PROJECT PROPOSAL PRESENTATION

Large Language Models for Software Engineering
03/11/2025

1. MWAHAHA: A COMPETITION ON HUMOR GENERATION

Subtask A: Text-based Humor Generation

Given a set of text-based constraints, generate a joke. This subtask will be conducted in English, Spanish, and Chinese.

Constraints:

Each generated joke must respect one of the following constraints, designed to make it difficult to retrieve existing jokes from the web:

- **Word Inclusion**: Must contain two specific words (from a list of rare word combinations).
- **News Headline**: Must be related to a given news article headline (it could be a punchline, or a joke inspired by the headline).

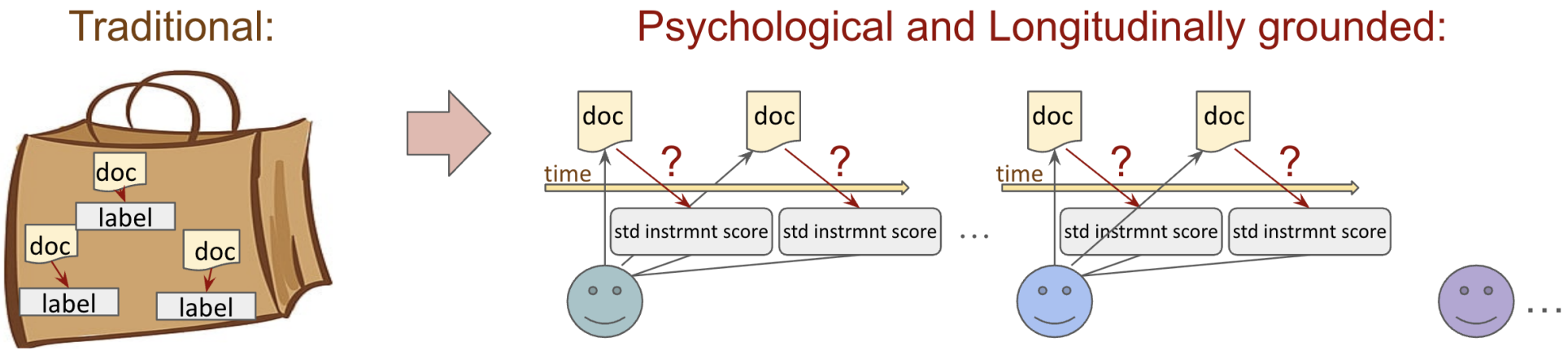
The evaluation will be based on human preference judgments.

Since sub-task B would require the use of VLM, we do not recommend actually doing it.

Predict (and forecast!) how emotions evolve over time within a text. Given an essay, the system must output:

- **Valence → positive or negative emotion**
- **Arousal → calm or excited state**

Predictions are made at the sentence or segment level, forming a time series of emotional variation across the essay



3. DIMENSIONAL ASPECT-BASED SENTIMENT ANALYSIS

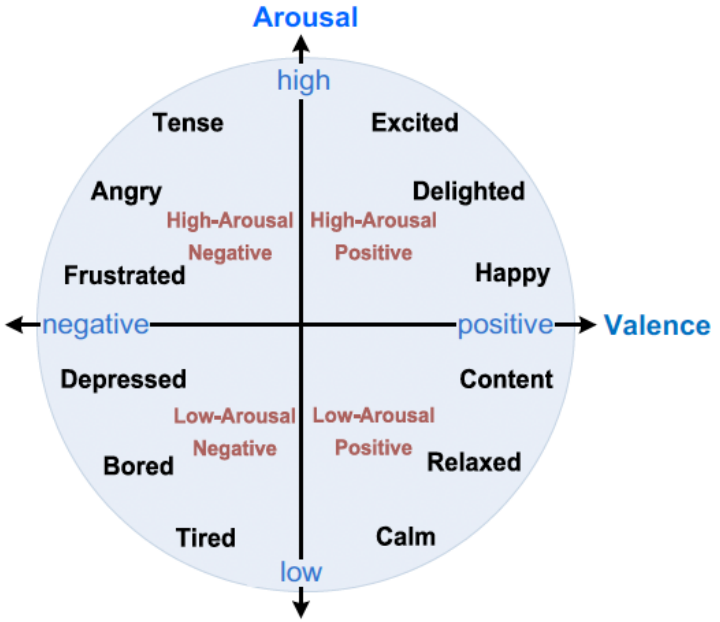
Similar to the task 2, with the difference that here we are talking about a multilingual and multidomain dataset that changes across the two different proposed subtasks

Track A - DimABSA

Focuses on opinions about specific things (aspects) in text.
Example: "The food was great but the service was slow."
The system must find:

What is being talked about (*food*, *service*)
How positive and intense the feeling is → e.g.
food → *8.0#7.5*, *service* → *3.5#5.0*

Used for reviews: restaurants, laptops, hotels, etc.



Track B - DimStance

Focuses on people's opinions about topics or issues.
Example: "I support clean energy policies."
The system predicts how positive and strong that opinion is → *7.5#6.8*.
Used for stance detection in politics or social debates.

4. NARRATIVE STORY SIMILARITY AND NARRATIVE REPRESENTATION LEARNING

In this task, you are asked to identify stories that are narratively similar. Three core similarity components determine narrative similarity:

1. **Abstract Theme**: The ideas and motives of the story.
2. **Course of Action**: The sequence of central events, turning points, etc.
3. **Outcomes**: The results of a story.

⚓ Anna loses her purse. She is terrified because there are important documents in it. She retraces her steps but cannot find it. Dan finds it and helpfully returns it to her.

A Brian lost his backpack. He did not care too much, as only a water bottle was in it. After an hour of search, he finally found it.

B Alex loses his engagement ring while swimming. He freaks out, and after hours of diving for it, he still cannot find it.

5. RATING PLAUSIBILITY OF WORD SENSES IN AMBIGUOUS SENTENCES THROUGH NARRATIVE UNDERSTANDING

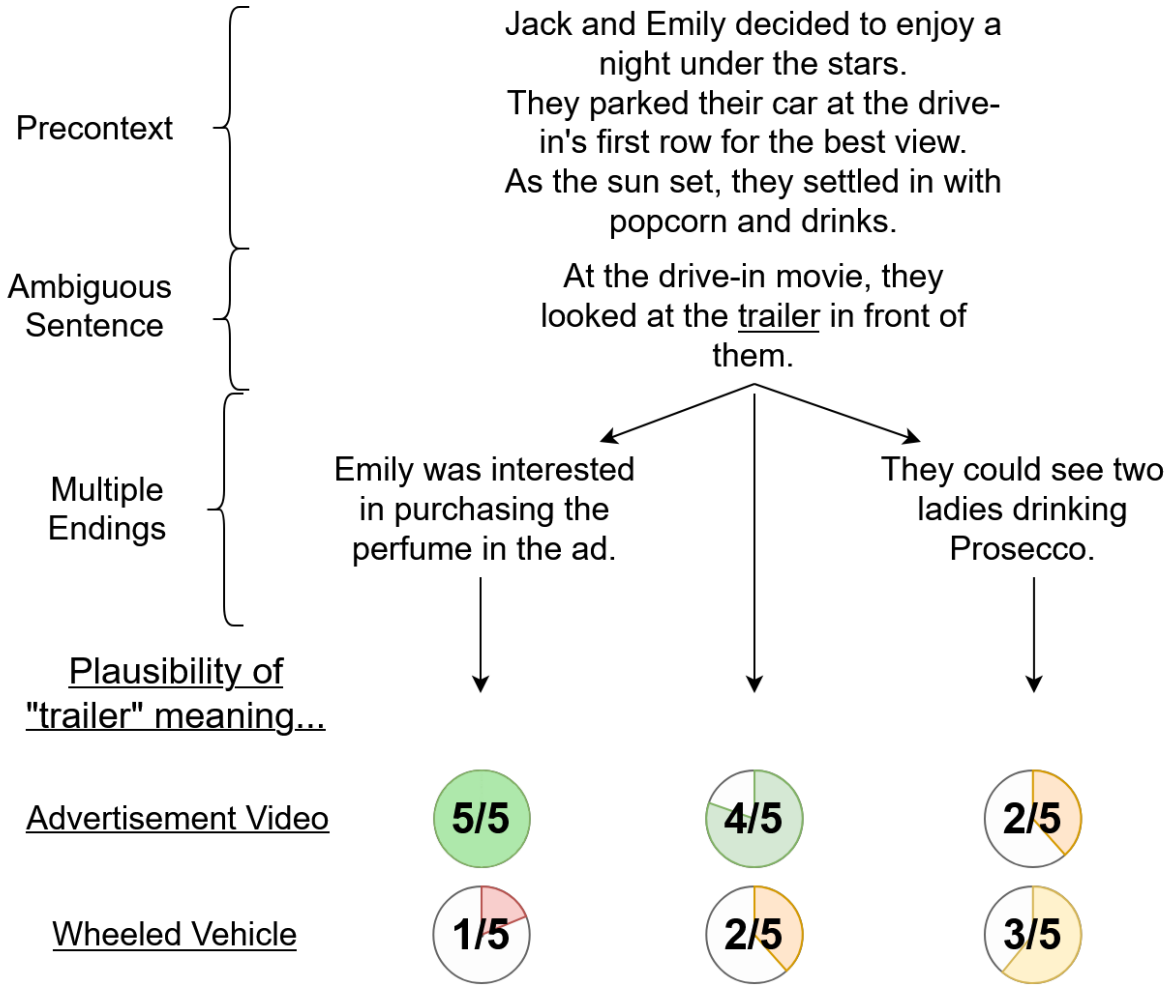
Study how humans and models interpret ambiguous words in context, when more than one meaning can be plausible, not just "correct" or "wrong."

Task:

Given a short 5-sentence story, determine which sense of an ambiguous word fits best.

Each story has:

- **Precontext**: 3 sentences that set the scene
- **Ambiguous sentence**: contains a homonym (e.g., trailer)
- **Possible endings**: each implying a different meaning



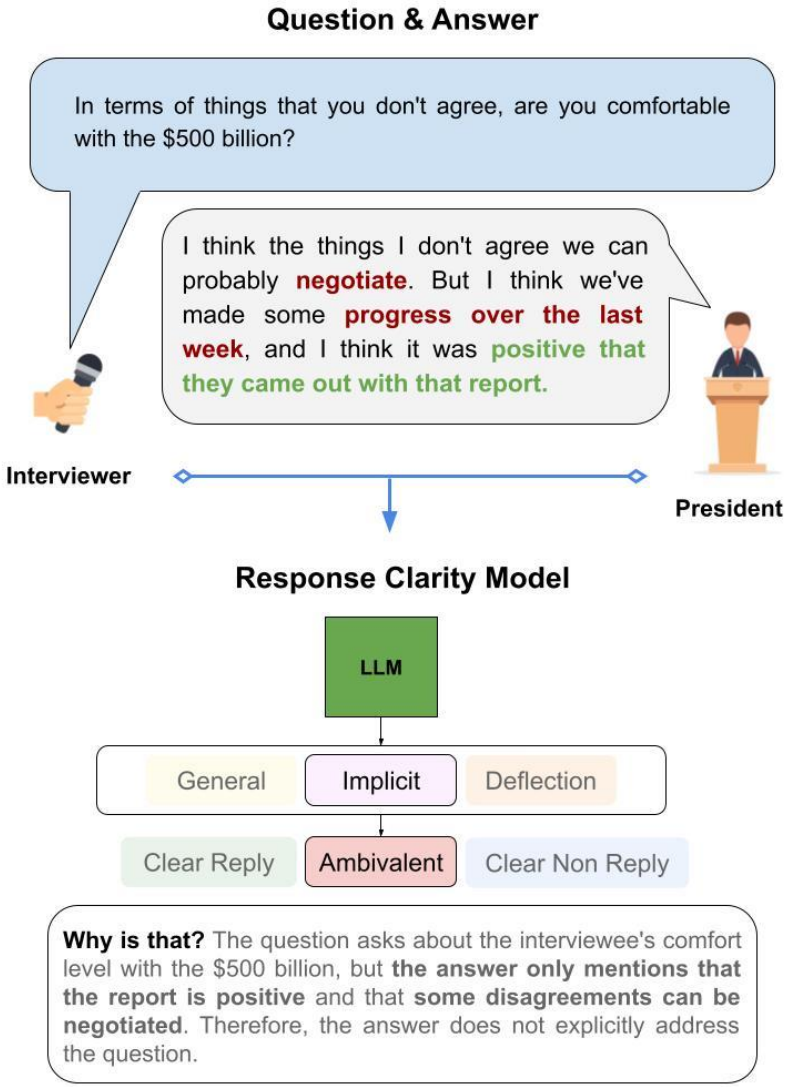
6. UNMASKING POLITICAL QUESTION EVASIONS

Task 1 - Clarity-level Classification

Given a question and an answer, classify the answer as Clear Reply, Ambiguous or Clear Non-Reply.

Task 2 - Evasion-level Classification

Given a question and an answer, classify the answer into one of the 9 evasion techniques.



7: EVERYDAY KNOWLEDGE ACROSS DIVERSE LANGUAGES AND CULTURES

Evaluate how well language models understand everyday, culture-specific knowledge across 26 languages and 30 countries

Tasks:

1. Short-Answer Questions (SAQ):
Questions and answers in 26 local languages.

2. Multiple-Choice Questions (MCQ):
English questions with four culturally distinct answer options.
The model must pick the most appropriate answer for a given country/region.

Area	Language (Region)
Africa	**Arabic (Algeria)**, **Amharic (Ethiopia)**, **Hausa (Northern Nigeria)**, Afrikaans (South Africa), Arabic (Egypt), Arabic (Morocco)
Asia	**Assamese (Assam)**, **Azerbaijani (Azerbaijan)**, **Chinese (China)**, **Indonesian (Indonesia)**, **Persian (Iran)**, **Korean (North Korea)**, **Korean (South Korea)**, **Sundanese (West Java)**, Arabic (Saudi Arabia), Japanese (Japan), Thai (Thailand), Bengali (India), Tagalog (Philippines), Tamil (Sri Lanka), Taiwanese Mandarin (Taiwan), Singaporean Mandarin (Singapore), Malay (Singapore)
Europe	**Greek (Greece)**, **Spanish (Spain)**, **English (UK)**, French (France), Bulgarian (Bulgaria), Swedish (Sweden), Irish (Ireland)
America	**English (US)**, **Spanish (Mexico)**, Spanish (Ecuador)

8: EVALUATING MULTI-TURN RAG CONVERSATIONS

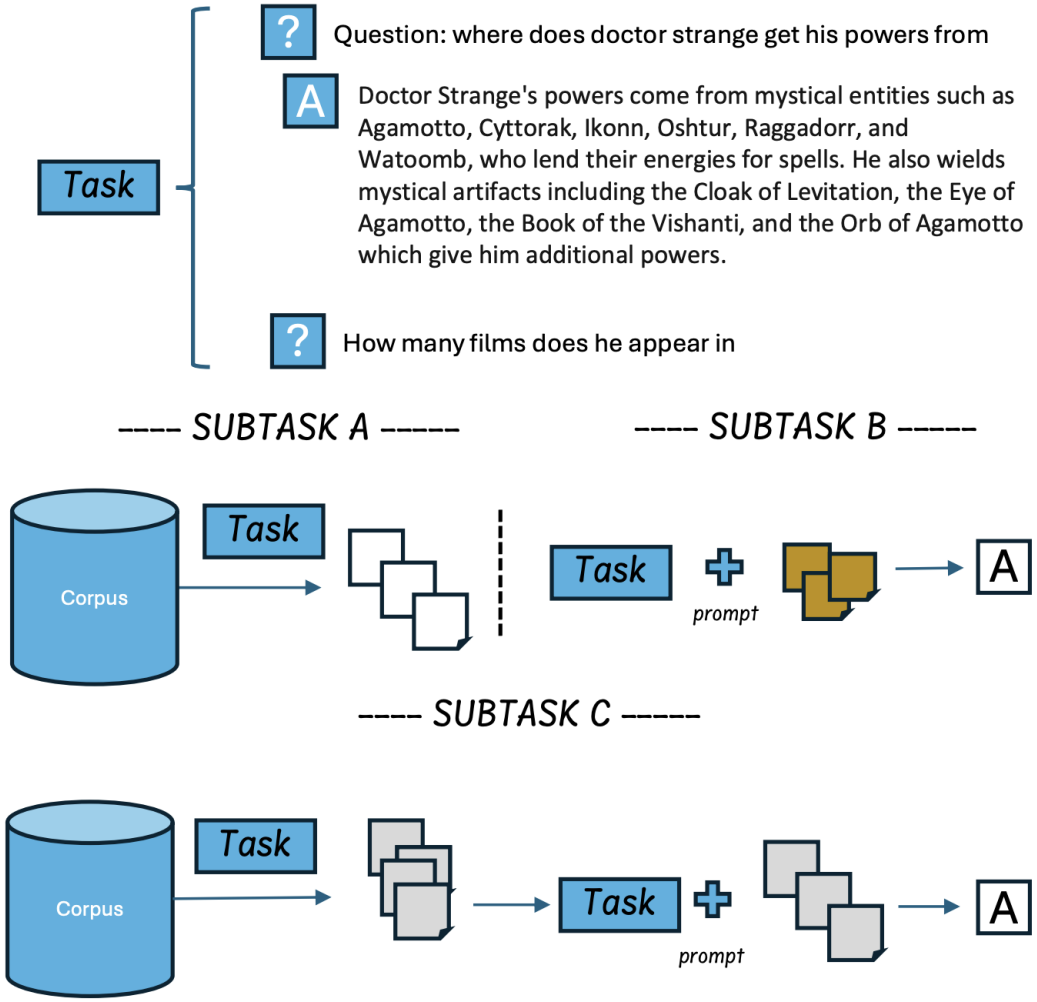
Evaluate how well language models handle multi-turn information-seeking conversations using Retrieval-Augmented Generation (RAG)

Tasks:

1. Retrieval:
Given a conversation turn, retrieve the most relevant passages.

2. Generation (with Reference):
Given gold reference passages, generate the correct answer.

3. Full RAG:
Retrieve passages and generate an answer from them (end-to-end evaluation).



Evaluate how well models can detect and understand polarization in social media discussions across 22 languages, reflecting diverse cultural and linguistic contexts.

Tasks:

1. Polarization Detection

Determine whether a message expresses polarized attitudes or not.

2. Polarization Type Classification

Identify the primary targets of polarization, including political, social, or identity-related divisions. Each post may include more than one type of polarization.

3. Manifestation Identification

Detect the linguistic and rhetorical forms through which polarization appears, such as generalization and hostility.

10: PSYCHOLINGUISTIC CONSPIRACY MARKER EXTRACTION AND DETECTION

Focus on understanding and detecting conspiracy-related content in online conversations. It's built from Reddit submission statements.

Tasks:

1. Marker Extraction

Identify psycholinguistic components of conspiratorial thought (such as references to agents, victims, or consequences), revealing how conspiracy narratives are linguistically constructed.

2. Conspiracy Detection

Classify whether a given text reflects conspiratorial thinking, based on its underlying reasoning style and discourse structure, not just keywords or topics.

11: DISENTANGLING CONTENT AND FORMAL REASONING IN LANGUAGE MODELS

Assess whether language models can perform pure logical reasoning (evaluating the formal validity of syllogisms) independently of world knowledge or plausibility.

Tasks:

1. English Syllogistic Reasoning
Evaluate validity judgments in English; measure both accuracy and content bias (intra-, cross-, and total content effect).

2. English Reasoning with Irrelevant Premises
Same as (1) but includes distracting premises; models must also select relevant information (F1 on premise selection).

3. Multilingual Syllogistic Reasoning
Test cross-lingual generalization: predict validity in several target languages and measure multilingual content effect.

4. Multilingual Reasoning with Irrelevant Premises
Combine multilingual validity prediction and premise selection under noisy conditions.

Evaluate how well language models can reason about causes, not just describe or predict events. Given an observed event and related documents, the model must infer the most plausible and direct cause, simulating human reasoning.

Task:

Formulated as a multiple-choice reasoning task. For each instance, the model receives:

- **Event**: brief real-world outcome
- **Context**: retrieved supporting and distracting documents
- **Options (A–D)**: candidate explanations, including one stating "insufficient information."

The model must choose the correct option(s) based on reasoning over the context.

Build a system that can detect machine-generated code under diverse conditions by evaluating generalization to unseen languages, generator families, and code application scenarios.

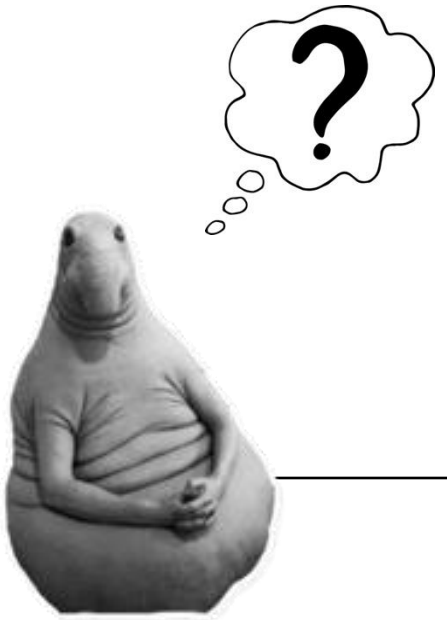
Task:

1. **Binary Machine-Generated Code Detection**
Given a code snippet, predict whether it is human-written or machine-generated.

2. **Multi-Class Authorship Detection**
If the code is machine-generated, also infer the actual family that generated it.

3. **Hybrid Code Detection**
Classify each code snippet as 1) human-written 2) machine-generated 3) hybrid or 4) adversarial (generated via adversarial prompts or RLHF to mimic humans)

THANK YOU FOR YOUR ATTENTION!



All the challenges are available at the following link