Electric Twin: Platform Accuracy

What Electric Twin does

Electric Twin addresses a fundamental challenge in behavioural science: the inability to make reliable and reproducible predictions about human behaviour. The platform uses survey data from **leading global survey companies** and large language models to create **representative synthetic populations** that can predict how real-world populations would respond to various scenarios.

Key applications include:

- Test market responses before product launches
- Assess public opinion on complex topics
- Evaluate policy impacts across different demographics
- **Product teams**: Test personalized experiences and predict user engagement
- Understand consumer preferences and decision-making patterns
- **Marketing teams**: Evaluate campaign effectiveness across audience segments
- Sharpen brand messaging and communications

How it works: synthetic audiences

We create synthetic audiences using large language models (such as GPT-5 or Claude) that adopt specific characteristics based on survey data including demographics and contextual information. We construct the audiences based on information from the survey data, combine this with the responses the model generates in order to produce a representative distribution of the question being asked. Each audience functions as a synthetic agent capable of responding to questions in ways that mirror real individuals from target populations.

STEP 1

Construct the audiences based on information from the survey data

STEP 2

Combine this with the Large Language Model inference

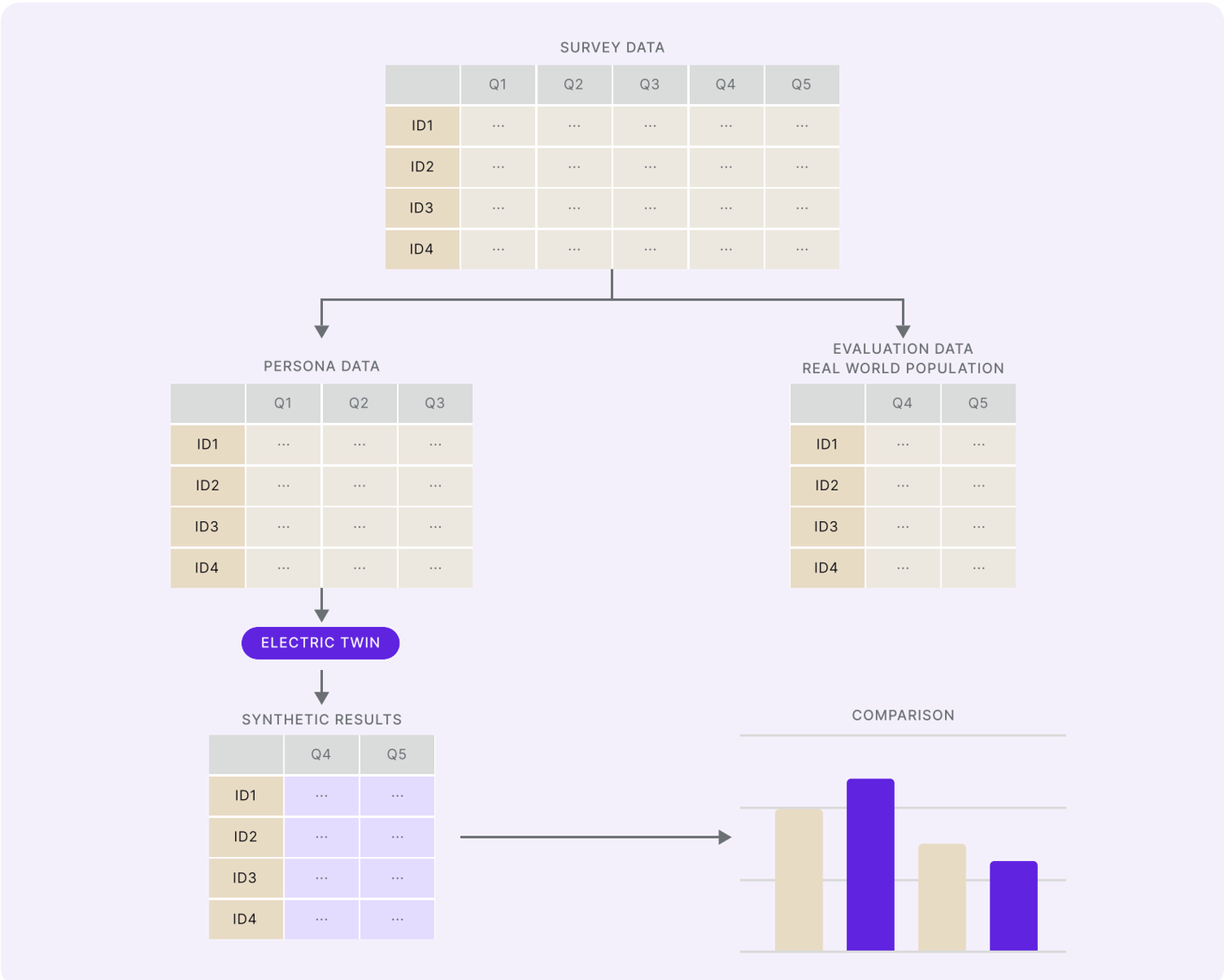
STEP 3

Produce a representative distribution of the question being asked

Quantitative evaluation

Validation methodology

Electric Twin employs strict data separation to prevent data leakage and ensure unbiased accuracy assessment:



01 **Dataset partitioning**: each dataset splits into:

Audience creation data
Used to construct synthetic personas with demographic and contextual information

Evaluation data
Completely held out during persona creation and used solely for testing predictions

02 **Hold-out testing**: Evaluations questions are never included in persona creation and are 'held-out', ensuring that the personas generate responses based on their characteristics rather than memorised answers.

03 **Comparative analysis**: Synthetic populations answer the same held-out questions, then responses are compared directly to real-world population answers.

Internal survey evaluation

We use **NDAM** (an accuracy score to compare distributions). We have tested across internal datasets spanning UK and US populations. We have improved an average of 20% across time attributing improvements to:

01 Better persona construction techniques, including information which is most helpful only

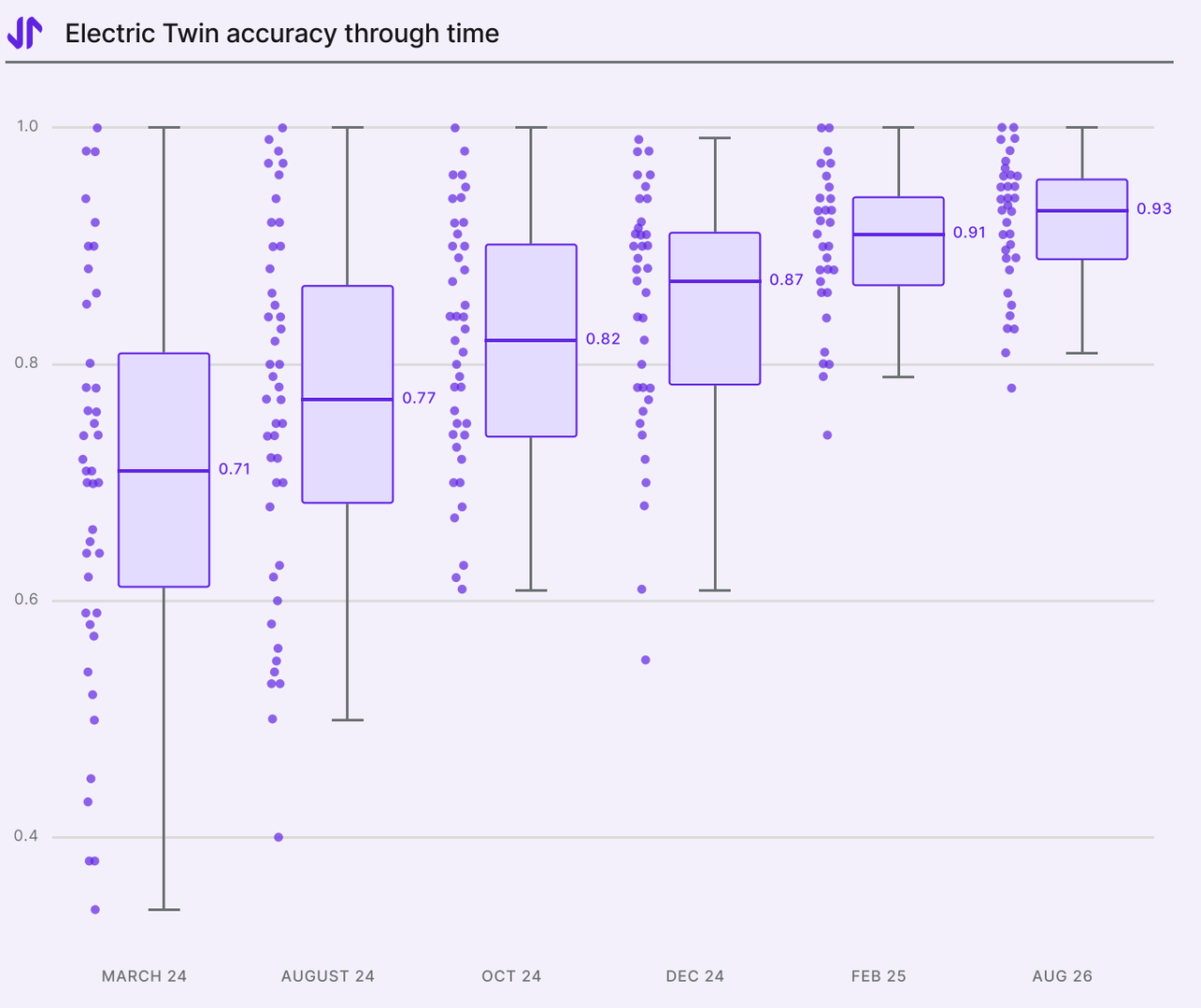
02 Better post-processing of LLM outputs

03 General model improvement through time from APIs

92% VS **94%**

ELECTRIC TWIN · NDAM — HUMAN NOISE LEVEL

Our current accuracy score sits at 92% NDAM, compared to the human noise level of 94% (asking the same question twice in a survey).

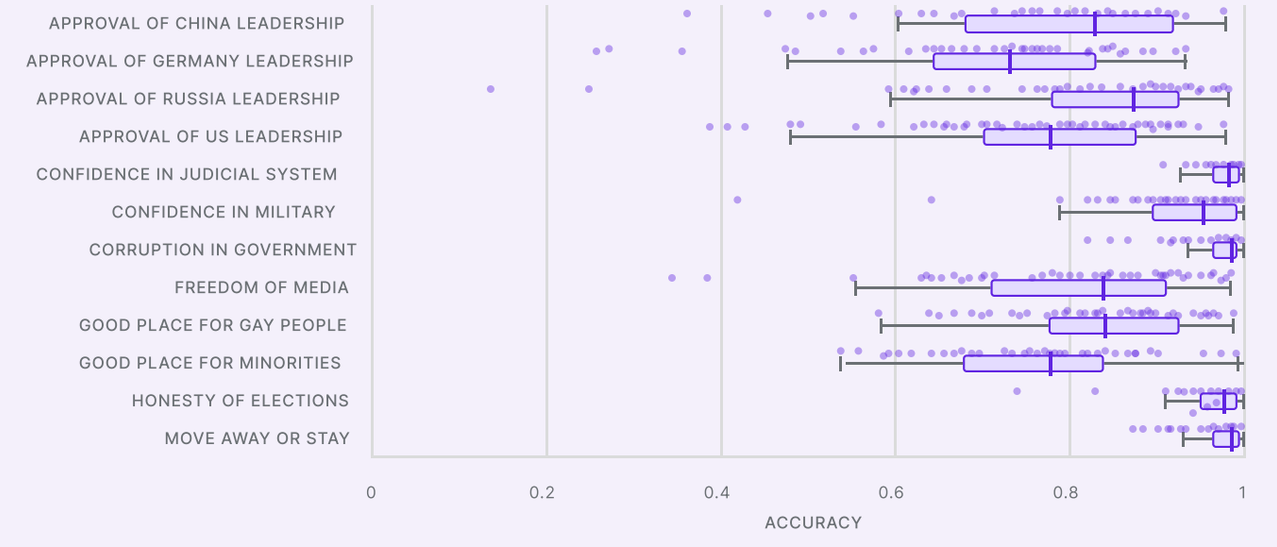


Comparison to advanced LLMs only

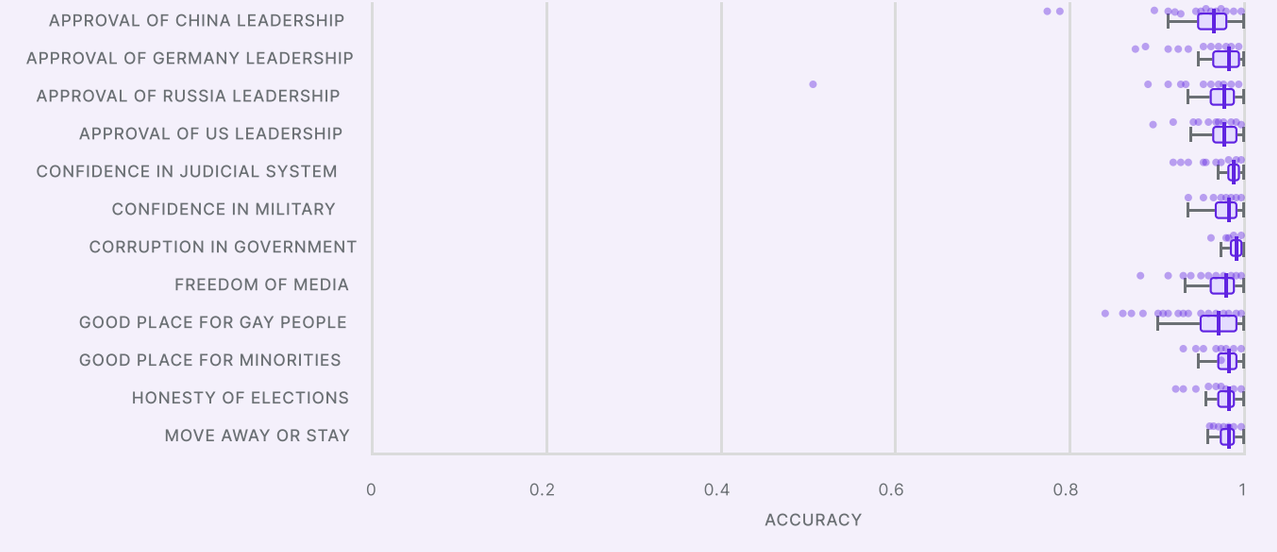
When compared to asking ChatGPT to predict survey response distributions, Electric Twin demonstrates i) consistently higher average accuracy across all questions and countries using the **Gallup World Poll**, ii) much smaller variance in predictions. Here we compare the distributional accuracy (based on Mean Absolute Error) between ChatGPT results and our internal Electric Twin Engine results, **with Electric Twin winning across all question categories and on average across all countries.**

This comparison validates that Electric Twin's proprietary methodology provides **substantial improvement over baseline LLM capabilities**.

ChatGPT results



Electric Twin results



Visual asset evaluation

Electric Twin's synthetic audiences successfully predict real-world human reactions to visual assets with a mean **NDAM of 85.8%** (Images = 86%, Videos = 85%). Performance was consistent across industry and question types.

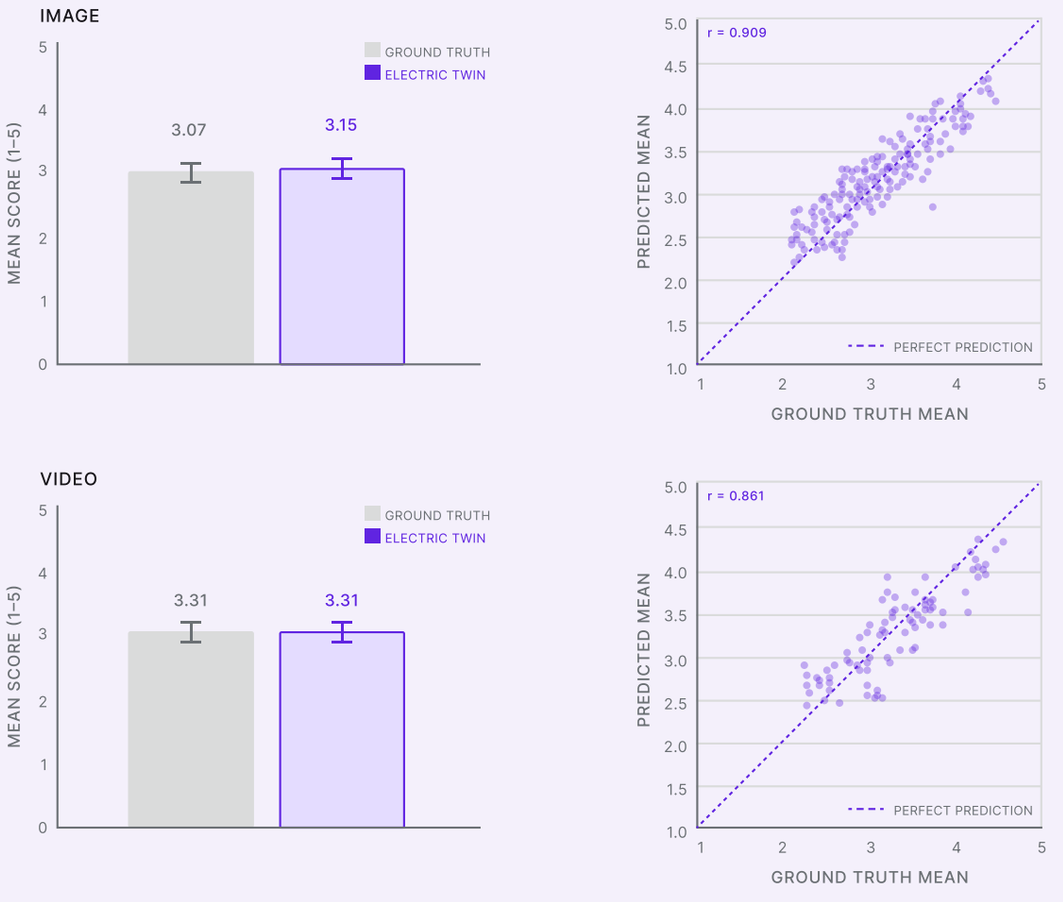
This was validated against a large-scale human benchmark of **4,000 participants**, evaluating 18 images and 6 videos up to 3 minutes in length, with strict data isolation enforced; no opinions or attitudes about the images were used to construct personas.



Beyond distributional accuracy, synthetic predictions show consistent alignment with human sentiment, showing a minimal difference in mean Likert scores and a **strong correlation between real and synthetic responses**.

Visual asset evaluation: sentiment alignment

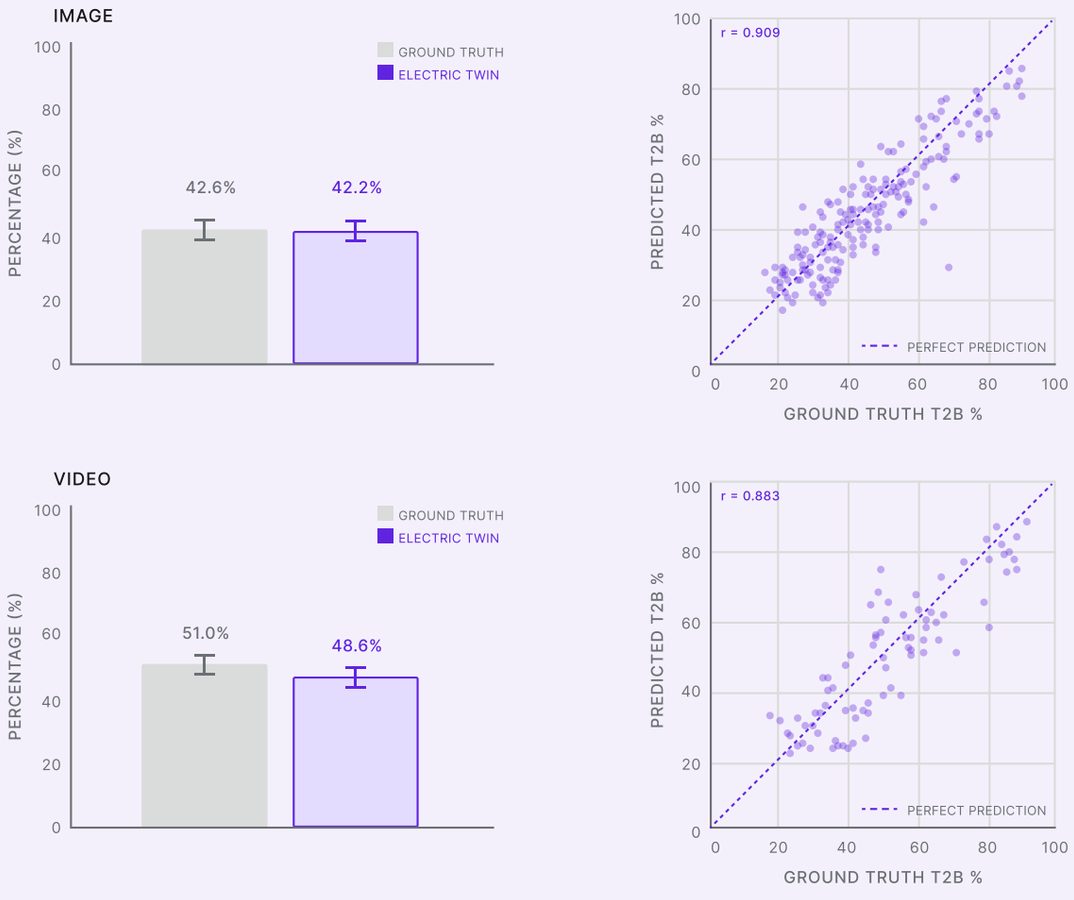
Mean Likert score: ground truth vs. Electric Twin



This was mirrored for **Top-2 Box Scores**, where models estimated this metric with a degree of accuracy within **0.3% for images and 1.4% for video**.

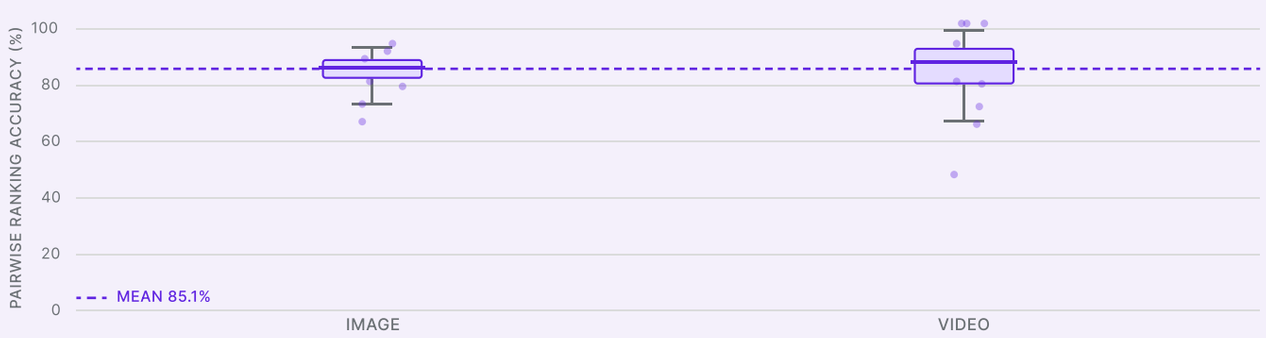
Visual asset evaluation: preference & ranking

Top-2 Box score: ground truth vs. Electric Twin



In head-to-head comparative matchups where meaningful consumer preference existed (more than 2% difference in Top-2 Box score), our personas correctly identified the "winning" image in **~85% of instances**.

Pairwise ranking accuracy (%)



Qualitative evaluation

Debate feature

Debate uses AI-powered **synthetic audiences** to simulate focus group debates, aiming to bridge the gap between scalable quantitative surveys and expensive qualitative focus groups.

By having synthetic participants debate topics over multiple rounds with an AI moderator, the Debate feature generates qualitative insights — themes, drivers, and representative quotes — that can be compared against real population data.

Debate is evaluated on two dimensions: whether the topics it surfaces match real survey distributions, and whether its synthetic quotes sound authentically human.

Quantitative match

Reverb's qualitative outputs were compared against nationally representative UK surveys across politics, climate, media, and consumer topics. The same datasets as the quantitative evaluation above. We achieve an average **NDAM score of 0.88**, demonstrating strong alignment between the topic distributions emerging from synthetic debates and actual polling data.

Turing test of voice

Three scenarios (UK issues, AI banking, football club ownership) were presented as blind comparisons to AI judges (Gemini 2.5 Pro, Claude 3.7 Sonnet, ChatGPT 4.5). Notably, in **6 out of 9 judgments** the synthetic group was rated as more authentic than the real one.

Focus group

We evaluate focus groups differently to Debate because different outputs are exposed. Here, we expose the free-text conversation so the focus is on evaluating these directly.

We generate answers from 15 focus groups with 5 participants each across each of our 9 whitepaper datasets (same questions as the Internal survey evaluation). The personas are again created only with a hold out dataset. We pose the evaluation questions to the focus groups without restricting them to what they would answer. Once we have the free text answers, we convert these to quantitative versions based on what the original survey options are. This is then compared to the ground truth with NDAM. The average score across the same questions as the quantitative score (0.92% NDAM) above is **0.86% NDAM**.