

Split-Second Decisions: Handing a Voice Agent's Small Decisions to a Decision Model

Madhu

Kansas City · Independent

ABSTRACT

A voice agent makes several small, closed decisions in every turn of a call: whether the caller has finished, whether a sound is an interruption, whether to hand over to a person, whether an action needs confirmation, whether a drafted sentence is safe to speak. We handed these to a small decision model that returns a typed answer with a probability, and measured accuracy on labelled cases against a time allowance. On the first run, with plain instructions and no tuning, about 39 decisions in 40 were right, a typical decision took under a quarter of the time allowed, and two kinds of decision had no errors. Checking a drafted sentence against a policy was the weakest, at about nine in ten. End to end, the agent still answered in nearly three times the target for natural conversation, and the decisions account for only about a tenth of that wait. This brief gives the results in relative terms with their limits. The figures are in the full results paper, shared on request. The system is proprietary: nothing here says which models were used or how the decisions are worded.

1 The question

On the phone, the time to reply is a conversational pause. A large language model can make every decision a call needs, but slowly, in prose, and without saying how sure it is. The question was whether the small decisions can be separated from the talking and answered by something small, fast and checkable.

2 How it was judged

For each kind of decision there is a set of labelled situations from a phone-support setting. The score is the mean accuracy across the kinds. Each decision has a time allowance for the typical case and a larger one for the slowest cases, and the cost of a thousand decisions is reported. Cases were written by one model and labelled blind by another, and kept only where the two agreed. End-to-end timing was taken separately, on a scripted call with a synthesised caller.

3 What was found

Table 1: Results in relative terms, first run.

Measure	Result
Decisions right, mean across kinds	about 39 in 40
Kinds of decision with no errors	two: routing, and confirm-before-acting
Weakest kind: checking a drafted sentence against a policy	about nine in ten
Time per decision, typical case	under a quarter of the allowance
Time per decision, slowest cases	about an eighth of the allowance
Share of the caller's wait taken by all decisions in a turn	about a tenth
Caller's wait, end to end, against the target	nearly three times

The reflex-like decisions are close to solved at this difficulty. The decision that requires holding several rules against one sentence is not, and in a live test call it was also over-cautious. The end-to-end wait is dominated by the reply's first word, then by turning speech into text, then by deciding that the caller has finished.

4 Limits

The labelled cases are machine-written, machine-labelled, clean text, and on the easy side; no person has checked them. The accuracy is from one run in one setting, before any tuning. The end-to-end timing is from a scripted call on a busy machine and is pessimistic, but it is not a measurement of real callers. No result has been reproduced by a third party.

5 The figures

The accuracy for every kind of decision, the timing of every stage and the cost are in the full results paper. It is shared on request to the author.