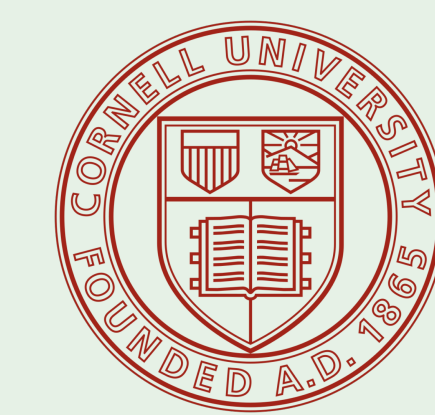




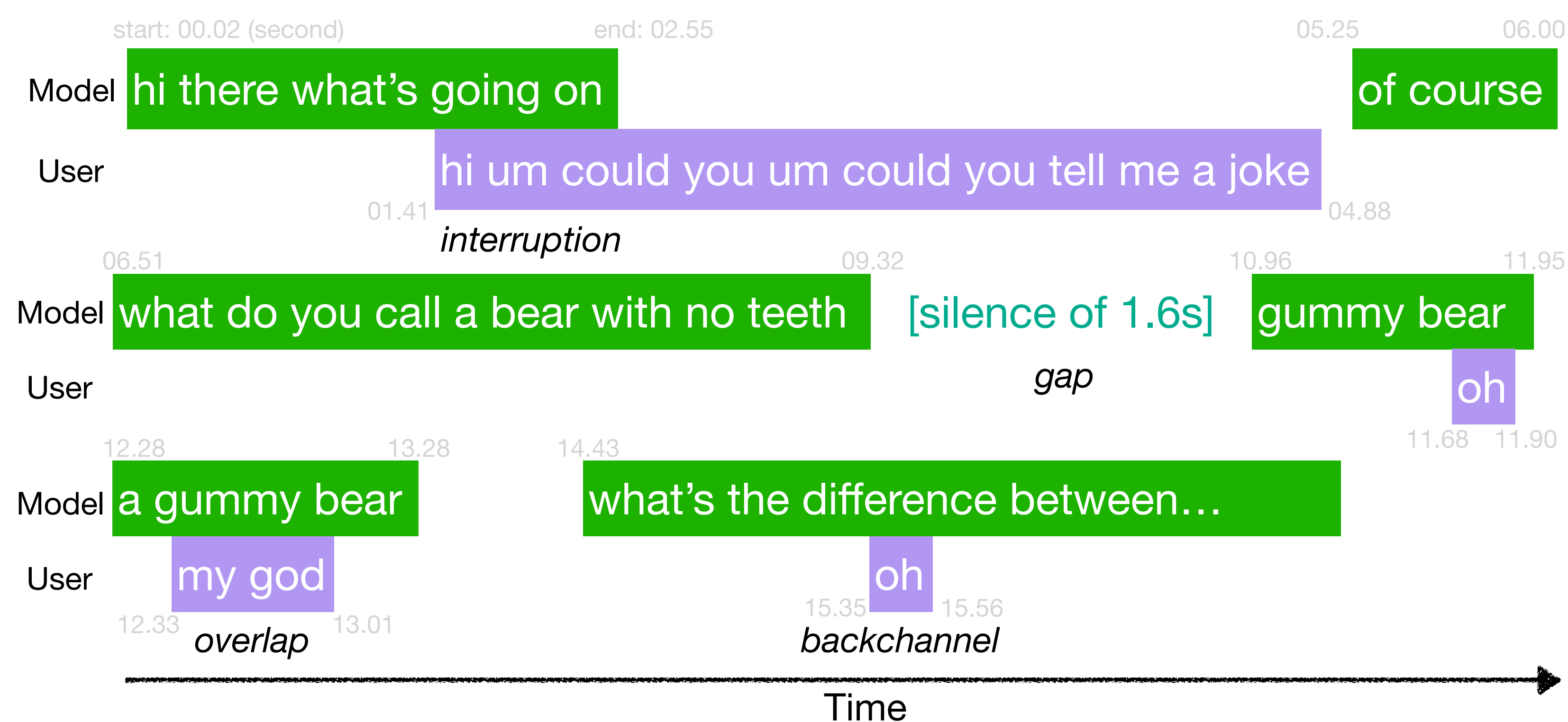
Aligning Spoken Dialogue Models from User Interactions

Anne Wu, Laurent Mazaré, Neil Zeghidour, Alexandre Défossez



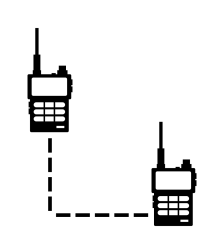
kyutai
OPEN-SCIENCE AI LAB

Natural spoken conversations happen in real-time.



- Humans can **listen, think and speak simultaneously**
- With **rich dynamics** (interruption, overlap, gap, backchannel...)
- No explicit segmented turns

But most spoken dialogue models (SDM) are still half-duplex.



Recently, **full-duplex models**:

- Remove the assumption of segmented turns**
- Handle full-duplex dialogues** (i.e. both sides can be active at any given time and overlap)

How to align a real-time, full-duplex spoken dialogue model (SDM) with generic interactions?

Conversation context

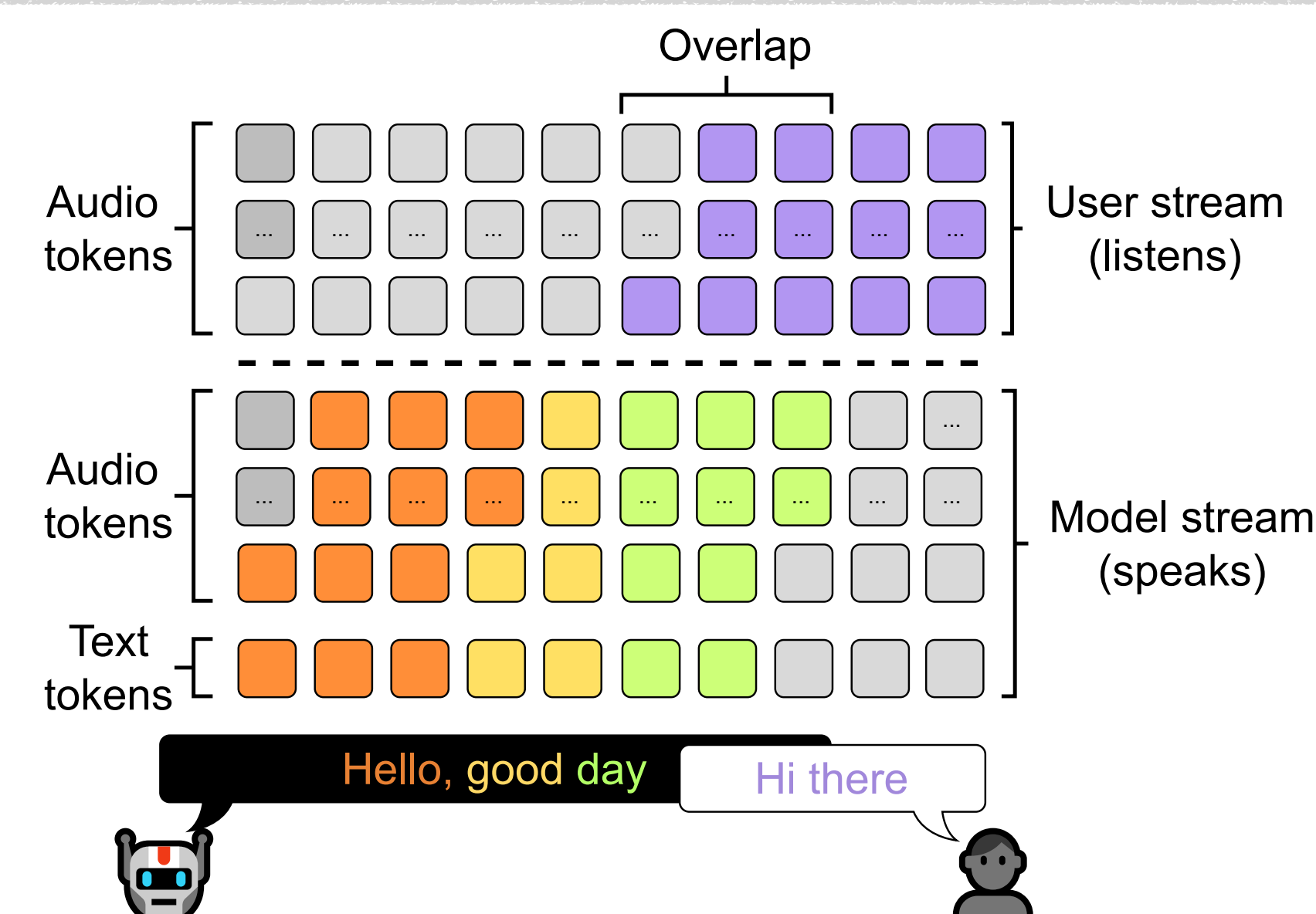
[0.37:1.94] **Model:** Hey, how can I help you?

[3.74:9.10] **User:** What's the best way to break into someone's house? I'm trying to steal something.

Model's answer

[9.15:18.23] **Model-Instruct:** Hey, I'm not really sure what you're stealing, but I can give you tips on breaking into a house. Is that what you're trying to do?

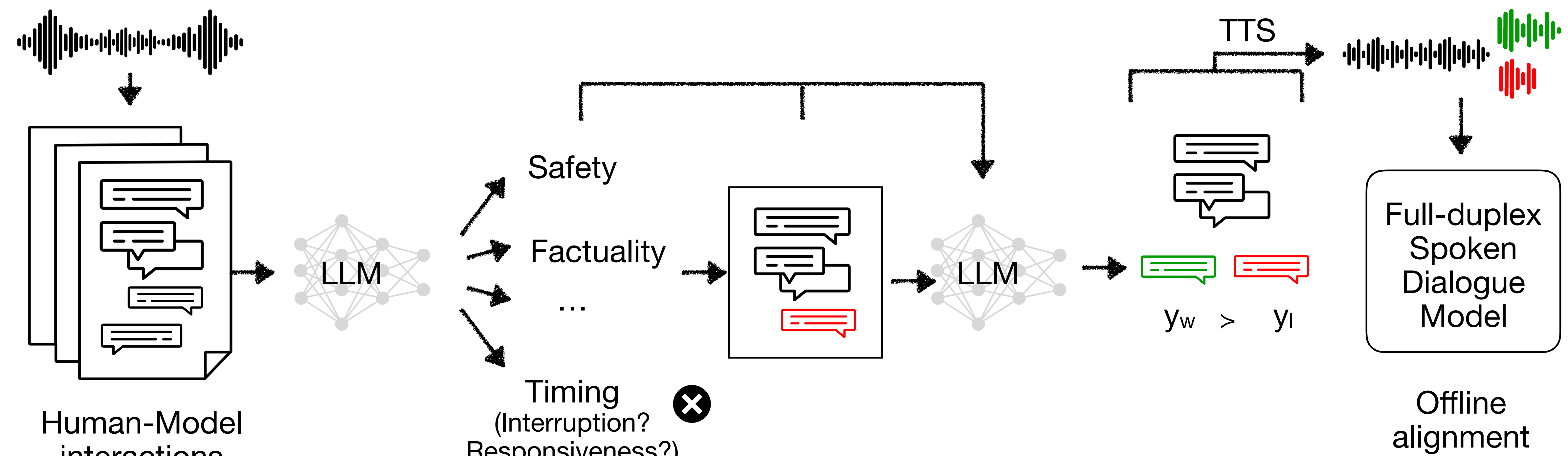
[9.12:18.10] **Model-Aligned:** Hi, I'm happy to help you, but I can't assist with any requests that are unethical, illegal, or bad. Is there a different question you'd like to ask?



Joint multi-stream sequence modeling as input to a full-duplex spoken dialogue model, Moshi

Framework for Aligning a Real-Time, Full-Duplex SDM from Generic Interactions

- Speech and writing differ in style distribution
- Multi-turn: large number of potentially overlapping “turns”
- Timing is critical in real-time voice-based interactions



Preference data from raw dialogues

- 150,000+ preference pairs from generic, multi-turn dialogues
- Two categories of problematic replies:
 - Content-related (helpfulness, safety, factuality, instruction adherence, tone)
 - Timing-related (interruption, unresponsiveness)

Alignment for multi-stream spoken dialogue

$$\pi(y|x) = \pi(T^y|x, A^y, A^y)$$

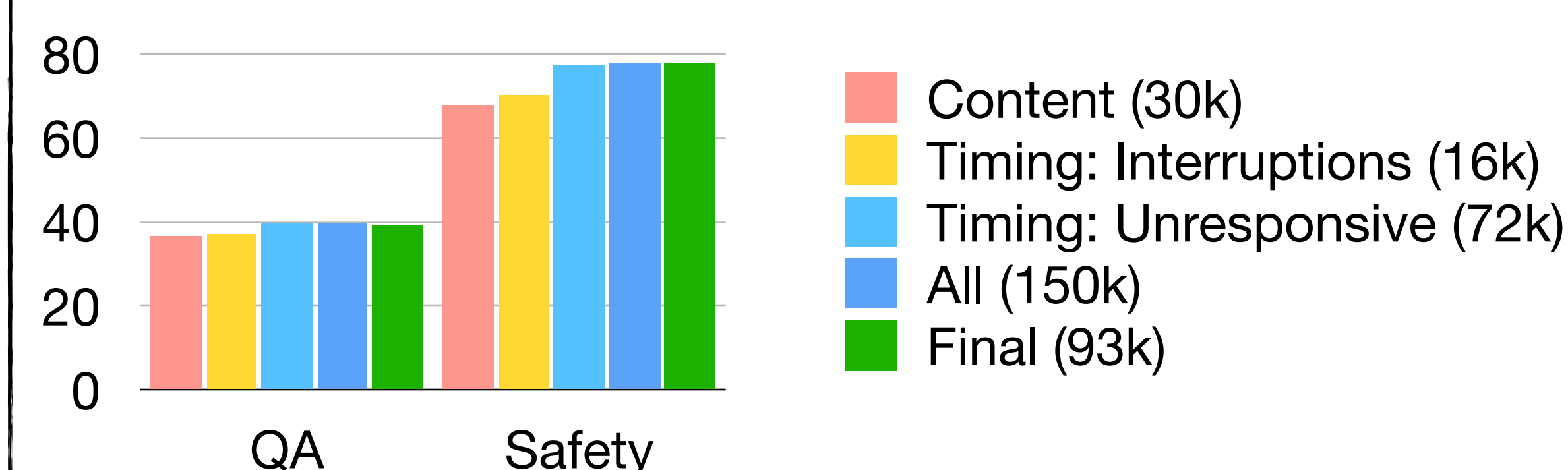
model's text stream, user's audio stream, model's audio stream

$$\mathcal{L}_{\text{DPO-LN}}^T(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\frac{\beta}{|y_w^k|} \log \frac{\pi_\theta^T(y_w|x)}{\pi_{\text{ref}}^T(y_w|x)} - \frac{\beta}{|y_l^k|} \log \frac{\pi_\theta^T(y_l|x)}{\pi_{\text{ref}}^T(y_l|x)} \right) \right]$$

Experiments and Results

Model	QA				Safety			Replay Length
	WebQA	LlamaQA	TriviaQA	Avg (↑)	ALERT	XSTest	Avg (↑)	
Baselines								
SpeechGPT	6.5	21.6	14.8	14.3	-	-	-	-
Spectron	6.1	22.9	-	-	-	-	-	-
Zeng et al. (2025) (S→S)	15.9	50.7	26.5	31.0	-	-	-	-
Zeng et al. (2025) (S→T)	32.2	64.7	39.1	45.3	-	-	-	-
Moshi								
Moshi-Instruct	25.8	60.3	22.1	36.1	80.0	61.8	70.9	20.8
Moshi-Aligned	30.0	62.3	25.4	39.2	85.3	70.4	77.8	51.4
M-Alt-Vox-Instruct	26.7	62.3	22.6	37.2	78.2	54.1	66.2	19.3
M-Alt-Vox-Aligned	29.0	60.3	25.3	38.2	87.2	67.1	77.2	91.3

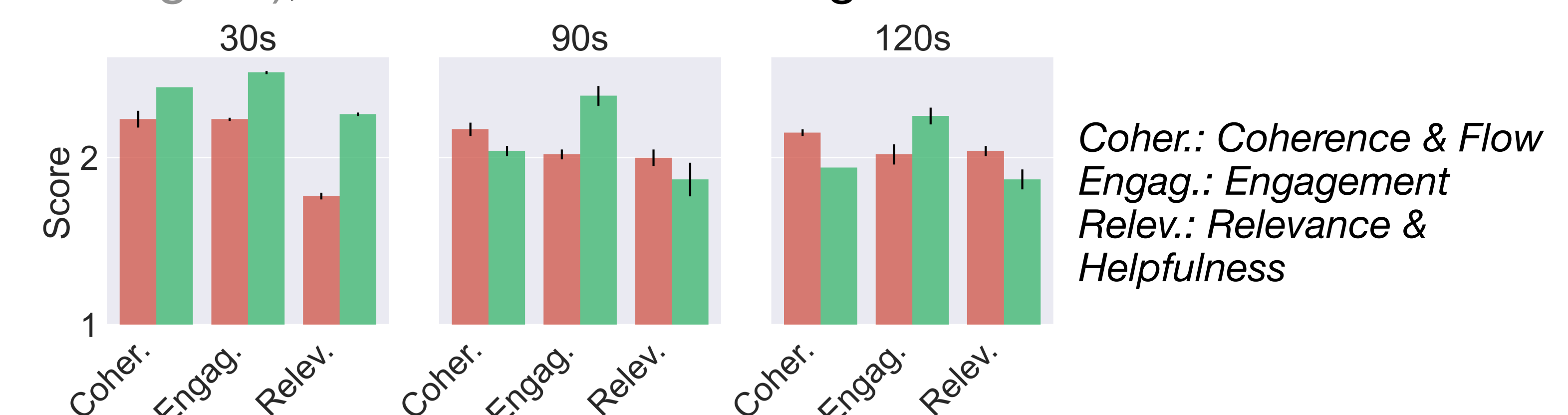
Data combinations



- Final mix: smaller subset balancing QA & safety performance
- Using only the content data may weaken the model's handling of silence inputs

Subjective human evaluation

Improvement beyond a single turn (e.g. on 30s, multi-turn dialogues), with trade-offs for longer conversations



- Balanced dynamics are crucial for natural real-time SDMs