

Twelve Questions, Then a Twin: Fidelity of a Person's Twin Built from a Short Voice Interview

Madhu

Kansas City · Independent

ABSTRACT

We report results for Think Like Me, a system that turns a twelve-question voice interview into a twin: a system that answers new questions as the interviewed person would, and speaks in that person's voice. The twin is scored on 42 questions per person that the interview does not touch: 18 ratings, 12 forced choices and 12 open questions. On twelve simulated people it reaches a fidelity of 77.5 on a scale of 0 to 100, which is 86.4 percent of how consistently those people agree with themselves when asked twice. The largest published study of this kind reports 85 percent for two-hour interviews of 1,052 real people [1]; the two figures are not like for like, because the people here are simulated and are more consistent than humans. In a spoken rehearsal the twin begins its reply a median of 2.1 seconds after the caller stops. No real person has been scored, and the likeness of the voice has not been measured. The system is proprietary: this paper reports what it measures and not how it is built.

1 Introduction

A transcript records what a person said. A twin has to answer what the person was never asked. The difference matters for any use in which a model stands in for one individual, and it can only be measured on questions that the model could not have rehearsed.

Park et al. [1] showed that a model conditioned on a two-hour interview reproduces a person's survey answers at 85 percent of that person's own two-week consistency. Two hours is a long time to ask of anyone. This work asks what a twin can do after twelve questions, and reports the result with the conditions it was measured under.

This is a results paper. It states what was measured, on whom, how the measure is defined, and where the result is weak. It does not describe the interview or the design of the twin.

2 Setting

2.1 What goes in and what comes out

The input is one short voice interview of twelve questions. The output is a twin that can be spoken to in real time and that answers in a voice taken from the person's own interview answers.

2.2 The people

The benchmark has twelve people who vary in age, culture, class and personality. They are simulated: each is a written character played by a language model. The twin never sees the written character. It hears only what the person said in the interview.

2.3 The test questions

Each person answers 42 fixed questions that are held out from the interview: 18 statements rated on a scale of 1 to 5, 12 situations with a forced choice, and 12 open situations answered in the person's own words. Each person answers every question twice, so that their agreement with themselves is known. A run in which an interview question overlaps a test question is void.

3 Measure

Fidelity is the mean of three agreements, multiplied by 100. For ratings, agreement on an item is $1 - |\text{twin} - \text{person}| / 4$. For forced choices, agreement is accuracy. For open answers, a separate judge that does not know which answer is whose rates from 0 to 10 whether the twin's answer takes the same stance, decision and priorities as the person's; wording is not scored.

Normalized fidelity divides each part by the people's own agreement between their first and second answers, following [1]. It reads as the share of a person's own consistency that the twin reaches.

4 Results

4.1 Fidelity

Table 1: Twelve simulated people, 42 held-out questions each. The ceiling is the people's agreement with their own second answer.

Part	Questions	Twin	Ceiling
Ratings, 1 to 5	18	0.829	0.99
Forced choices	12	0.868	0.99
Open answers	12	0.627	0.72
Fidelity	42	77.5	

Part	Questions	Twin	Ceiling
Normalized fidelity		86.4	

The twin is within one step of the person on 89 percent of ratings. Across the twelve people, fidelity ranges from 71.7 to 81.3.

4.2 By kind of question

Agreement is highest on moral lines (0.89) and on how the person decides (0.83), and lowest on life story and personal stances (0.67). The open questions that are missed most often concern habits and priorities in situations the interview did not reach.

4.3 Systematic errors

On ratings, about half of the error is a one-step miss toward the middle of the scale. On statements worded in the negative, the twin tends to rate the underlying trait and not the statement, and so answers at the wrong end of the scale.

4.4 Speed and cost

In a real-time rehearsal over the full audio path, with a synthetic caller and on a single laptop, the median delay from the end of the caller's speech to the start of the twin's reply is 2.1 seconds. On the written test the twin's median time to answer a question is 1,065 ms. One full scoring pass over the twelve people costs \$0.87.

4.5 Noise and the record

The same configuration scored twice gave 77.7 and 77.5. Differences of under about one point are therefore not resolved by this benchmark. The record holds 28 measured experiments: 3 changes were kept, 17 were discarded, 7 did not complete and 1 was void.

5 Comparison

The reference result is 85 percent normalized, for two-hour interviews of 1,052 real people [1]. The normalized 86.4 reported here is from a twelve-question interview, but it is not a like-for-like comparison: the people here are simulated, and they agree with themselves 0.99 of the time on ratings and choices where real participants in a public dataset [2] agree 0.82 to 0.85. Absolute numbers on simulated people are optimistic.

6 Limitations

The people are simulated. No real person has been scored. A second test set has been built from 30 real participants and 40 questions of a public dataset [2]; on it, answering every question with the neutral option scores 63.4 and the participants' own consistency is 83.6, and no twin has been scored on it. The system has not been used by a real person on a real microphone. The likeness of the twin's voice to the person's has not been measured. The benchmark is small: twelve people, and a resolution of about one point.

7 Availability

The questions of the interview and the design of the twin are not in this paper. The system is proprietary and is available on request to the author.

References

1. Park, O'Brien, Cai, Morris, Liang, Bernstein. Generative Agent Simulations of 1,000 People. arXiv:2411.10109, 2024.
2. Toubia et al. Twin-2K-500. arXiv:2505.17479, 2025.