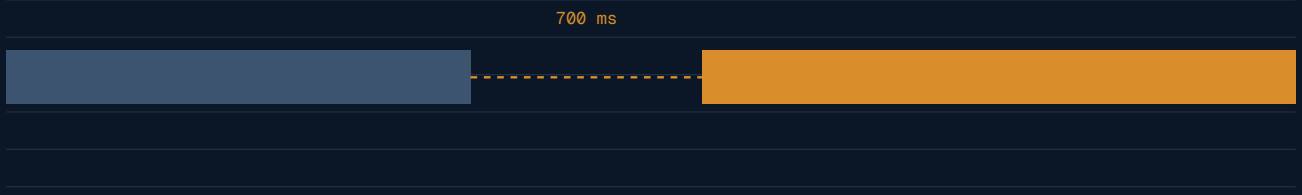
PHYSICAL AI FRONTIERS

ROBOTICS TRACK · EDITION R05 · JULY 2026

The Pause That Sounds Like Thinking



Conversational timing in social robots — why human turn transitions average 208 milliseconds, why one laboratory study preferred about 700 milliseconds for a robot, and how endpointing shapes the delay users actually experience.

Lise DELOBEL

GenAI project manager & consultant — Physical AI, digital twins, and applied ML for industrial and extreme-environment operations. MIT IDSS Data Science & Machine Learning certificate.

LISEDELOBEL.FR · LISE.DELOBEL@ME.COM

Faster is not the target

The argument in one line: in conversational robotics the design target is not minimum delay but an appropriate and predictable turn transition — and a large, often unexamined share of it is spent deciding that the person has finished.

Human conversation runs on a remarkably tight clock. Across ten languages drawn from indigenous communities to major world languages, the mean transition between conversational turns is 208 milliseconds, with each language-specific mean falling within about 250 milliseconds either side of that figure. Given that, the intuitive engineering target for a talking robot is obvious: get as close to 208 milliseconds as possible.

That intuition deserves testing rather than inheriting, and when it was tested the answer came back very different. It is also harder to reach than it looks, for reasons that have little to do with model speed. Note that these are answers to two different questions: one describes how people coordinate with people, the other estimates an acceptable response point for a particular robot task.

FINDING 01

The preferred delay is far slower than human

In a psychophysical study, 210 participants judged robot responses at fifteen delays from 100 to 1500 ms. The preferred delay was approximately 700 ms — more than three times the human turn gap — and the estimated preference point did not vary significantly with communication style or robot visibility in that study.

FINDING 02

Between humans, that same pause is long

Conversation analysis treats a 700 ms gap between people as lengthy, and longer-than-expected gaps may be read as hesitation, dispreference or uncertainty depending on question and context. People evidently apply different expectations to machines — which makes human conversational norms a reference point, not a specification.

FINDING 03

Much of the budget is spent before generation begins

Endpointing — deciding the user has finished — is often one of the largest controllable terms in a well-tuned streaming pipeline, particularly where a fixed trailing-silence threshold is used. A 700 ms threshold adds 700 ms to every turn before the system commits the turn and launches response generation.

FINDING 04

Unbounded context stretches the latency tail

In published practitioner traces, time to first token rises sharply from median to 95th percentile as carried-forward context grows. For a companion robot this matters: history is what makes it feel like a companion, and unless the architecture retrieves selectively, summarises and caches, it is also what lengthens the tail. Stored memory and prompt context are not the same thing.

Two definitions used throughout

Endpointing is the decision that the speaker has finished their turn; a fixed-threshold implementation waits for a set duration of silence, while a predictive one estimates turn completion from prosody and syntax. Time to first audio (TTFA) is the interval from the end of user speech to the first sound the robot emits — the only latency the person actually experiences, and not the same as the time to produce a complete answer.

Timing became the binding constraint

Conversational quality stopped being limited by what a robot could say. It is now limited by when it says it.

208 ms
HUMAN TURN GAP
Mean response offset across ten languages, all within ~250 ms of that figure

~700 ms
PREFERRED, ONE STUDY
Point of subjective equality across 210 participants and 15 tested delays

4×
TTFT TAIL GROWTH
Reported rise in time to first token from median to 95th percentile as history grows

01 Language stopped being the bottleneck

For bounded interactions, current systems produce substantially more coherent dialogue than earlier generations. Timing and turn management have therefore become bottlenecks alongside language quality rather than after it — the hesitation before a reply starts, the failure to stop when interrupted, the pause that lands as absence.

02 The pipeline became measurable

Production voice stacks now instrument each stage separately — endpointing, transcription, first token, first audio chunk, transport. That decomposition frequently reveals something uncomfortable: a large share of perceived latency comes from an avoidable waiting policy rather than computation.

03 Predictive turn-taking became practical

Rather than waiting for a fixed silence, systems can estimate when a turn is complete from falling intonation and completed clauses. A 2026 endpoint-anticipation system reports forecasting turn ends up to 2.56 seconds ahead and cutting average latency by 505 ms through speculative execution, at the cost of a 28.4% increase in speculative computation.

04 Companion robotics raises the stakes

In a transactional voice agent, latency is an annoyance. In a companion or care setting, timing is read as an emotional signal — attention, interest, presence. The same delay that makes a booking agent feel slow makes a companion feel absent.

Sources: Stivers et al., PNAS 106(26), 2009 · International Journal of Social Robotics, 2026 · “Endpoint Anticipation for Low-Latency Spoken Dialogue,” arXiv:2606.13450 · practitioner latency-budget documentation, 2026. Pipeline figures are practitioner-reported illustrative traces, not cross-platform benchmarks.

Three reference zones, three evidence bases

Silence in conversation is not empty. These are reference points drawn from different kinds of study — they are not a single validated scale, and should not be read as one.

PAUSE 01

~200 ms — human-human cross-language mean

The interval people actually use with each other. Transitions this short are generally understood to require anticipatory planning: listeners use grammatical, prosodic and pragmatic cues to project when a turn may end and begin preparing before it does.

PAUSE 02

~700 ms — preferred point in one robot video experiment

Participants were most likely to judge responses near this point as neither too fast nor too slow. The study did not establish that the delay was interpreted specifically as deliberation or thought — that reading is this paper's hypothesis, not the study's finding.

PAUSE 03

Beyond ~1 s — practitioner warning zone

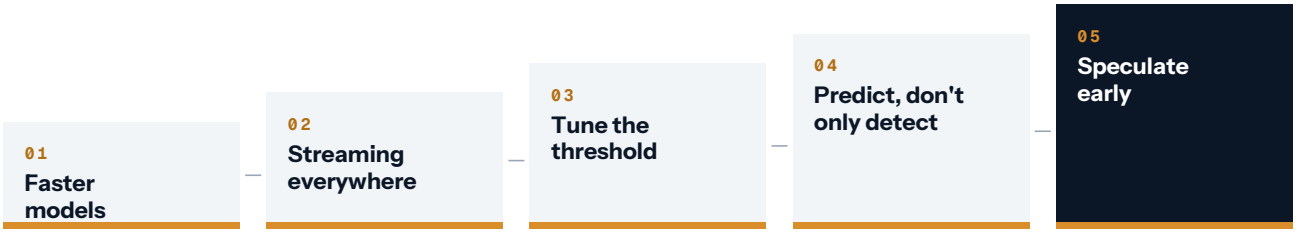
Practitioner guidance converges on users noticing above one second and assuming failure above two. This is stack- and context-dependent rather than a measured threshold, and in an embodied setting the interpretation becomes social as well as technical.

PAUSE 04

Filled versus empty — an untested hypothesis

Embodied cues may make a pause more legible by signalling continued engagement, and embodiment offers a channel a voice-only agent lacks. The effect has not been validated at conversational timescales and should be tested for the specific robot, task and population.

THE LADDER OF RESPONSES



Swap in a quicker model. Real, bounded, and leaves the waiting policy untouched.

Overlap the stages so latency resembles the slowest one rather than their sum.

Shorten the silence window to the smallest value that does not cut people off.

Add turn-completion modelling on top of endpoint detection, not instead of it.

Begin reversible downstream work as turn-end probability rises; cancel if speech continues.

The strategic read

Almost every team starts at rung one, because model choice is the most legible lever and the easiest to procure. The larger gains usually sit further along, where the question stops being how fast the system computes and becomes when it decides to start. These combine rather than replace each other: production systems run voice activity detection, acoustic endpointing, semantic cues, turn prediction and barge-in together. A further move costs no compute at all — designing what fills the remaining pause, which changes what the delay means rather than how long it lasts.

Measuring what people actually prefer

A psychophysical study of robot response delay — 210 participants, 15 delays

Most latency targets in conversational AI are inherited assumptions: humans respond in roughly 200 milliseconds, therefore robots should too. A recent study tested the assumption directly rather than reasoning from it, using a psychophysical method to estimate the point at which responses were equally likely to be judged too fast or too slow.

THE EXPERIMENT

The method

210 participants watched 150 videos in which a person asked a yes/no question and a robot responded after one of fifteen delays between 100 and 1500 milliseconds, while exhibiting one of four communication styles — authoritarian, submissive, neutral or childish — or hidden behind a curtain as a control. For each, participants judged only whether the response was too fast or too slow.

THE FINDING

The result

The preferred response delay was approximately 700 milliseconds, and the estimated preference point did not vary significantly with communication style or with whether the robot was visible. Submissive and childish styles produced significantly flatter psychometric curves — the preference point did not move, but tolerance for deviation widened. The authors note a previous study had identified one second as the preferred robot response time.

THE LIMIT

What a second study did not find

In a follow-up, 420 participants rated robots across four styles at 200, 700 and 1500 ms. Response delay had no effect on the measured dimensions, while communication style significantly influenced perception — the authoritarian style scored lower on attractiveness, enjoyment and sociability and higher on anxiety. Participants detected the delay variation without it registering on the measured social dimensions — timing judgements and personality judgements were separable in this video task.

So what

Two implications, one narrow and one broad. Narrowly: a 700 ms budget is roughly three times more achievable than a 200 ms one, which changes what architecture is required — though this is one laboratory task using video stimuli and yes/no questions, not a general specification. Broadly: this is an argument for testing inherited targets. The 200 ms figure was real, correctly cited, and about humans talking to humans — which turned out not to be the same question.

Source: “Talking to Robots: Does a Robot’s Personality Change How we Feel About its Response Delay?” International Journal of Social Robotics, 2026. Laboratory study using video stimuli rather than live interaction.

Where the milliseconds actually go

Decomposing the pipeline, and the term nobody budgets for

Once the target is known, the question becomes whether it is reachable. Decomposing a production voice pipeline produces a consistent and counterintuitive answer: the dominant term is not a computation.

THE BUDGET

The stages, and their share

In published practitioner traces, stage budgets for a well-tuned streaming pipeline fall roughly at: transport 30–80 ms, endpointing and turn-taking 150–300 ms, model time to first token 150–400 ms, synthesis to first audio 100–200 ms. One illustrative configuration lands near 630 ms to first audio. These are ranges from different stacks rather than one benchmark, and because stages overlap under streaming they cannot simply be summed.

A LEADING TERM

Endpointing dominates

Endpointing is often one of the largest controllable terms, especially where a fixed trailing-silence threshold is used. The mechanism is stark: waiting 700 ms of silence before declaring the turn over adds 700 ms to every turn. Trailing windows above 400 ms are cited as a common source of tail latency, with 500–800 ms defaults typical. In less optimised stacks, recognition, synthesis, tool calls or transport may dominate instead.

THE TAIL

The tail is worse than the median

Median latency flatters the system. In one published production trace, time to first token climbed from around 566 ms at the median to roughly 2,246 ms at the 95th percentile as conversation history grew. Tail percentiles matter because occasional very slow turns disproportionately shape perception; report them alongside the median and alongside interruption rate.

So what

The uncomfortable implication for a companion robot is that its most valuable property is also its Accumulated history is what lets it recall what you said last week; carried forward wholesale into the prompt, it is also what stretches the tail. The distinction that matters is between stored memory and active context — a robot can hold a great deal of the first while sending little of the second, through selective retrieval, summarisation and caching. Treat context as a latency budget line, not only as a capability.

Sources: practitioner latency-budget documentation, 2026, including published stage breakdowns and production trace figures. These are illustrative traces from named practitioner sources rather than cross-platform benchmarks; hardware, region, model, context size and whether stages overlap all vary and are not always stated.

When the pause is the message

Timing as a social signal in companion and care settings

In transactional voice AI, latency is a usability metric. In companion robotics it is closer to a behavioural one, because people read timing as intent — and they do so whether or not the designer intended it.

THE PARADOX

The same interval means opposite things

Conversation analysis classes a 700 ms gap between humans as lengthy, with prolonged latency associated with uncertainty or deception. Yet 700 ms is the measured optimum for a robot. The resolution is that people apply different expectations to machines — which means human conversational norms are a starting reference, not a specification.

THE EVIDENCE

Delay changes behaviour, not only opinion

In a study of turn-taking with a humanoid during a game, participants spent significantly more time gazing at the robot's face and hands during delay periods, and increased face gaze was associated with lower fluency ratings — timing problems surfacing in behaviour before they surface in self-report. Important caveat: those delays were 4 and 10 seconds and concerned robot movement, not sub-second speech. It evidences monitoring during noticeable delay; it does not validate a 700 ms speech target.

THE HARD CASE

Group settings amplify everything

A study of multi-party open-ended conversation with a social robot found the setting remained challenging, particularly under speaker overlap, rapid turn changes, response relevance and system latency. Adding participants multiplies the turn-taking decisions without adding time to make them.

So what

If timing is read as attention, then evaluating a companion robot on task success and response quality alone misses the dimension users are most sensitive to. The measurable proxies already exist — gaze during pauses, overlap rate, interruption rate, turn-transition offset distribution — and they are cheap to instrument compared with the qualitative studies most programmes commission instead.

Sources: Scientific Reports 15, 2025, gaze and movement adaptation during delayed robotic movement · multi-party open-ended conversation with a social robot, arXiv:2503.15496 · CHI 2026 conversation-analysis study of human-agent latency in VR.

The budget is spent before you compute

Endpointing in a conversational pipeline — the largest term is a wait

THE CHALLENGE

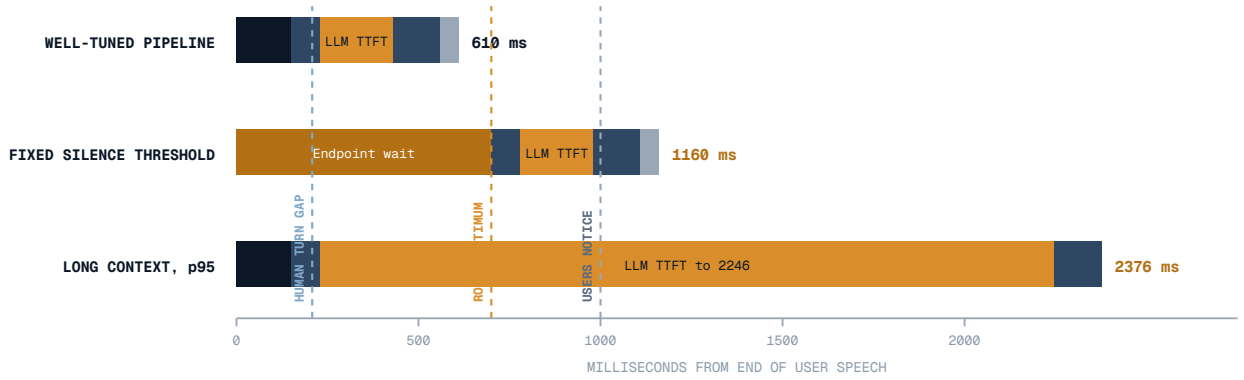
A robot cannot commit to answering until it believes the person has finished. The conventional mechanism is a silence threshold: detect that the audio has gone quiet, wait a fixed interval to be sure, then commit the turn and launch response generation. Streaming transcription may already be running throughout — but that waiting interval buys no progress toward a reply, and it is charged to every single turn.

The arithmetic is unforgiving. The preferred delay measured in the cited study is around 700 milliseconds, and typical implementations default to trailing silence windows of 500 to 800 milliseconds. A 700-millisecond threshold can therefore consume the whole post-utterance budget before the turn is even committed. Every downstream optimisation is then competing for time already spent.

Why the obvious solutions fail

Shortening the threshold trades one failure for a worse one. A short window makes the robot quick but risks cutting the person off during a natural pause for breath or thought — and in companion and care interactions premature interruption may be particularly damaging to perceived respect and inclusion — though the relative cost of interrupting versus waiting should be measured for the intended population rather than assumed, since in urgent contexts waiting is worse. Faster models help less than expected: once recognition and synthesis are fully streaming and colocated, endpointing and generation become the more prominent terms. The threshold is a guess about human behaviour dressed as a configuration parameter.

FIG. 01 — ILLUSTRATIVE LATENCY BUDGETS; NOT A CROSS-PLATFORM BENCHMARK



1 Bars: illustrative practitioner budgets
2 A fixed threshold spends the budget first
3 TTFT bar ends at 2246; +TTS to 2376
4 Blue line: measured study reference
5 Amber line: preferred point, one study
6 Grey line: practitioner design threshold

Predict the ending instead of waiting for it

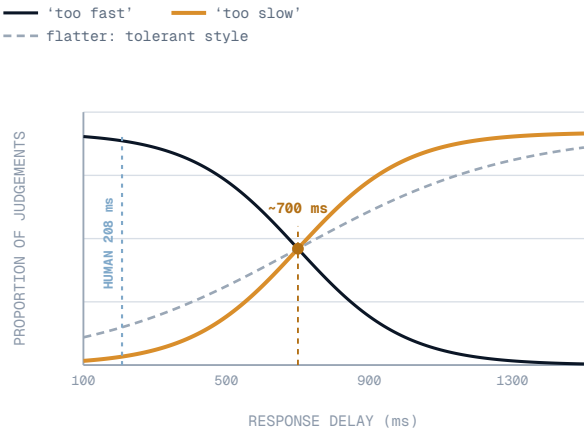
The design move, the trade-off accepted to ship, and what the target actually is

THE RESOLUTION

The turning point is to augment reactive endpoint detection with predictive turn-completion modelling. Silence is a lagging indicator: by the time enough has accumulated to be convincing, the moment has passed. A turn ending is signalled before the sound stops — by falling intonation and syntactic completion. Prediction does not replace detection; it sits on top of it, alongside confidence thresholds, barge-in and repair.

The stronger version goes beyond a better trigger: begin reversible downstream work as turn-end probability rises, then cancel if the person keeps talking. A 2026 endpoint-anticipation system reports forecasting up to 2.56 seconds ahead and cutting average latency by 505 ms, at the cost of a 28.4% rise in speculative computation — which is close to what human listeners appear to do.

FIG. 02 — JUDGEMENTS BY DELAY



REFERENCE FIGURES — THREE SOURCES

Human turn transition, mean	**208 ms**
Preferred robot delay	**~700 ms**
Participants, study 1	**210**
Delays tested	**15, 100–1500 ms**
Typical silence window	**500–800 ms**
TTFT, p50 → p95	**566 → 2246 ms**

Human figure from a ten-language corpus study; robot optimum from a laboratory psychophysical study using video stimuli; pipeline figures practitioner-reported. Curve shapes in Fig. 02 are illustrative of the reported method and optimum, not plotted from the published data.

The trade-off accepted to ship it

A prediction can be wrong, and its failure modes are not symmetric. Predicting too early interrupts the person; too late reintroduces the delay it was meant to remove. Where interruption is the more costly error, the predictor is deliberately biased toward waiting, so the achievable gain is smaller than the theoretical one. What is bought is control rather than elimination: the remaining interval becomes a designed quantity instead of a threshold nobody revisited.

THE OUTCOME — A DIFFERENT TARGET, NOT JUST A FASTER ONE

~700 ms
PREFERRED, ONE STUDY
Point of subjective equality; did not vary significantly with style or visibility

3.4×
SLOWER THAN HUMAN
Ratio of the robot optimum to the 208 ms human turn gap

~630 ms
ILLUSTRATIVE BUDGET
One well-tuned practitioner configuration to first audio — not a platform-independent standard

Sources: Stivers et al., PNAS 106(26):10587–10592, 2009 · International Journal of Social Robotics, 2026 · practitioner latency-budget documentation, 2026. Figure 02 illustrates the psychophysical method and reported optimum; curve shapes are schematic.

Designing for timing rather than speed

Durations are indicative and depend on deployment setting, language coverage and whether the platform exposes stage-level instrumentation.

PHASE 01 · WEEKS 0-4

Instrument the stages separately

Measure endpointing, transcription, first token, first audio and transport as distinct spans, and report the 95th percentile alongside the median. Teams that track only end-to-end average latency cannot tell whether they have a compute problem or a waiting problem — and it is usually the second.

PHASE 02 · MONTHS 1-3

Tune the threshold, and measure what it costs

Shorten the silence window to the smallest value that does not truncate natural pauses, then measure the interruption rate in production rather than estimating it. Interruption rate and latency are the two sides of the same setting, and only one of them usually gets watched.

PHASE 03 · MONTHS 3-6

Add prediction on top of detection, carefully

Introduce turn-end prediction alongside existing endpoint detection, tuned with an explicit cost function that generally penalises premature interruption more heavily, with the balance varying by task and safety context. Validate across the accents, speech rates and hesitation patterns of your actual population — a predictor tuned on fluent, standard-accent speech will systematically interrupt older adults, non-native speakers and people with speech or cognitive impairments, who in a care setting are the people it exists to serve.

PHASE 04 · MONTHS 3-9, THEN CONTINUOUS

Speculate, and design the residual pause

Add speculative execution: begin reversible work as turn-end probability rises and cancel if speech continues. Separately, whatever delay remains will be interpreted, so give it something to carry — a breath, a gaze shift, a postural settle. Consider tracking time to first acknowledgement separately from time to first substantive answer: a robot can signal it is listening long before it has anything to say.

> **Organisational readiness**
>
> Conversational timing falls between three owners: the platform team owns the pipeline, the product team owns the persona, and nobody owns the interval between them. The practical fix is to make turn-transition offset a tracked product metric with a named owner, in the same way error rate or task completion already is — because a quantity that appears on no dashboard will be optimised by no one.

What goes wrong, and what it costs

Ranked by how often it is missed rather than by severity.

RISK 01

Optimising the wrong term

Recognition and synthesis are the most legible components and among the smallest contributors. Endpointing and first-token latency dominate, and neither is fixed by procuring a faster speech engine.

RISK 02

Median-only reporting

Average latency flatters a system by smoothing away exactly the slow turns users remember. Report tail percentiles alongside the median, together with interruption rate, premature and missed endpoint rates, overlap rate and recovery latency.

RISK 03

Borrowed human targets

The 200 ms figure is real and well sourced, and it describes humans talking to humans. Experimental evidence puts the robot optimum around 700 ms, so inheriting the human number sets an unreachable target and the wrong one.

RISK 04

Interruption treated as free

Shortening the silence window trades latency for cutting people off. In care and companion settings interrupting is the more damaging failure, and it rarely appears in the same dashboard as latency.

RISK 05

Predictors tuned on fluent speech

Turn-end prediction learned from typical speakers will misfire on slower, disfluent or accented speech — systematically disadvantaging older adults, people with speech or cognitive impairments and non-native speakers, who are frequently the intended population.

RISK 06

Memory as an unbudgeted cost

Conversational history is what makes a companion feel like one and what stretches the latency tail. Without selective retrieval, the system gets slower precisely as the relationship deepens.

RISK 07

Silence left undesigned

The residual pause will be interpreted whether or not anyone designed it. Dead air reads as absence; a breath or gaze shift of identical duration reads as thought.

RISK 08

Evidence read past its scope

The optimum comes from a laboratory study using video stimuli, and the pipeline figures are practitioner-reported. Both are useful design references; neither is a measurement of your deployment.

08 / WHERE TO START

Three questions worth asking this quarter

01 Do we know how much of our latency is waiting rather than computing?

02 Are we measuring interruption rate alongside response time?

03 Which of our targets did we inherit rather than test?

The third question generalises well beyond conversation. Most systems carry at least one number that was correct about something adjacent.

CLOSING THE ROBOTICS TRACK

Five editions, one argument

R01 learned the actuator it could not model. R02 measured the compressibility it could not see. R03 declared the solver parameters it could not standardise. R04 assembled parts finer than the machine could aim. R05 found the latency target itself was borrowed from a different question. The recurring finding is that the binding constraint is usually a quantity nobody was measuring.

SOURCES & METHOD

Primary: Stivers, Enfield, Brown, Englert, Hayashi, Heinemann, Hoymann, Rossano, de Ruiter, Yoon and Levinson, “Universals and cultural variation in turn-taking in conversation,” PNAS 106(26):10587-10592, 2009 · “Talking to Robots: Does a Robot’s Personality Change How we Feel About its Response Delay?” International Journal of Social Robotics, 2026.

Supporting: “Endpoint Anticipation for Low-Latency Spoken Dialogue,” arXiv:2606.13450 · Scientific Reports 15, 2025, on gaze adaptation during delayed robot movement · multi-party open-ended conversation with a social robot, arXiv:2503.15496 · CHI 2026 conversation-analysis study of agent latency in VR.

Method note: the 208 ms figure is a mean turn-transition offset between human speakers, not a measure of how quickly people compose answers. The ~700 ms figure is a preferred delay estimated in one laboratory task using video stimuli and yes/no questions; the same study found no effect of delay on measured social dimensions, so the reading that this interval signals deliberation is this paper’s hypothesis rather than the study’s finding. The gaze study used 4 and 10 second movement delays and does not validate a sub-second speech target. Pipeline budgets are illustrative practitioner traces, not cross-platform benchmarks. Figure 02 curve shapes are schematic rather than plotted from published data.

Cross-references Edition R01 on measured versus assumed behaviour.

Lise DELOBEL

GenAI project manager & consultant — Physical AI, digital twins, and applied ML for industrial and extreme-environment operations. MIT IDSS Data Science & Machine Learning certificate.

LISEDELOBEL.FR · LISE.DELOBEL@ME.COM

PHYSICAL AI FRONTIERS · ROBOTICS TRACK · EDITION R05 · JULY 2026