

57.6k h

Synthetic training speech

54.4k h dyadic + 3.2k h multi–party

32.6 min

Average real conversation

104 samples, 56.5 h total, 43.6 min maximum

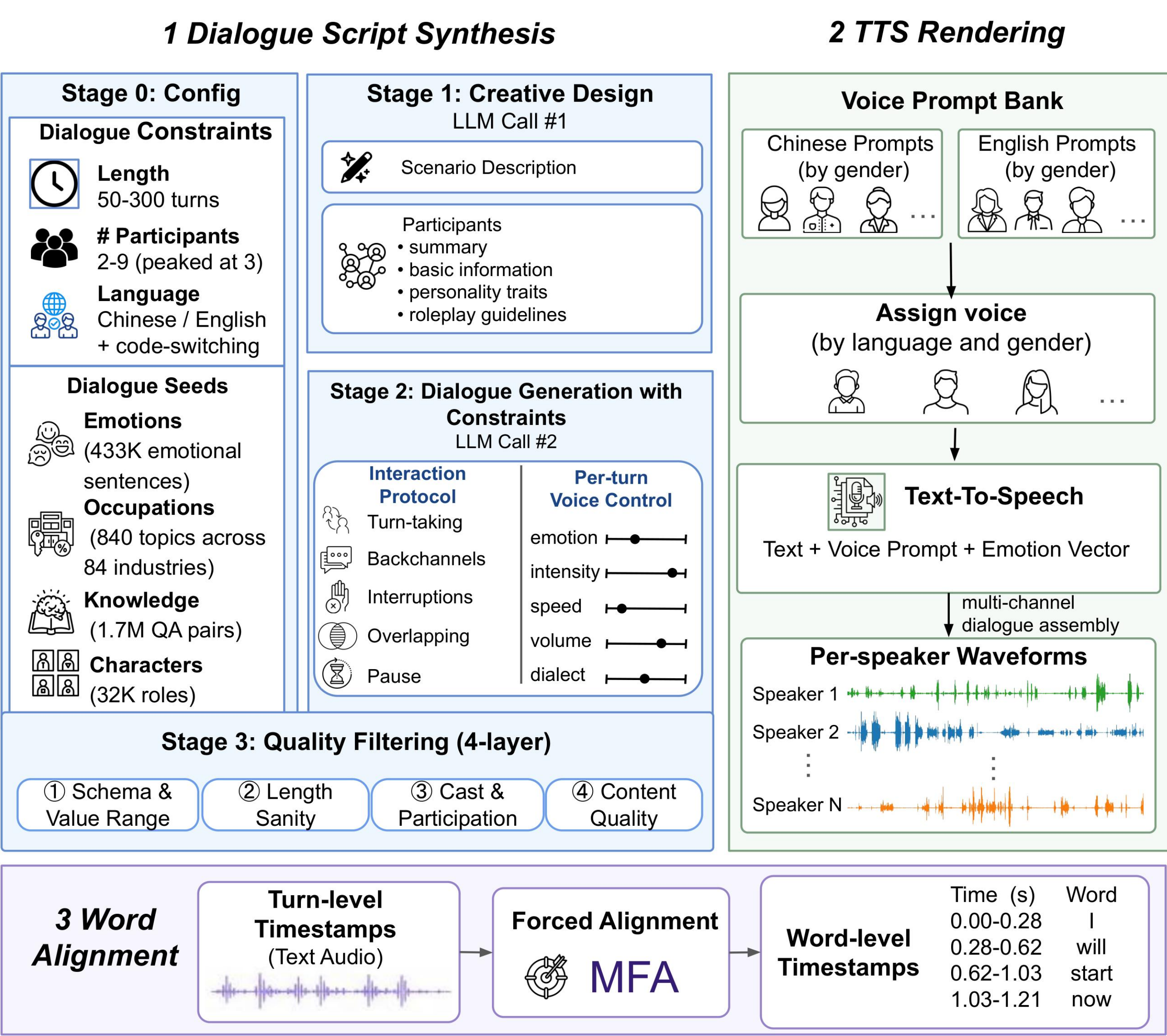
+182.2%

Final Score over Moshiko–7B

4.66 to 13.15 on MultiTalkBench

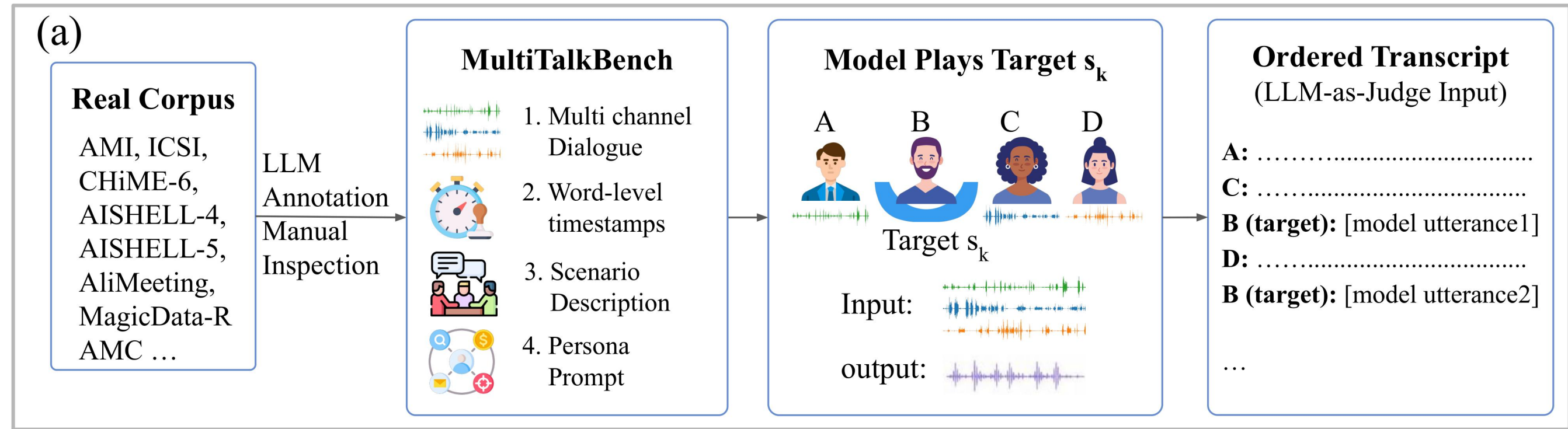
Data engine and training

2–9 participants, 50–300 turns, English/Chinese.
Two–pass scripts, TTS and word alignment.



MultiTalkBench

Real human conversations. The model takes the role of a named meeting participant.



What the benchmark measures

- General
- Clarity, relevance, agenda fit and contribution
- Multi–Party IQ
- Speaker tracking, addressee selection, information integration and group behavior

Role–conditional behavior

Facilitator, driver, collaborator or evaluator

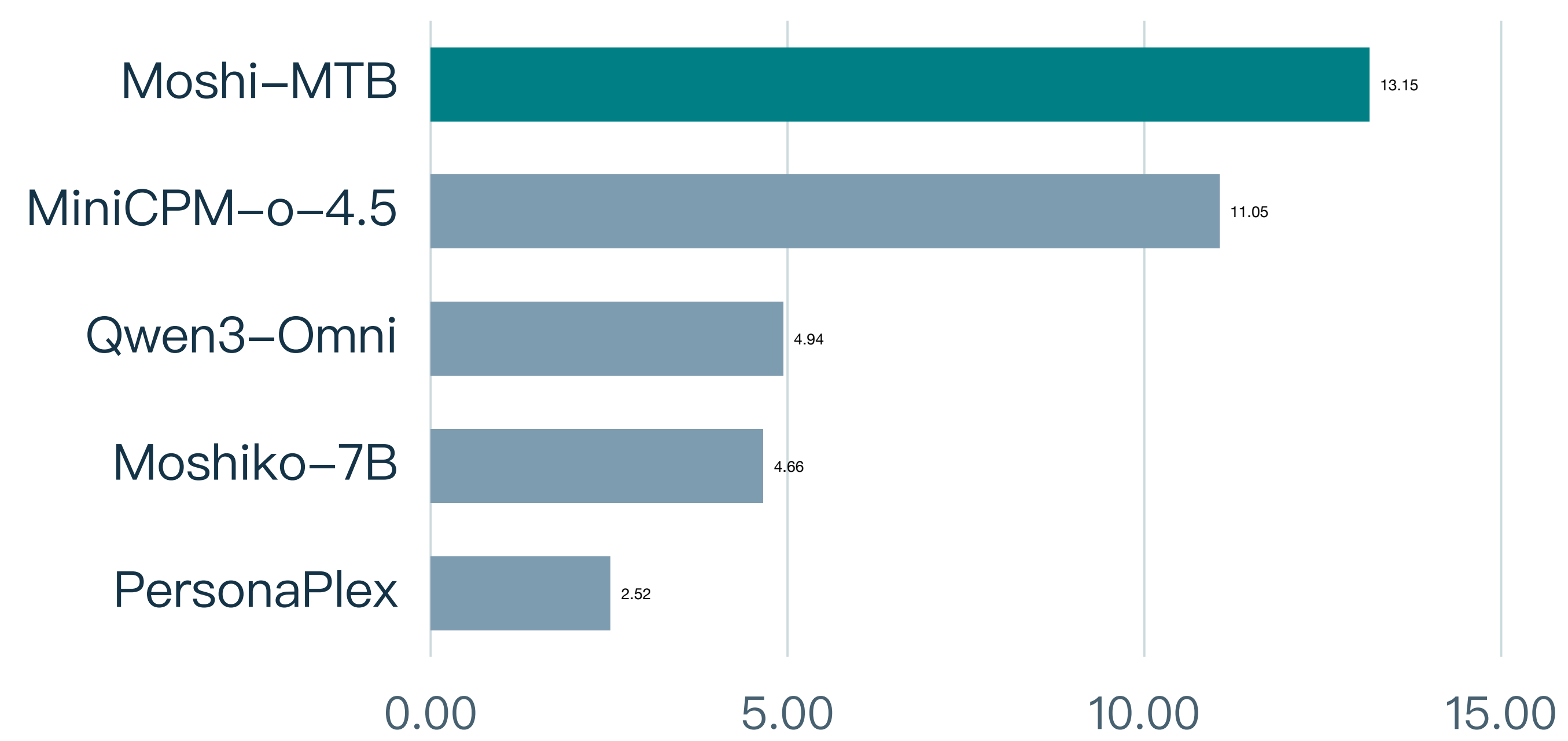
Evaluation protocol

Equal weight for the three content groups.
Participation is a separate span–ratio measure.
Offline judging excludes interaction timing.

Human agreement: Spearman ρ = 0.898
on Final Score (100 reviewed samples).

Results on real conversations

Moshi–MTB leads the four evaluated baselines.
Final Score (higher is better)



Human reference: 68.74

Bilingual and multi–party gains

Metric	Moshiko	Moshi–MTB
English (Final)	7.82	10.87
Chinese (Final)	0.36	16.27
Multi–Party IQ	2.41	8.56
Role–conditional	6.62	12.16

What remains difficult

MiniCPM–o–4.5 leads Multi–Party IQ (9.00 vs. 8.56).
All models remain far below the human reference.