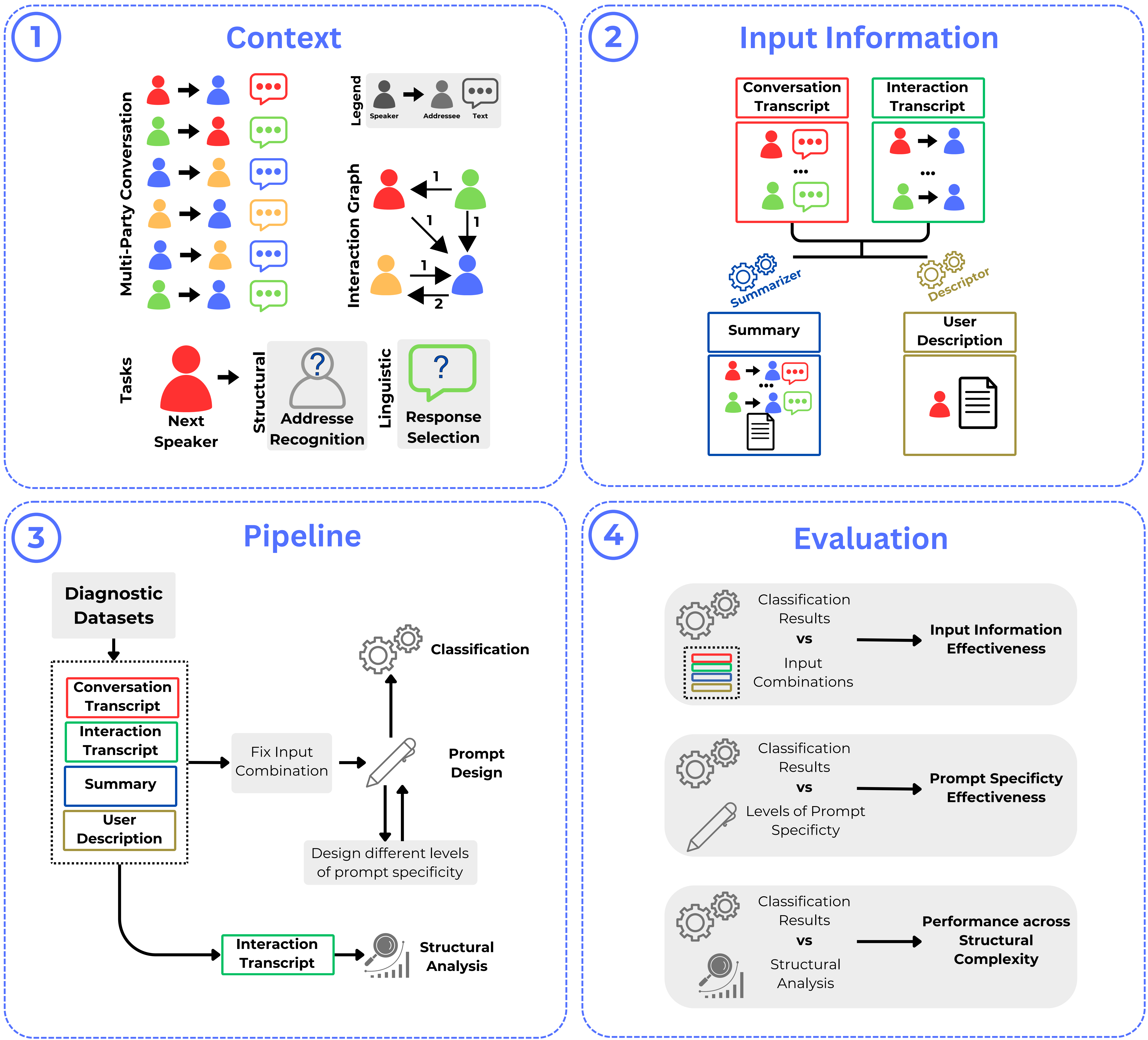


Evaluating LLMs on Multi-Party Conversations: a diagnostic pipeline

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



FINDINGS

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

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)

