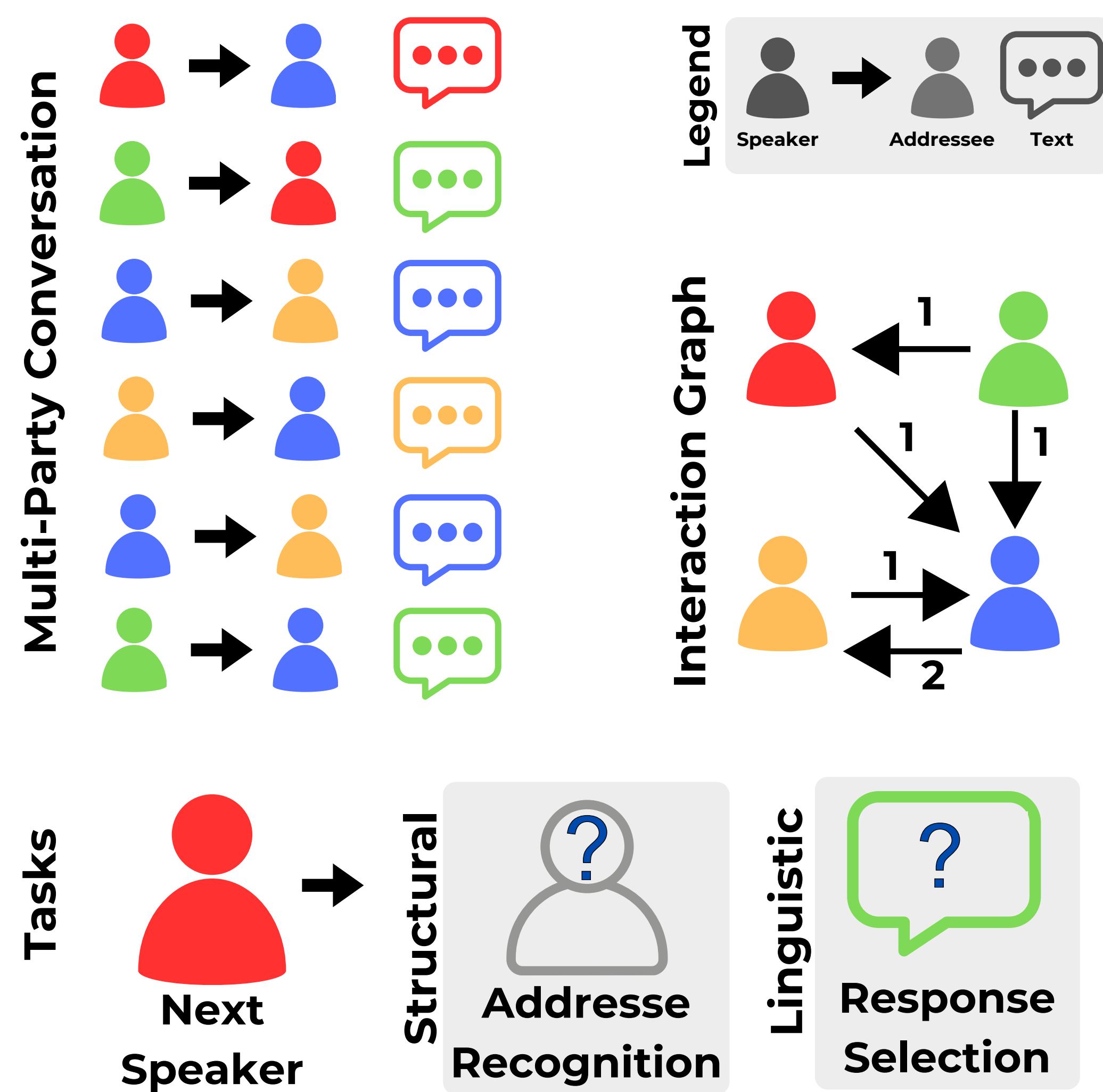


# Evaluating LLMs on Multi-Party Conversations: a diagnostic pipeline

Nicolò Penzo, Maryam Sajedinia, Bruno Lepri, Sara Tonelli, Marco Guerini

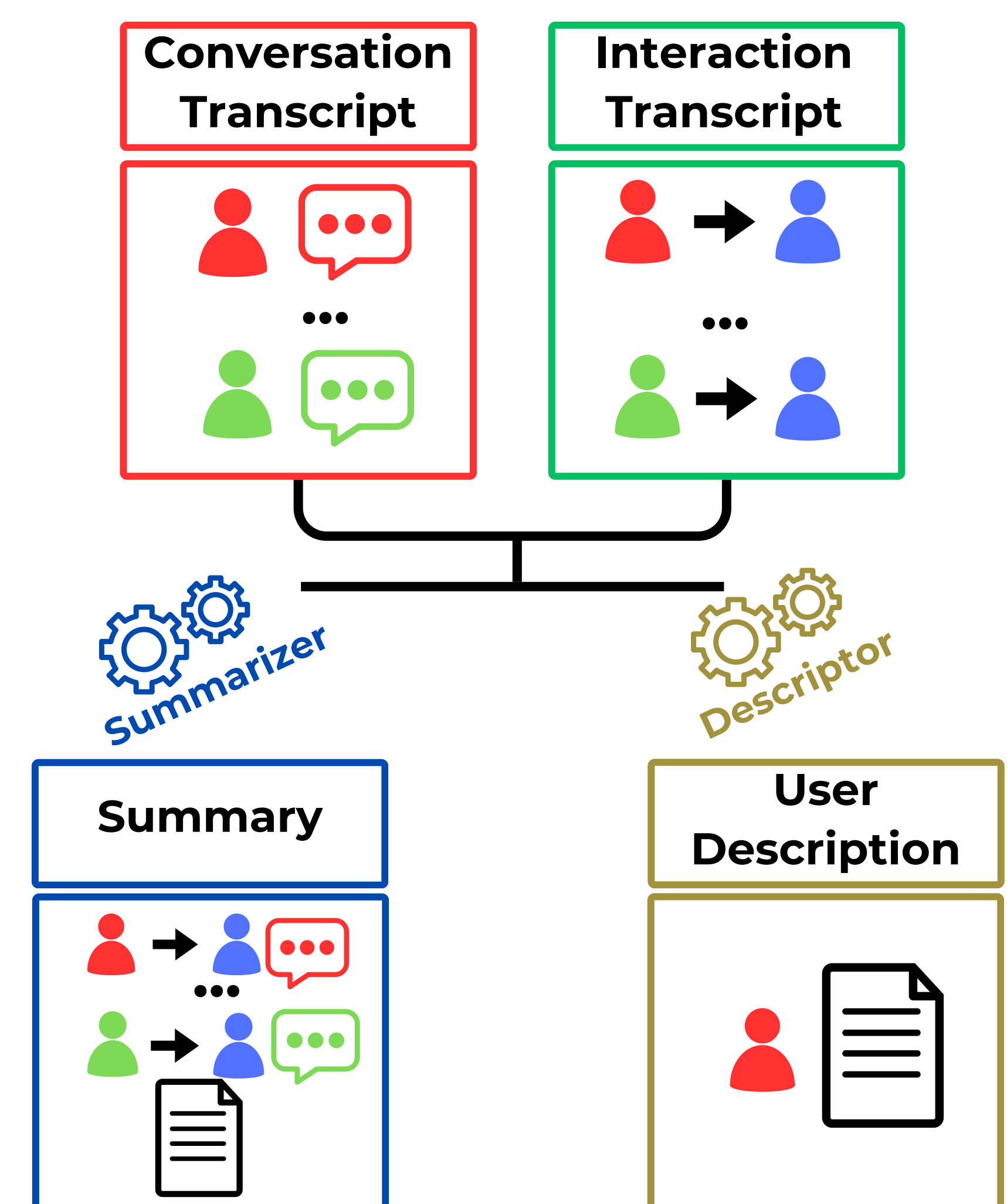
1

## Context



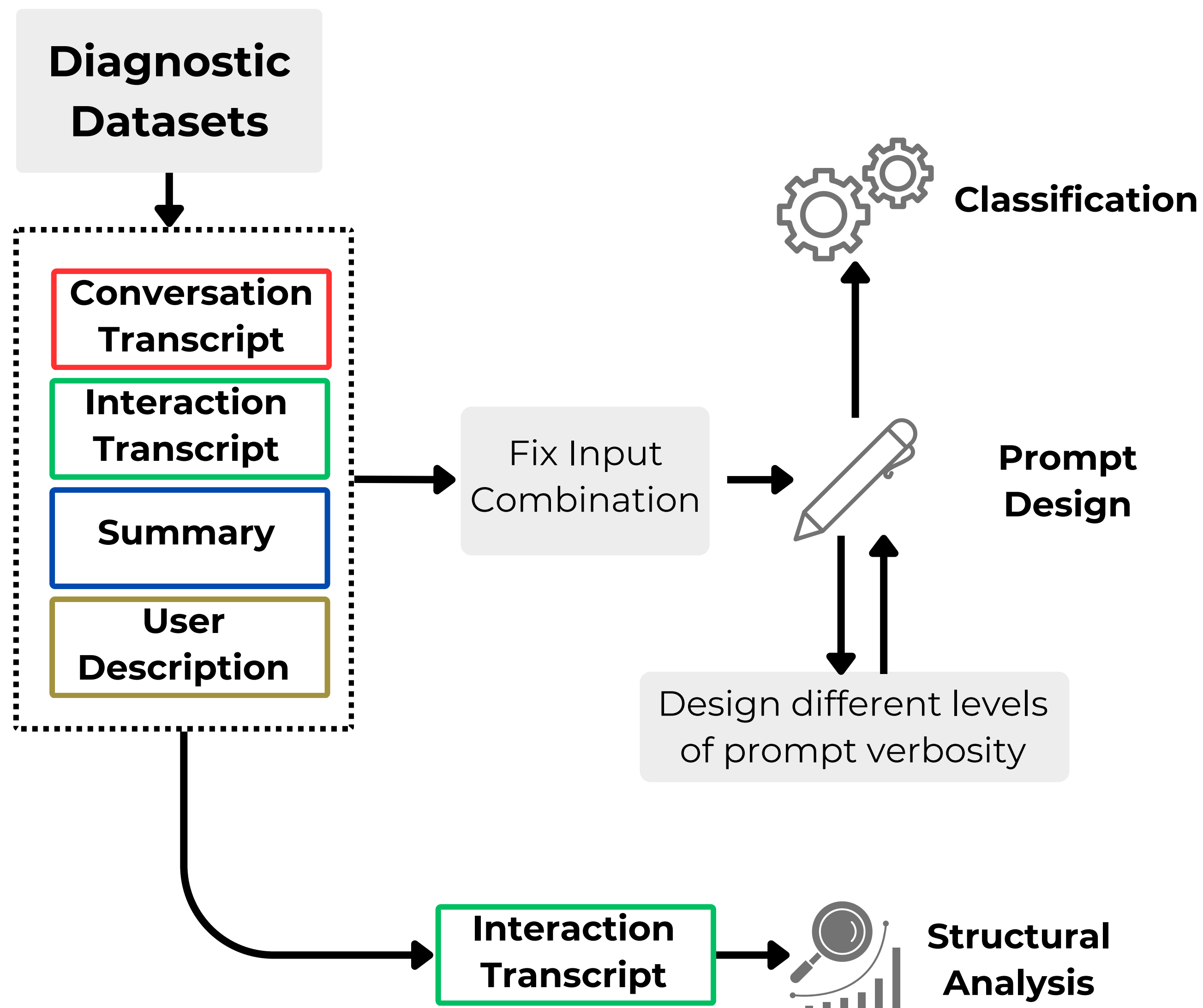
2

## Input Information



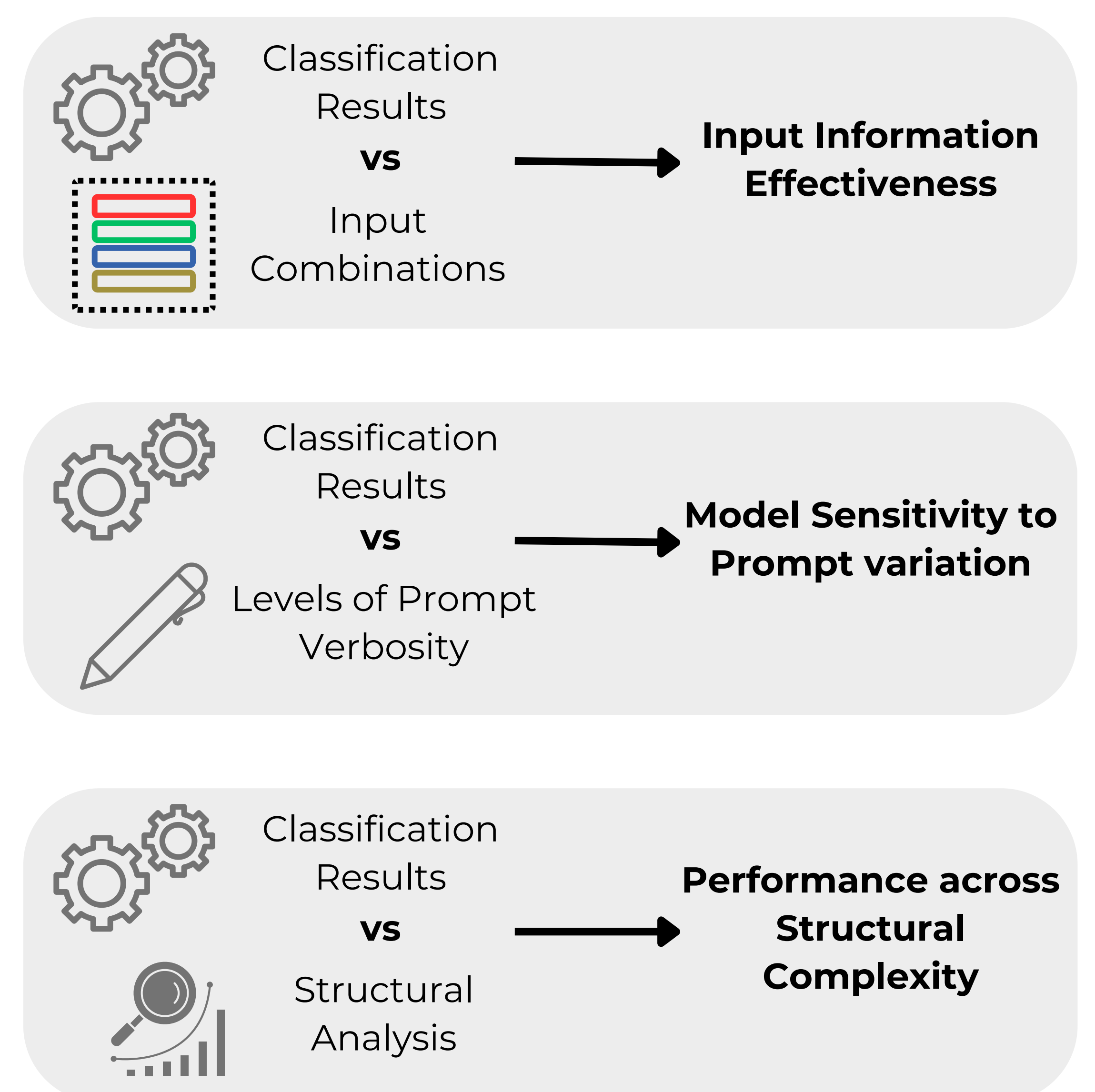
3

## Pipeline



4

## Evaluation



## FINDINGS

- **Interaction Transcripts** are **crucial** for the **structural** task, but **not** for the **linguistic** task.
- The **structural** task is more sensitive to **prompt variation** compared to the linguistic task.
- In the **structural** task, **performance differences** across input combinations (always including Interaction Transcript) is due to **improved outcomes** on items of **low structural complexity**.

POSTER



Nicolò Penzo

✉ npenzo@fbk.eu  
✉ @penzo\_nicolo  
🌐 nicolopenzo.github.io

Affiliations

University of Trento, Trento (Italy)  
Fondazione Bruno Kessler, Trento (Italy)  
University of Turin, Turin (Italy)

