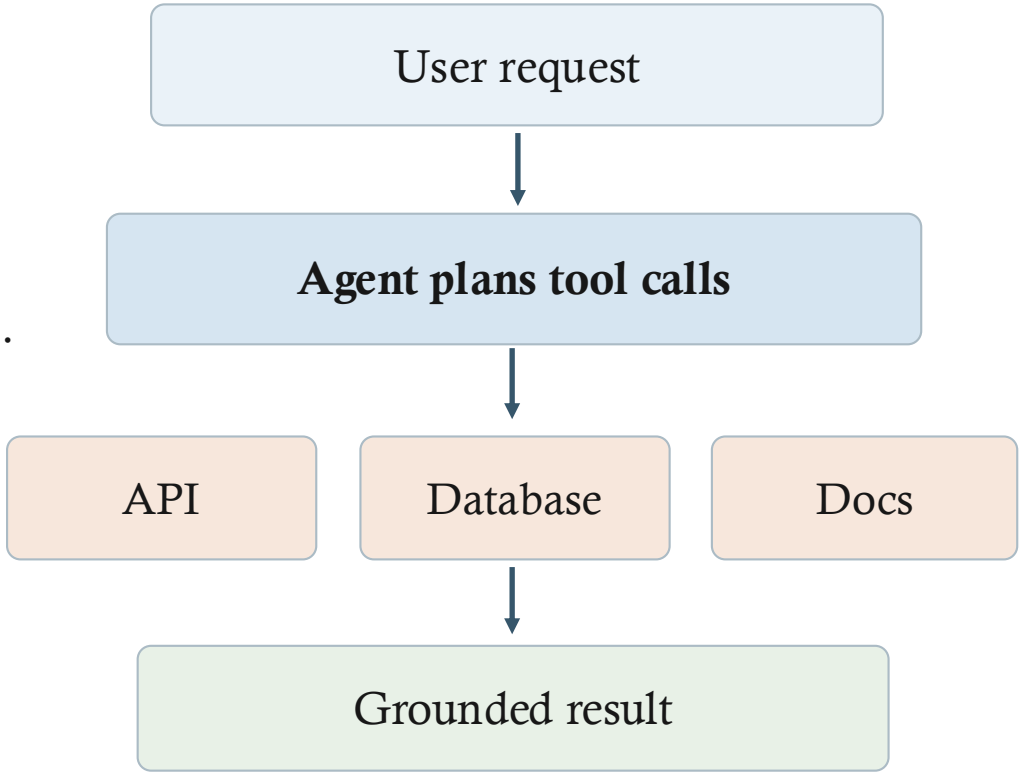
Regression + Classification + Tool Selection

11. VAKRA-ADVANCED: MULTI-HOP TOOL CALLING

Solve complex requests by selecting and executing multiple tools, combining data from different sources.

Four subtasks:

1. API Chaining: execute connected API calls.
2. Tool Selection: choose the relevant tools.
3. Multi-hop API Reasoning: reuse earlier results.
4. Multi-source Reasoning: combine APIs and documents.



Requires an executable tool environment, not just text generation.

Tool Selection + Agentic Planning

Thank you for your attention!



All SemEval 2027 tasks are available at the following link