Universal Library of Innovative Research and Studies Research Article

ISSN: 3065-0003 | Volume 3, Issue 4

Open Access | PP: 27-33

DOI: https://doi.org/10.70315/uloap.ulirs.2026.0304005

Modular Prompt Architecture in a Production Outbound Voice Agent and Its Effects on Conversation Length, Scenario Progression and Cost

Nadezhda Shiroglazova

MS in Data Science, Northwestern University, Chicago, USA.
ORCID: 0009-0006-9011-212X

Abstract

Many outbound voice agents built on large language models run a sales scenario from a single long prompt, and the effect of splitting that prompt into a graph of dialogue blocks has received little measurement in production traffic. The study compares two architectures for an outbound sales scenario, deployed on the same telephony platform, targeting a cold base of former customers in a European country. The corpus covers 14,155 dials over twenty-seven dialing days, with 5,892 answered calls and 5,119 dialogues in the main sample. Scenario depth was measured using a language-model labeler, validated against block telemetry, across 4,572 dialogues. The multi-block agent kept answered calls longer, with a median of 44.8 s against 31.5 s, and the gap of 13.3 s stayed between 12.4 and 14.0 s without transferred calls, on a shared base model and on days with matched answer rates. Survival analysis identified a divergence between the 10th and the 30th second of the call, with the single-prompt agent ending 17.1% of answered calls within the first 30 s, compared with 0.6% for the modular agent. Dialogues reaching need discovery rose from 14.7% to 53.1% after correcting labeler errors against telemetry, and transfers to a human operator rose from 5.1% to 10.7% of dialogues. Modularity carried a measurable price: the platform charge per answered call doubled, tokens per turn grew by 60%, and the 90th percentile of response latency grew by 27%, while the cost per transfer stayed level at $2.41, unchanged from $2.44. The findings frame modular prompt design as a trade-off between conversational robustness and per-call expense.

Keywords: *Conversational AI, Voice Agents, Modular Prompting, Prompt Architecture, Outbound Sales, LLM-As-A-Judge, Dialogue Telemetry, Human Handoff.*

INTRODUCTION

This study is based on firsthand engineering of an outbound voice agent for a telephone sales company that operates distributed contact centers in Europe and Latin America. The agent calls former customers with no current intent to buy, qualifies their interest and passes warm customers to human operators. The scenario underwent a direct redesign from a single continuous prompt to a graph of dialogue blocks, and the comparison reported here uses the platform records of this agent before and after the redesign. The design logic behind every block boundary is therefore known from the design work itself, and each measured transition corresponds to a block with a known instruction.

Voice-based conversational agents have moved from scripted menus into open dialogue, and a natural field experiment in a call center has shown that a voice agent changes customer behavior and service outcomes in measurable ways (Wang et al., 2023). Generative models also alter the work of human support staff, with productivity gains concentrated among less experienced workers (Brynjolfsson et al., 2025). An outbound sales agent faces a harder task, since they have to keep a stranger on the line, obtain consent, identify a need, and hand the customer to an operator within a few minutes of a call the customer did not ask for.

The way the scenario is presented to the model is an engineering choice with consequences for each step. Chaining several prompts, each responsible for a narrow subtask, makes a language-model system more transparent and easier to control than one monolithic prompt, because each step can be inspected and revised in isolation (Wu et al., 2022). A

Citation: Nadezhda Shiroglazova, "Modular Prompt Architecture in a Production Outbound Voice Agent and Its Effects on Conversation Length, Scenario Progression and Cost", Universal Library of Innovative Research and Studies, 2026; 3(4): 27-33. DOI: https://doi.org/10.70315/uloap.ulirs.2026.0304005.

single prompt also carries a cost that grows with its length. Language models use information placed in the middle of a long context worse than information placed at its edges (N. F. Liu et al., 2024). Their reasoning accuracy declines as the input grows, long before the context window fills (Levy et al., 2024). A sales script with seven stages and transfer rules occupies thousands of characters, and the instruction that matters at a given turn sits somewhere inside that mass.

Research on task-oriented dialogue shows that general-purpose models produce fluent replies yet lose track of the dialogue state across turns when no explicit structure holds it (Hudeček & Dušek, 2023). Agent architectures built on language models address the same difficulties in planning, memory, and tool modules, which split a task into parts the model can handle one at a time (Wang et al., 2024). Spoken interaction imposes timing constraints, since delays at turn transitions shape how users perceive a spoken system (Skantze, 2021), and every additional model decision at a block boundary is introduced as audible delay.

The opening seconds of an outbound call also decide whether persuasion can begin. In automated customer conversations, small initial commitments increase compliance with later requests (Adam et al., 2021), turning the consent stage of a sales call into the point at which the rest of the scenario is won or lost. Measuring how far each call progresses through such a scenario requires annotation at scale. Language models have matched crowd workers on text annotation tasks (Gilardi et al., 2023), and evaluation schemes that provide models with explicit criteria reach close agreement with human judgment (Y. Liu et al., 2023). A model labeler still needs a reference to be trusted, and the transition logs of a block-based agent can serve as one.

In the reviewed literature, no study was found that measures outbound production traffic in which the conversation of a voice agent breaks down stage by stage, or that examines how a block-graph architecture changes that pattern, along with its latency and monetary cost. The study therefore compares a single-prompt and a multi-block version of one outbound sales scenario on one platform with respect to conversation length, progress through the seven scenario stages, transfers to human operators, and the price of modularity in tokens, latency, and platform charges. A previous study of the same production system addressed the pacing of call launches against operator availability (Shiroglazova, 2026), and the present work leaves that control problem aside.

MATERIALS AND METHODS

The Deployment and the Two Architectures

The agent was built for repeat sales to a cooled base of former customers, a low-priority segment for the business. Both architectures called this base, so the absolute level of every outcome reflects a difficult segment, and the quantity of interest is the difference between the arms. The single-prompt arm ran the entire scenario from a single continuous instruction text, which grew from 7,840 to 10,875 characters across its revisions. The multi-block arm decomposed the same scenario into a graph of conversational blocks, increasing the number from 9 to 13, with instructions totaling 10,896 to 23,107 characters. Each block carried the instruction for one part of the conversation and belonged to one of the seven scenario stages, and the platform moved control along defined transitions with compact state variables.

Both arms followed the same script with the same seven stages, the same agent role and the same two tools, one to end the call and one to transfer the customer to a human operator. The stages run from greeting and identity confirmation through consent to talk, need discovery without and then with the product, a summary and a short presentation of the offer and transfer. At the switch, the redesign kept the stages, the role, and the tools of the script, and changed how the prompt was assembled and delivered to the model.

Data and Anonymization

The records come from a direct export of the call log maintained by the telephony platform for the agents described above. The corpus contains 14,155 dials made over twenty-seven dialing days, which span the period from twenty-one days before to ninety-two days after the architecture switch, with one row per dial. Days carry the codes D01-D27. Each row holds the arm, the base model tier, the end reason, duration, turn and character counts, the scenario stage from two sources, response latency, tokens per turn, tool calls, and the platform charge in US dollars. Phone numbers and the names of the company, platform, country, product, and base models are absent from the export, and agent and call identifiers are salted irreversible hashes. Each arm ran as several agent instances on the platform, two in the single-prompt arm and nine in the multi-block arm. The base models appear as model A and model B, and the single-prompt arm used model A on every call. The multi-block arm used model A and model B in different periods and, on days D21, D23, and D24, both models in parallel.

Definitions and Sample Construction

A call counts as answered when a person picks up and the duration exceeds zero. A dialogue is an answered call in which the customer spoke at least one utterance. A transfer is a call that ends with a completed handoff to a human operator; an announced intent to transfer does not qualify. The analysis uses dialogues as the primary denominator for transfers and reports answered calls and dials alongside them. Daily answer rates on campaign days ranged from 97.0% to 15.5% as telephony routes and list quality changed, and all comparisons of conversational behavior therefore run within answered calls.

Days with fewer than 50 dials were manual test sessions, and the main comparison excludes them. These days held 22 dials in the single-prompt arm and 87 in the multi-block arm. With the tests included, the single-prompt transfer rate per dialogue rises from 5.1% to 6.9%, and its platform cost per transfer falls from $2.44 to $1.87, because one test session produced 10 transfers on 21 dials. The main sample holds 1,203 dials from the two single-prompt campaign days and 12,843 dials from the multi-block campaign days.

The multi-block period overlaps with the launch of predictive dialing and with 70 unmarked revisions of the graph, so the differences reflect the transition to the modular architecture and its subsequent development. The exported records contain neither sales outcomes nor utterance text. The correction of the labeler error described below assumes the same error pattern across transcripts from both arms.

Measuring Scenario Progress

The platform records the scenario stage for every multi-block call because it logs each transition between blocks, including the block name and a timestamp. The single-prompt arm has no physical stages, and its progress can be reconstructed from the transcript alone. A language-model labeler read each transcript on the platform side before anonymization and assigned the furthest stage reached on the seven-stage scale; the export retains this label without the text. The same labeler ran on both arms, so the comparison uses one scale, and stage shares use the dialogues with a labeler stage: 532 of 547 in the single-prompt arm and all 4,572 in the multi-block arm. The accuracy of the labeler was assessed against block telemetry for all 4,572 multi-block dialogues in the main sample. Exact agreement on the seven-stage scale reached 56.8%, agreement within one stage reached 92.7%, the quadratically weighted kappa was 0.85, and agreement on the binary question of reaching the presentation stage was 95.6% with a kappa of 0.84. The labeler is unreliable on large steps in the funnel and confuses adjacent stages 2 and 3, assigning stage 3 to 725 of the 1,655 dialogues that telemetry placed in stage 2.

To remove this bias from the single-prompt funnel, the confusion matrix estimated on multi-block dialogues served as a model of misclassification, and an expectation-maximization unfolding recovered the distribution of true stages from the observed single-prompt labels. Applied to the multi-block labels, the procedure reproduced the telemetry funnel to within 0.2 percentage points at every stage. A joint bootstrap of the confusion matrix and the single-prompt labels, with 1,000 replicates, yielded the interval for the corrected estimate.

Statistical Analysis

Duration comparisons rely on medians, the Mann-Whitney test, the rank-biserial effect size r, and bootstrap intervals for the difference in medians with 2,000 resamples. Proportions were compared using Fisher's exact test, and odds ratios were calculated. Survival was described through the share of answered calls active at each second and through the interval hazard, the share of calls ending within an interval among those active at its start.

Adjusted estimates were obtained from median and logistic regression, with the daily answer rate, the logarithm of daily volume, and the model tier as covariates, and standard errors clustered by day. Robustness checks repeated the main comparison without transferred calls, on multi-block traffic served by model A, on multi-block days whose answer rate fell within a band of 44% to 56% around the single-prompt rates of 50.2% and 53.7%, and without day D16, a low-volume day with 69 dials above the test threshold. Two placebo comparisons set model A against model B on the days when both served the multi-block agent and the first single-prompt campaign day against the second. The platform charge per answered call was decomposed on a logarithmic scale into components attributable to conversation length and to the price per minute of airtime.

RESULTS

Sample

The main sample contains 620 answered calls and 547 dialogues in the single-prompt arm, compared with 5,272 answered calls and 4,572 dialogues in the multi-block arm, with 28 and 491 transfers, respectively. Answered calls made up 51.5% and 41.0% of dials, a gap that follows the drift of telephony routes and list quality.

Conversation Length and Shape

The multi-block agent held answered calls longer. The median duration was 44.8 s versus 31.5 s, a difference of 13.3 s, with a 95% bootstrap interval of 10.2 to 17.6 s, $r = 0.22$, and $p < 0.001$. A multi-block dialogue carried 18.1 turns against 12.5 in the single-prompt arm, and the agent spoke 640 characters per dialogue against 374. Customers spoke more as well, with 286 characters per dialogue against 143 and 9.9 turns against 6.3, while the median customer share of all characters stayed at 23.7% and 25.2%.

Fig. 1 traces the moment when the two arms part. The survival curves of answered calls in panel a remain within 3 percentage points of each other during the first 10 s, diverge between the 10th and the 30th second, and then descend in parallel up to the second minute. At 30 s, 71.1% of multi-block calls remained active, compared with 50.8% of single-prompt calls. Among calls active at 10 s, 45.3% of single-prompt calls ended before 30 s, compared with 25.4% of multi-block calls, yielding an odds ratio of 0.41 ($p < 0.001$). Between 30 and 120 s, the hazards match, for example, at 29.2% and 30.0% between 30 and 45 s. Beyond 120 s, fewer than 100 single-prompt calls remain active, and their faster ending widens the gap in panel a.

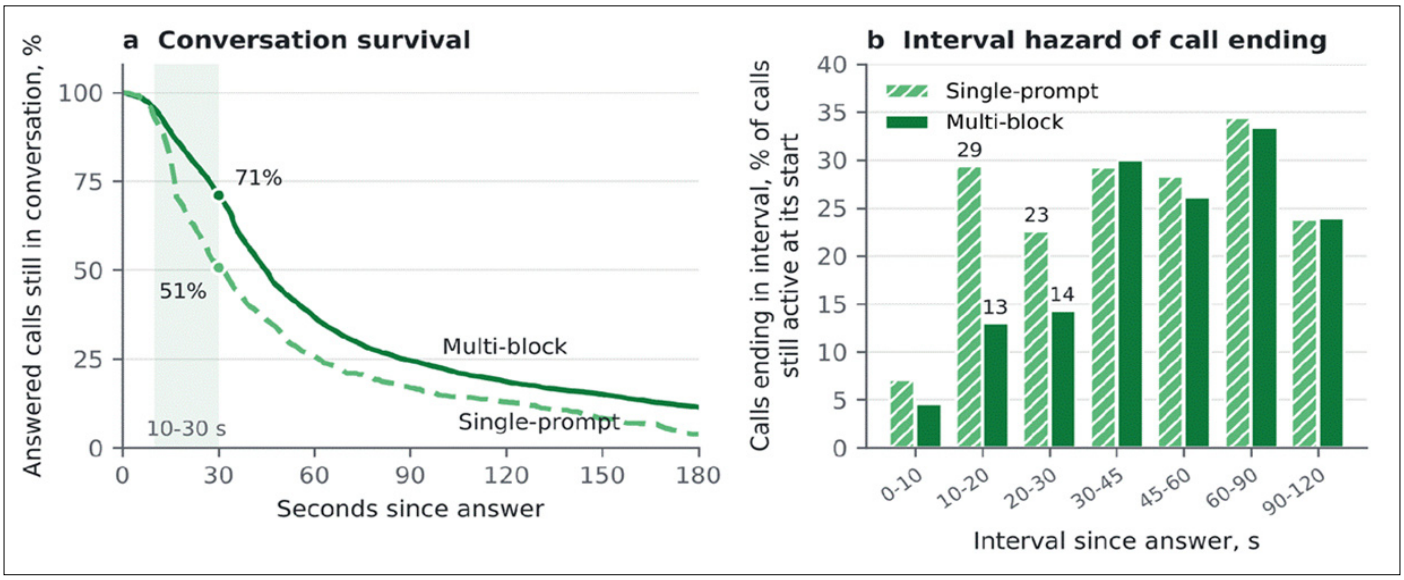


Fig. 1. Survival of answered calls and interval hazard of call ending in the single-prompt and multi-block arms

End reasons point to the agent as the party that closes many of these calls. The single-prompt agent ended 26.1% of answered calls on its own initiative, compared with 8.6% for the multi-block agent (odds ratio 3.76, $p < 0.001$). Additionally, 17.1% of its answered calls were agent hang-ups within the first 30 s, compared with 0.6% for the multi-block agent. Single-prompt hang-ups had a median duration of 20.1 s, and the labeler placed 58% of them at the greeting stage, which ties the window between 10 and 30 s to the opening stages of the scenario. Multi-block hang-ups came later, at a median of 71.8 s. The adjacent campaign days around the switch, D03 under the single prompt and D05 under the first multi-block version, differed in median duration by 5.1 s, with an interval from 0.4 to 9.3 s and $p = 0.016$, so part of the length effect appeared on the first day of the new architecture.

Scenario Progression and Block Telemetry

The multi-block agent moved further through the scenario. On the labeler scale, which carries the same error pattern in both arms, the mean stage reached was 3.15 versus 2.36, and 63.3% of multi-block dialogues reached need discovery at stage 3 or later, compared with 24.2% of single-prompt dialogues. The gap opened in the middle of the funnel and narrowed to 12.7% against 10.0% at the final stage.

Fig. 2 sets the labeler funnels beside the block telemetry and the corrected single-prompt estimate. Telemetry places 53.1% of multi-block dialogues at stage 3 or later, 10.2 percentage points below the labeler, which confirms the upward bias at the boundary between stages 2 and 3. After correction with the same error model, the single-prompt share at stage 3 or later falls to 14.7% with a bootstrap interval from 11.5% to 18.4%. The difference between arms survives the correction at 38.4 percentage points, and the ratio between the arms grows from 2.6 to 3.6.

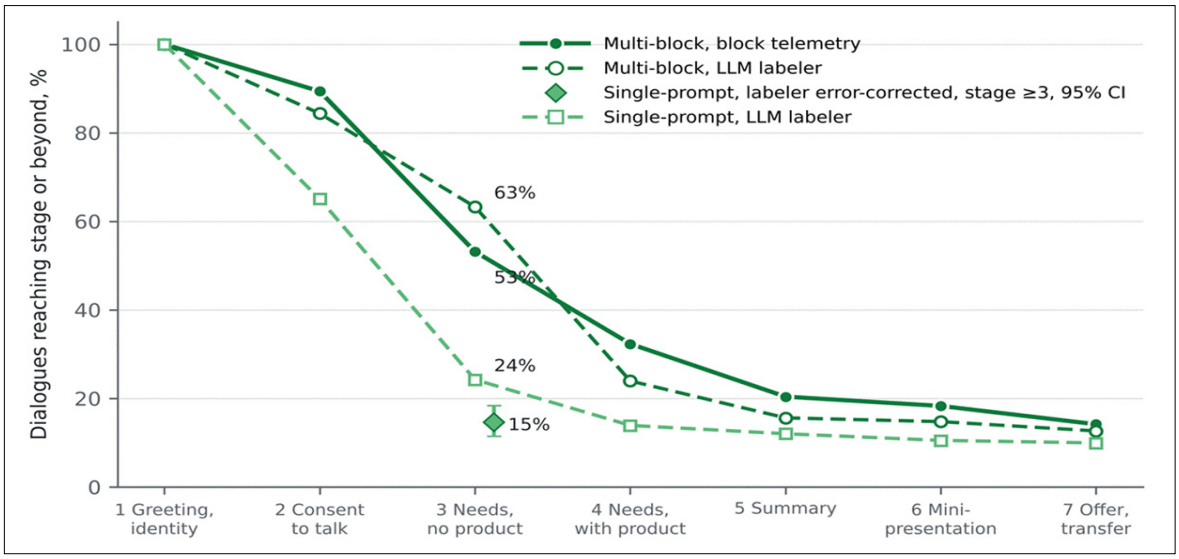


Fig. 2. Share of dialogues reaching each scenario stage by arm and measurement source

Block telemetry shows where multi-block conversations dissolve. Of the dialogues that entered a block, 89.3% progressed from greeting to consent, 59.5% progressed from consent to need discovery without the product, 60.8% progressed to need discovery with the product, and 63.1% reached the summary. The later transitions kept 89.7% and 77.4%. Attrition therefore concentrates in the three transitions that cross need discovery, where each step keeps three dialogues in five. Every multi-block transfer originated in the offer block and ended in a transfer in 75.8% of the dialogues that reached it.

Depth grew as the graph developed. On day D05, the first multi-block campaign day, 27.2% of dialogues reached stage 3 or later by labeler, within 3 percentage points of the single-prompt baseline of 24.2%. From day D19 onward, every multi-block day reached at least 60%. The daily share correlated with the number of days since the switch (Spearman $\rho = 0.74$, $p < 0.001$).

Transfers to Human Operators

Transfers to human operators doubled. Per the dialogue, the rate rose from 5.1% to 10.7%, with an odds ratio of 2.23 (1.51-3.30) and $p < 0.001$. Per answered call, the rates were 4.5% and 9.3%. On model A traffic, the multi-block agent transferred 10.6% of 1,750 dialogues. The daily transfer rate showed no monotonic trend ($\rho = 0.06$) after a slow start of 3.9% on day D05. On the adjacent days D03 and D05 the transfer rates were 5.2% and 3.9% with $p = 0.79$.

The gain has two sources. By labeler, 10.0% of single-prompt dialogues and 12.7% of multi-block dialogues reached the offer stage. Among those who reached it, the transfer was completed in 52.8% of single-prompt cases (28 of 53) and in 75.5% of multi-block cases (437 of 579), yielding an odds ratio of 2.75 with $p < 0.001$. On the logarithmic scale, deeper progression accounts for 40% of the increase in transfers and completion at the offer stage accounts for 60%. A further 54 multi-block transfers carried a labeler stage below the offer stage and stay outside this decomposition.

Robustness

Table 1 collects the robustness checks. Across all variants of the main comparison, the difference in median duration stayed between 10.9 and 14.0 s, the odds ratio for transfers between 2.16 and 2.23, and the share of multi-block dialogues reaching stage 3 between 63.3% and 64.4%. The adjusted models retained all three effects. Neither placebo comparison showed statistically significant differences across the evaluated outcomes ($p \geq 0.15$ throughout).

Table 1. Main effects and robustness checks

Comparison	Answered calls, S/M	Δ median duration, s, 95% CI	Transfers, OR, 95% CI	Stage 3 or later, %
1. Main comparison	620 / 5,272	+13.3, 10.2 to 17.6	2.23, 1.51 to 3.30	24.2 → 63.3
2. Transferred calls excluded	592 / 4,781	+12.4, 8.2 to 15.2	n/a	n/a
3. Model A traffic only	620 / 1,964	+14.0, 10.5 to 18.5	2.19, 1.45 to 3.30	24.2 → 64.4
4. Answer rate of 44% to 56%	620 / 2,262	+13.7, 10.2 to 18.2	2.16, 1.44 to 3.24	24.2 → 64.0
5. Day D16 excluded	620 / 5,225	+13.4, 10.4 to 17.9	2.17, 1.47 to 3.22	24.2 → 63.8
6. Covariate-adjusted models	620 / 5,272	+10.9, 7.1 to 14.7	2.18, 1.52 to 3.13	OR 4.18
7. Placebo, model A → B	1,063 / 870	−2.3, −6.7 to 2.9	1.03, 0.77 to 1.39	68.8 → 65.5
8. Placebo, D01 → D03	248 / 372	−3.0, −8.6 to 4.5	1.03, 0.47 to 2.25	27.4 → 22.0

Note: S and M denote the single-prompt and multi-block arms, and in rows 7 and 8 the first and the second compared group. OR is the odds ratio for transfers per dialogue, and row 6 also gives the odds ratio for reaching stage 3. Rows 1 to 6 reach $p < 0.001$ on every outcome, and rows 7 and 8 stay at $p \geq 0.15$.

All seven multi-block agents with at least 80 answered calls exceeded the single-prompt baseline, with medians of 35.2 to 52.2 s, stage-3 shares of 30.0% to 71.0% and transfer rates of 5.7% to 13.4%.

The Price of Modularity

Modularity raised the expense of each conversation. The platform charge per answered call rose from $0.109 to $0.223, a factor of 2.04, and by a factor of 1.78 on model A traffic. Tokens per turn grew from 3,588 to 5,736, an increase of 60%, and tool calls per dialogue grew from 0.32 to 1.38. Median response latency changed by 3%, from 1,654 to 1,710 ms, while the mean 90th-percentile latency grew from 2,437 to 3,100 ms, an increase of 27%. Inactivity timeouts ended 3.4% of answered multi-block calls, compared with 1.0% of single-prompt calls.

A logarithmic decomposition attributes 61% of the cost increase per answered call to longer conversations, with mean duration up by a factor of 1.55, and 39% to a higher price per minute of airtime, which rose from $0.127 to $0.168, an increase of 32%. The cost per transfer stood at $2.44 in the single-prompt arm and $2.41 in the multi-block arm; airtime per transfer fell from 19.1 to 14.3 min, and the platform charge for dialogues per dialogue that reached need discovery fell from $0.505 to $0.390. Unanswered dials accounted for 0.5% of the multi-block bill.

DISCUSSION

The results describe a change in the form of the conversation. A single prompt asks the model to locate the relevant

instruction inside a long text at every turn, and the evidence on long-context use predicts that instructions placed in the middle of such a text lose influence over the output (N. F. Liu et al., 2024). Tokens per turn grew by 60% in the multi-block arm, so a shorter input cannot explain the gain. The block graph changes the structure of the input, because at every turn the platform designates one active block whose instruction governs the reply, while the single prompt keeps the instructions of all seven stages in the input at every turn. The survival analysis places the gain in the opening stages, where the single prompt must shift the agent's goal while the customer is still deciding whether to listen, and text irrelevant to the current step degrades this kind of reasoning (Levy et al., 2024).

A single-prompt agent that ends 17.1% of its answered calls within 30 s of pickup is consistent with a misreading of the conversational state at the start of the call, the kind of state-tracking failure that research on task-oriented dialogue associates with unstructured prompting (Hudeček & Dušek, 2023). Block transitions give the state an explicit carrier, and the rate of early agent hang-ups fell to 0.6%. Initial commitments shape later compliance in automated customer conversations (Adam et al., 2021), and an agent that secures consent before moving on builds the commitment that need discovery depends on. Between the 30th and the 120th seconds, the interval hazards of both arms coincide, suggesting that the architecture determines who stays in the conversation.

Among the transfers that the labeler places at the offer stage, deeper progression accounts for 40% of the gain, and a higher completion rate at that stage accounts for 60%. Completion depends on the dedicated offer block with its own instructions and a holding node, and on a free operator at the moment of handoff. Prompt chaining allows one step of a pipeline to be refined without disturbing the others (Wu et al., 2022), and the growth of stage depth over the first weeks of operation reflects that use.

A block transition added a model decision, and the tail of response latency grew by 27% while the median remained unchanged. Research on turn-taking links pauses at turn transitions to the perceived quality of a spoken agent (Skantze, 2021), and the rise in inactivity timeouts from 1.0% to 3.4% aligns with a reading in which some customers encountered a long pause at a block boundary and fell silent. The monetary cost per call doubled, yet 61% of that increase covered the cost of longer conversations, and the cost per transfer stayed level. For an operation that values warm customers delivered to operators, the modular design yields twice the transfers per dialogue at the same price per transfer. In production, the operating team reported sales growth after the transition, and multi-block assembly has since become the only design in use.

A language-model labeler can achieve high agreement on coarse categories (Gilardi et al., 2023), and a criteria-based evaluation with a strong model can track human judgment in text generation tasks (Y. Liu et al., 2023). The present labeler met that standard on large-funnel steps, with a weighted kappa of 0.85, yet its systematic confusion between stages 2 and 3 inflates the multi-block share at stage 3 by 10.2 percentage points. Production telemetry made it possible to estimate the bias and correct the single-prompt funnel accordingly, which keeps the comparison between arms on a common footing. Field studies of voice agents have relied on outcome data (Wang et al., 2023), and block telemetry extends such evaluation to the internal course of each call.

LIMITATIONS

The unequal observation periods and limited number of baseline transfers constrain the precision of comparisons, particularly at the later scenario stages. The multi-block period overlaps with the launch of predictive dialing and with 70 unmarked revisions of the block graph, so the measured effects belong to the transition to modular assembly and its later development. Sales outcomes are absent from the records, and the transfer is the last step that the data can observe. The export holds no utterance text, which leaves the content of early agent hang-ups beyond direct inspection. The correction of the single-prompt funnel assumes that the labeler errs on single-prompt transcripts in the same pattern as on multi-block ones.

CONCLUSION

Splitting the prompt of a production outbound voice agent into a graph of dialogue blocks changed the shape of its conversations. Answered calls lasted 13.3 s longer at the median, customers spoke twice as many characters, and the divergence was concentrated between the 10th and the 30th second. The single-prompt agent ended 17.1% of answered calls within the first 30 s, compared with 0.6% for the modular agent. Dialogues reaching need discovery rose from 24.2% to 63.3% by labeler and stood at 53.1% against 14.7% after correction against block telemetry. Transfers to human operators doubled from 5.1% to 10.7% of dialogues, driven by deeper progression and a higher completion rate at the dedicated offer block. The modular design doubled the cost per answered call and lengthened the latency tail, while the cost per transfer held at $2.41, unchanged from $2.44. Block telemetry located attrition at the transitions through need discovery and supplied a reference for correcting a language-model labeler. For engineering teams that run outbound voice agents, the results support modular prompt assembly wherever completed transfers drive campaign value, since the doubled per-call expense leaves the cost per transfer unchanged, and they argue for logging block transitions from the first day of deployment.

REFERENCES

1. Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets, 31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7
2. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044
3. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120
4. Hudeček, V., & Dušek, O. (2023). Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 216–228). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.sigdial-1.21
5. Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15339–15353). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.818
6. Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638
7. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511–2522). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.153
8. Shiroglazova, N. (2026). Dynamic operator allocation for conversational AI voice-calling systems. *International Journal of Advanced Artificial Intelligence Research, 3*(7), 69–81. https://doi.org/10.55640/ijaair-v03i07-07
9. Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language, 67*, Article 101178. https://doi.org/10.1016/j.csl.2020.101178
10. Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2023). Voice-based AI in call center customer service: A natural field experiment. *Production and Operations Management, 32*(4), 1002–1018. https://doi.org/10.1111/poms.13953
11. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1
12. Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Article 385). Association for Computing Machinery. https://doi.org/10.1145/3491102.3517582

Copyright: © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.