SIGDIAL 2026 · POSTER #39

Ping-Ponder: Concurrent Dual-Agent Spoken Dialogue via Shared Belief State

Akiko Masaki-Kato **Haris Gulzar** NTT, Inc. — Musashino R&D Center, Tokyo, Japan · akiko.masaki@ntt.com

1 Responsiveness vs. Deliberation

- Spoken dialogue: silences beyond ~1 s feel unnatural (Skantze, 2017).
- LLM reasoning + tool use takes several seconds per turn.
- Sequential dual-agent systems (Talker-Reasoner, Chat-Supervisor) go silent exactly when planning is most visible.

7.1–7.3 s blocking silence per reasoning turn in the sequential baseline

2 Key Idea: Shared Belief State

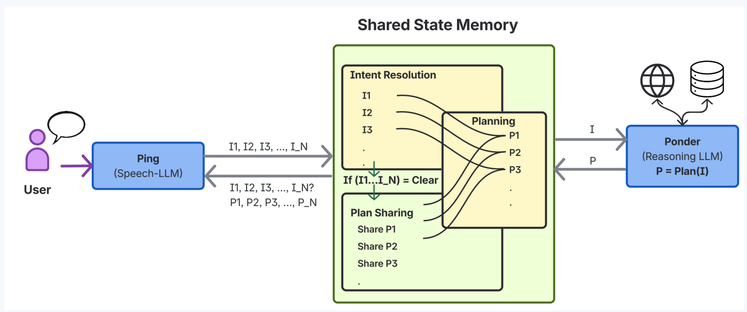

Fig. 1 — Ping writes intent slots; Ponder plans from partial intent and writes plans back.

- **Ping (System 1, Speech-LLM):** streams ASR/TTS, writes intent slots incrementally. Never waits.
- **Ponder (System 2, Reasoning LLM):** plans from partial intent — commits partial plans early.
- **Belief state B = {slots, status, plan drafts}:** a bidirectional pub/sub channel, not a per-turn DST snapshot.

3 Concurrency Design

- **Partitioned single-writer:** Ping owns slots; Ponder owns plans → no central arbiter.
- **Version counters:** stale plans are detected and re-planned.
- **Speculative prefetch:** Ponder runs against the anticipated next turn, off-turn.

4 Worked Example (App. B)

User *"I want to plan a vacation to Europe this summer. My budget is $3000."*
Ping *"For summer I'd suggest Italy, France or Switzerland. How long is your trip?"*

Slots	
destination	**Europe**
when	**Summer**
budget	**$3000**
duration	
people	

Plan (Ponder) countries: **Italy, France, Swiss**

Ponder ⇄ popular_countries(season, budget)

User *"Italy and France are good — I plan for 10 days."*
Ping *"I'd suggest 5 days in Italy, then 5 in France. How many people?"*

Slots	
destination	Europe
when	Summer
budget	$3000
duration	**10 days**
people	

Plan (Ponder) itinerary: **Italy 5 d + France 5 d**

Ponder ⇄ duration_recommendation(countries)

User *"Solo. Can you give me a complete plan?"*
Ping *"Italy: Colosseum, Pisa, risotto → France: Louvre, beef bourguignon."*

Slots	
destination	Europe
when	Summer
budget	$3000
duration	10 days
people	**1**

Plan (Ponder) **attractions + food per country → plan complete**

Ponder ⇄ attractions(·), food_recommend(·)

Slots fill up while planning proceeds — every plan fragment Ping verbalises comes from the shared state; the user never waits on a blocking call.

5 Experimental Setup

- **Two regimes:** live paired case study (single operator, EN + JA) + large-scale replay of logged user turns.
- **JA:** in-house spoken TOD corpus — 1,200 dialogues, 6 domains, 39 voice actors (App. C). Eval: 120 dialogues/cond., 8,765 judged turns.
- **EN:** SpokenWOZ, single-domain — 106 dialogues/cond., 5,067 judged turns.
- **Judge:** gpt-4.1 vs. gold human turn — Appr. / Faith. / Flu. (0–3), disfluency-aware rubric.

Condition	Supervisor	Coordination
Baseline	NL, blocking	serial, after chat
No-prefetch	{slots, prop.}	serial, before chat
Ping-Ponder	**{slots, prop.}**	**background prefetch**
Single-agent	none	—

One metric (App. B.2): planning latency = heavy-model call start → output available. Identical in every condition.

6 Live Case Study (N = 1)

Planning: EN 7.11 → 0.49 s (−93%) · JA 7.31 → 0.96 s (−87%)

Condition	Lang	Intent (s)	Plan (s)
Baseline	EN	0.42	7.11
Ping-Ponder	**EN**	**0.45**	**0.49**
Baseline	JA	0.70	7.31
Ping-Ponder	**JA**	**0.75**	**0.96**

Table 1 — paired single-operator session: an illustrative headline, not a statistical claim. The controlled replay study (→ ⑦) confirms the ordering at scale.

7 Replay Study at Scale

−31% JA planning: 1376 → 950 ms

−32% EN planning: 1029 → 703 ms

Time to first response (ms) — Table 2

Lang	Baseline	No-pref.	Ping-Pon.	Single
JA	1576 ± 99	2803 ± 121	**1526 ± 67**	1542 ± 70
EN	886 ± 50	1547 ± 81	**891 ± 43**	999 ± 68

- Ping-Ponder ≈ Baseline (−50 ms JA / +5 ms EN): prefetch absorbs the Ponder call.
- No-prefetch pays +1227 ms JA / +661 ms EN.

Planning latency, fired turns (ms) — Table 3

Lang	Condition	Planning (ms)	N
JA	Baseline	1376 ± 71	384
JA	**Ping-Ponder**	**950 ± 33 (−31%)**	**526**
EN	Baseline	1029 ± 30	255
EN	**Ping-Ponder**	**703 ± 38 (−32%)**	**25**

EN and JA agree to within one point. Fire-event definitions differ across conditions; the latency column is directly comparable.

8 Gains Land on Planning Turns

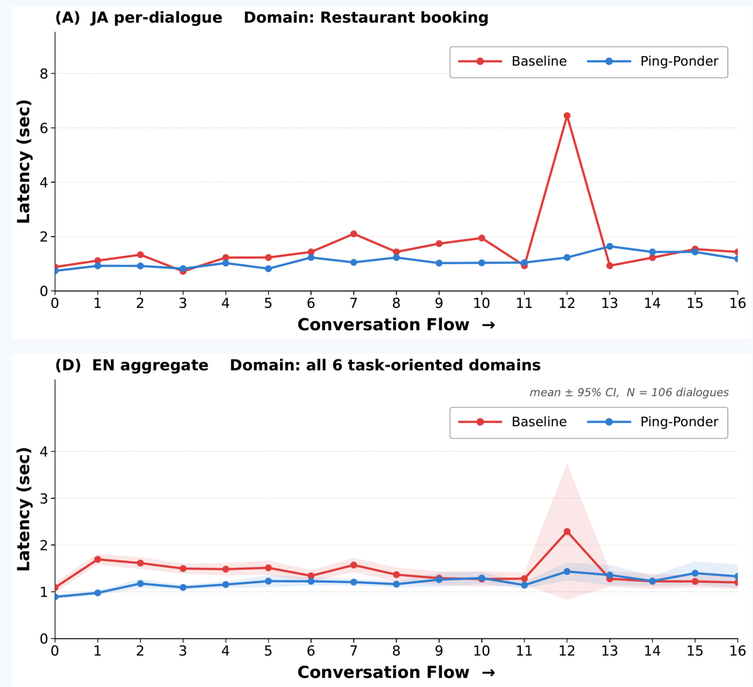

Fig. 2 — Baseline spikes to 6.5 s / 4.9 s on planning turns; Ping-Ponder stays ≈ 1 s.

Lang	Turns	Baseline	Ping-Pon.	Δ
JA	**fired**	**3257**	**2122**	**−35%**
JA	other	2106	2001	−5%
EN	**fired**	**2342**	**1176**	**−50%**
EN	other	1208	1166	−3%

Table 5 (ms) — fired-turn reduction is ≈7× (JA) / ≈17× (EN) the non-fired one.

9 Quality: No Degradation

	JA (8,765 turns)			EN (5,067 turns)		
LLM judge (0–3)	**Ap.**	**Fa.**	**Fl.**	**Ap.**	**Fa.**	**Fl.**
Baseline	2.52	1.97	2.89	2.21	1.60	2.96
No-prefetch	2.64	2.09	2.89	2.44	1.79	2.94
Ping-Ponder	**2.73**	**2.12**	**2.96**	**2.40**	**1.75**	**2.96**
Single-agent	2.65	2.09	2.99	2.45	1.78	3.00

- JA: Ping-Ponder tops Baseline on all three axes; EN: variants sit within judge noise → no speed–quality trade-off.
- **Blind human check (120 turns, App. A.1):** faithfulness ρ = 0.66; human ranking matches the judge's — with sharper separation:

Appr. (mean)	Base	No-pre.	P-P	Single
Human (blind)	1.20	1.83	**2.57**	2.70
LLM judge	2.13	2.53	**2.57**	2.57

Two LLM judges agree more with each other (ρ = 0.82) than with the human — human validation substantiates the claim.

10 Takeaways

The bottleneck is scheduling, not model speed — a versioned shared belief state decouples the two.

- **Limits:** N=1 live study; LLM-judged quality; EN/JA only; proprietary APIs (protocol is vendor-agnostic).
- **Next:** edge deployment; persistent user model; a third "critic" agent.
- **Release:** replay JSONLs, judge prompts, per-turn scores.

References

[1] G. Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. SIGDIAL 2017.
[2] K. Christakopoulou et al. Agents Thinking Fast and Slow: A Talker-Reasoner Architecture. NeurIPS 2024 Wksp. on Open-World Agents (arXiv:2410.08328).
[3] OpenAI. Realtime Agents — Pattern 1: Chat-Supervisor (our baseline). github.com/openai/openai-realtime-agents, 2024.
[4] S. Si et al. SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents. NeurIPS 2023.