



CLiC-it 2026

Palermo - Italy, 14-16 September, 2026

Plausible but Different

Comparing Human and LLM-Simulated Open-Ended Survey Responses in Italian

Marco Rospocher

with

M. Bombieri (University of Trento)

S. Cramerotti, D. Ianes and B. Zagni (Fondazione Scuola e Ricerca + Centro Studi Erickson)



UNIVERSITÀ
di **VERONA**

Dipartimento
di **LINGUE**
E LETTERATURE STRANIERE



Motivating question

- LLMs can generate plausible survey answers
- But surveys measure population-level distributions
- Open-ended answers add extra complexity: style, length, framing, vocabulary
- **Do LLMs reproduce human-like distributions of open-ended answers?**

Case study

- Italian school-inclusion online survey (*La qualità dell'inclusione scolastica e sociale 2025*)
 - Domain: school inclusion
 - Respondents: teachers, support teachers, educators, school principals, parents, ...
 - Format: **2 open-ended questions**, 14 word associations, 3 Likert-scale closed questions
 - Q1: successful inclusion experience: *Recall a situation in which you experienced full inclusion at school. What worked well?*
 - Q2: obstacles to inclusion: *Thinking about your specific situation in the last school year, what obstacles did you encounter for inclusion?*
- Human sample: 831 valid submissions

Generation protocol

- Profile-conditioned simulation

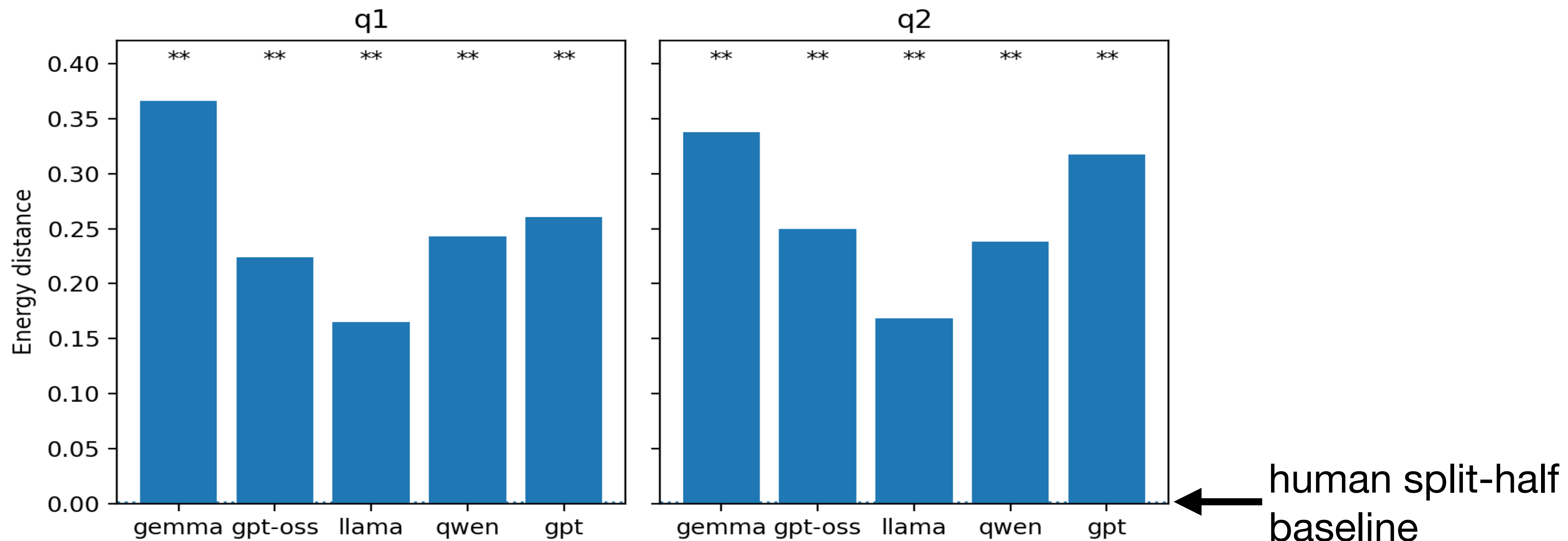


- 5 Italian-capable LLMs:
 - Gemma-3-27B-it, gpt-oss-20b, Llama-3.1-8B, Qwen3-30B-A3B, and GPT-4.1 Mini
- 831 submissions for each model, one synthetic submission for each profile
- Instructions: simulate a plausible respondent, online survey-completion context
- Zero-shot prompt: just (anonymized) profile and questionnaire

Evaluation strategy

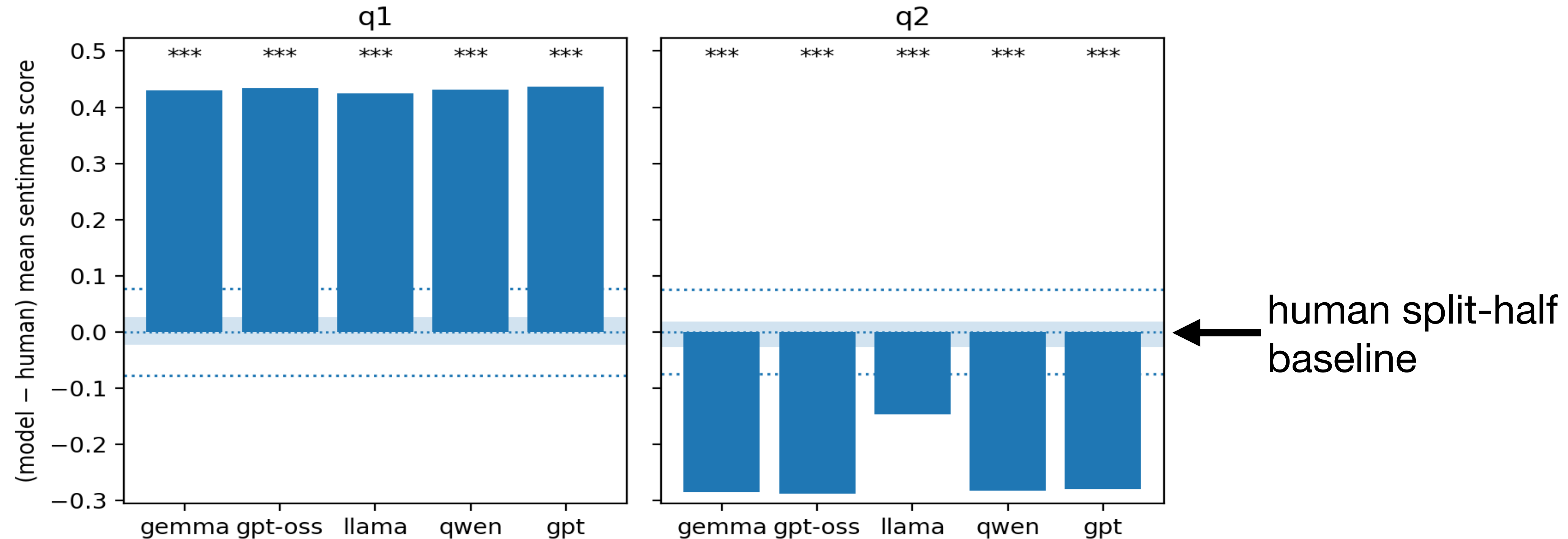
- We propose four complementary views
 - Semantic distributional distance
 - *Are model and human answers located similarly in embedding space?*
 - Sentiment mean shift
 - *Do model answers receive more positive or more negative sentiment scores?*
 - Response length
 - *Do models answer with the same level of brevity or verbosity as humans?*
 - Distinctive lexical choices
 - *Which words are characteristic of human answers, and which of generated answers?*

Results: semantic distance



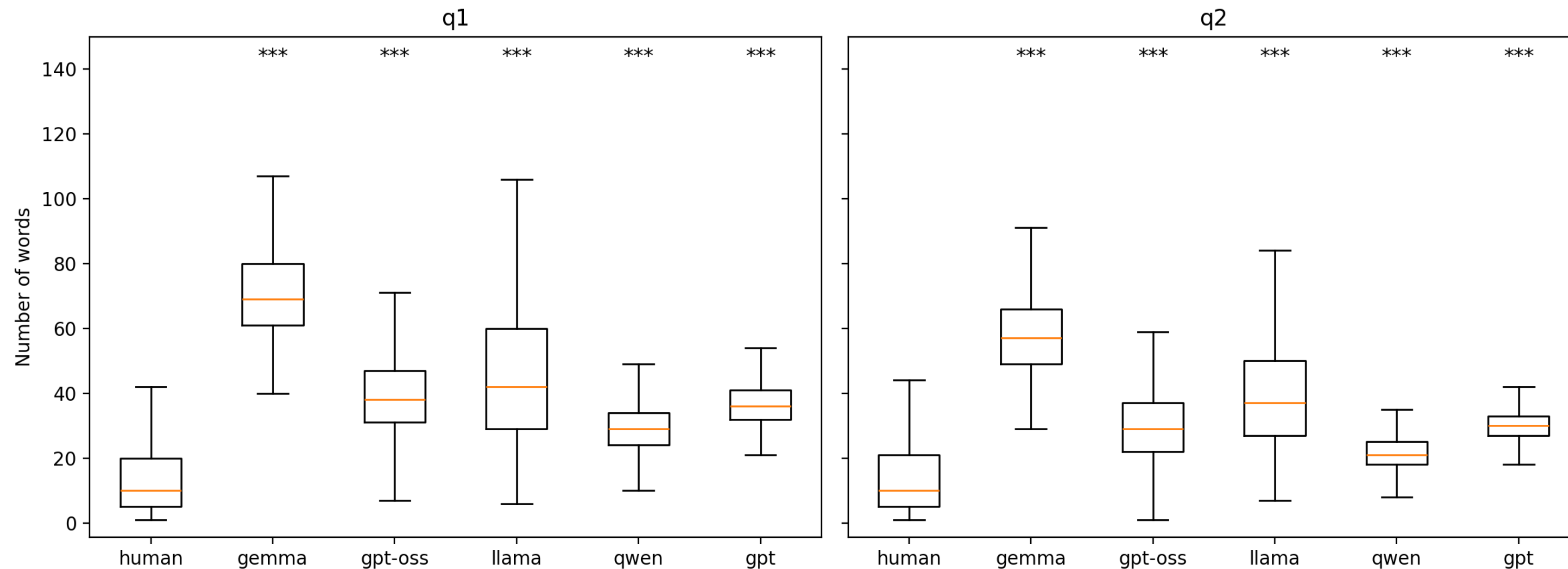
- Generated answers remain semantically separated
 - All models exceed the human split-half baseline
 - Answers are on topic, but not distributionally human-like

Results: sentiment shift



- Models amplify the question's affective orientation
 - Q1: model answers are more positive
 - Q2: model answers are more negative
- Interpreted as textual framing, not underlying affect

Results: response length



- Models are systematically more verbose
 - All models produce longer answers, often as self-contained mini-narratives
 - Humans often answer compactly or elliptically

Results: lexical framing

Q. Human-distinctive lemmas	Model-distinctive lemmas
Q1 docente (teacher), collega (colleague), team, condivisione (sharing), progettazione (planning), rete (network), relazione (relationship), curricolare (general-education), scuola (school), rapporto (relationship)	progetto (project), partecipare (participate), ricordare (remember), supporto (support), alunno (student/pupil), attivamente (actively), grazie (thanks), creare (create), ambiente (environment), bene (well)
Q2 docente (teacher), collega (colleague), curricolare (general-education), sostegno (support), classe (class), poca (little), bambino (child), gruppo (group), confronto (discussion), venire (come)	manca (lack), risorsa (resource), specifico (specific), tempo (time), supporto (support), difficoltà (difficulty), adeguato (adequate), formazione (training), personale (staff), principale (main)

- Situated human accounts vs generic model accounts
 - Humans: actors, roles, relationships
 - Models: institutional and evaluative register

Conclusions

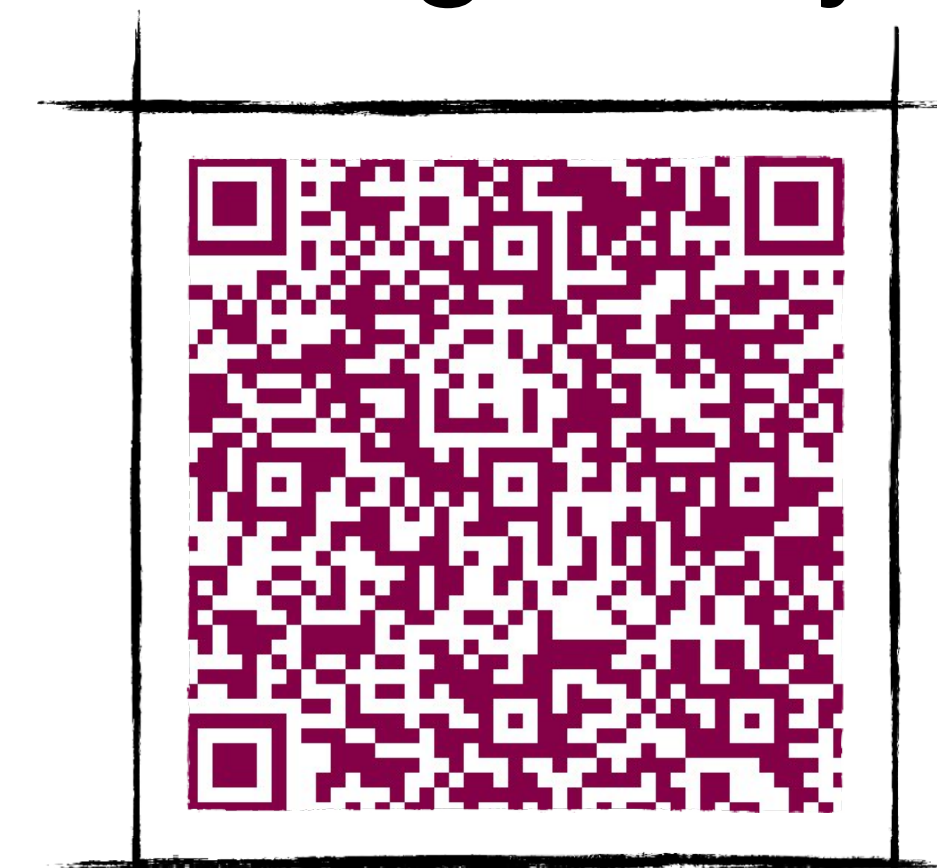
- **Plausible $\neq$ distributionally interchangeable**
 - Generated answers are fluent and on topic
 - But differ in semantic distribution, sentiment framing, length, and vocabulary
 - Validation must be population-, format-, and protocol-specific
- Several limitations: one domain, one language, two questions, one human sample, LLM-based sentiment analysis, one prompting protocol, ...

Preview of follow-up work

- Different formats, different failure modes
 - **Open-ended answers (2)**: semantic mismatch, verbosity, question-dependent sentiment shift
 - **Word associations (14)**: semantic mismatch, reduced lexical diversity, positivity inflation
 - **Likert items (3)**: ordinal distortions, higher skepticism toward full inclusion
- LLMs do not fail as survey respondents in one single way: the **failure mode depends on the question format**.

Preview of follow-up work

- Different formats, different failure modes
 - **Open-ended answers (2)**: semantic mismatch, verbosity, question-dependent sentiment shift
 - **Word associations (14)**: semantic mismatch, reduced lexical diversity, positivity inflation
 - **Likert items (3)**: ordinal distortions, higher skepticism toward full inclusion
- LLMs do not fail as survey respondents in one single way: the **failure mode depends on the question format.**



Data, Code, Analyses

Thank You!