SocialWorldBench: Benchmarking Social Understanding and Prediction

Anonymous authors
Paper under double-blind review

Abstract

Can AI understand a conversation and predict how it will end? We introduce SocialWorldBench, a benchmark of 19,107 questions drawn from six existing datasets. Its main test uses 396 human negotiation dialogues. For each dialogue, a model reads the same conversation excerpt and answers two questions: what negotiation strategies are being used, and how satisfied will a participant be afterward? We compare the satisfaction prediction with what the participant later reported. This lets researchers examine whether models that recognize negotiation strategies also make accurate predictions.

We test five large language models from one provider on a 1,401-question sample and the paired negotiation task. When scored by answer-label overlap (Jaccard), three models identify strategies worse than a simple method that always gives the most common answer. Changing the evaluation questions also reverses two models' rankings, even though their answers stay unchanged.

SocialWorldBench provides researchers with an evaluation dataset to test whether AI can understand conversations and predict their outcomes. The dataset is deposited on Zenodo (anonymous preview link).

1 Introduction

A conversation supports two questions. A static question asks what is already true in it; a forecasting question asks what happens next. We release SocialWorldBench, which asks both of the same conversation, stopped at the same point, and scores the forecast against what the people in it later did or reported. It is 19,107 items over six licence-cleared corpora, released with its adapters, manifests, audit and harness. We asked whether a model that reads a social situation well also forecasts how it goes. We could not find such a benchmark, so we built one.

One item, asked twice. SocialWorldBench puts both questions to the same episode: one conversation, cut at one turn, taken from one annotation source (§2.2). The cut is the point where we stop showing the conversation to the model, so it cannot read the answer off a later turn. A static and a forecasting question over one episode make a paired axis. In CaSiNo, our anchor corpus, two people negotiate over camping supplies. Item casino:0:t0010 cuts one such negotiation after eleven turns, ending "Let's shake hands on this: 3 firewood, 1 food for me… Is that a deal?" Of that one prefix we ask both questions.
Static (Intent). "Which negotiation strategies does mturk_agent_1's final turn use? List every one that applies." Gold: small-talk, from a ten-label inventory.
Dynamics. "After this negotiation closes, what satisfaction will mturk_agent_1 report?" Gold: 4, on the source's 1–5 scale.

Only the question differs, so the stimulus is paired. Each answer stays in its native format, the answer type and scale the source corpus already used, not rewritten as multiple choice. The forecasting gold answer here is the number the named participant filled in afterwards, not a label an annotator inferred. The two halves are not always about the same person: the static half is about whoever spoke the turn at the cut, the forecasting half about a surveyed participant, and they coincide in 191 of the 396 pairs (48.2%), this example among them (§2.2). In v1 that forecasting axis rests on one corpus, CaSiNo being the only cleared source

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

that reports outcomes from the people who took part. That thinness, and a baseline pool of five models from one provider, are priced in §6.

Existing static benchmarks ask only the first question. The flagship multimodal benchmark shows how complete the absence is. We screened all 8,817 Social-IQ 2.0 questions (Wilf et al., 2023) with a twelve-pattern filter for future-oriented wording (will, going to, what happens next, predict, and eight more). It returned 93 candidates; we read every one. All 93 were false positives, spatial or retrospective (Appendix R). The screen found none. A keyword screen only tells you about the questions it matched, so we also read 300 questions drawn at random from the 8,724 it skipped, and found none there either ($\leq$1.0% of that remainder, 95% one-sided; Appendix R).

The resources that do ask both ask them of other material. Several benchmarks have two tracks: one of comprehension questions and one of prediction questions (Kong et al., 2026; Tang et al., 2026; Zhou et al., 2025b). Two relate the two axes directly (Yadav et al., 2026; Lubrano et al., 2026). But Yadav et al. forecasts from a mental-state profile with the transcript withheld, and Lubrano et al. never holds one context fixed across the two questions. Holding one scenario fixed while varying the question type is not ours either, and we admit that plainly. Gu et al. (2026) ask a mental-state, a behaviour-prediction and a behaviour-judgement question of each story, with "the underlying scenarios …identical" across the three. So we do not claim that design. What differs here is the material. SimpleToM's scenarios are written by its authors, and its answers with them. A forecast can instead be scored against what a real participant went on to report.

We are also not claiming to have discovered the gap between static and dynamic performance. Six groups reported it before us: strong static social performance need not carry over to dynamic settings (Zhou et al., 2024; Gu et al., 2026; Li et al., 2025; Yadav et al., 2026; Gong et al., 2026), with priority to Riemer et al. (2025). It is our starting assumption, not our finding. Where the relationship has been quantified, the published associations vary with how the two axes are defined (Momentè et al., 2025; Hu et al., 2026; Ziems et al., 2026; Yadav et al., 2026; Lubrano et al., 2026). No two of those studies share items, model pool, statistic or dynamic-axis definition. So that variation is evidence about measurement rather than about models, and we do not try to settle it. The phrase "social world model" belongs to Tsvilodub et al. (2026) and Tang et al. (2026); we claim neither the phrase nor their integration hypothesis. Table 8 sets all seven axes beside each cited paper.

Why the items can be believed. A question is worth scoring only if answering it needs the reasoning it claims to test. A head is all items from one source sharing a question type and answer format. We attack every head with policies that do no social reasoning: answer the commonest label, answer without reading the conversation, copy a phrase from the passage. A head those policies partly solve is quarantined: dropped from the scored sets rather than reported, 3,064 items over four heads in this build. An audit bounds only the policies it runs (Feng et al., 2019), and ours caught itself recording an absence of measurement as a clean result, which §3.3 reports with the repair. Which items get scored is not innocuous either, so every set ships as a checksummed manifest (§3.4).

Contributions. (i) SocialWorldBench v1 (§2): 19,107 items over six licence-cleared corpora in four sets — All, Core, Gem, Dynamics. Every item states where its conversation was cut and keeps its target in the source's native format, and results are reported per source, with controls, manifests and harness. (ii) The paired set (§2.2): 396 independent CaSiNo pairs carrying both questions over byte-identical context, one gold an annotator's label and the other a surveyed participant's own report, with an action-conditioned forecasting track that asks what happens if a named next move is made, from a fixed move library and against three controls (§2.3). (iii) The construction method (§3): trivial policies run over every head on every build, the quarantine they produce, and the measurement that makes freezing the item list a requirement. (iv) Baselines for five models on the balanced Core sample and on the paired set (§5), each read against that head's measured floor, the score reachable on a head without doing the task.

Source	Mod.	Capabilities	Licence	All	Gem	Dyn
Included in v1						
CaSiNo (Chawla et al., 2021)	T	Intent, Struct, Dyn	CC BY 4.0	1,471	176	1,004
FANToM (Kim et al., 2023)	T	ToM, Struct	MIT	219	0	0
OpenToM (Xu et al., 2024)	T	ToM [§], Dyn	CC BY-NC 4.0 [†]	13,573	161	596
SOTOPIA (Zhou et al., 2024)	T	Interactive, Intent, Struct	CC BY-SA 4.0 [*]	450	0	0
MELD (Poria et al., 2019)	T [‖]	Affect	GPL-3.0 [‡]	2,601	146	0
Social-IQ 2.0 (Wilf et al., 2023)	T [‖]	Percept, Intent, ToM, Affect	MIT [‡]	793	17	0
Total				19,107	500	1,600
Overlap (Gem ∩ Dyn)	156					

Audited and excluded (modality at source; none of the three is served, so none has a served modality). Social Genome (Mathur et al., 2025) (V+A+T; Percept, Affect, Intent, ToM): ground truth withheld by design, EvalAI-only. Social-IQ 1.0 (Zadeh et al., 2019) (V+A+T; Percept, Intent, ToM, Affect): no working access route as of 2026-08-31. CraigslistBargain (He et al., 2018) (T; Intent, Dyn): licence unconfirmed on primary sources.

Table 1: Sources, modality and composition. Counts are items, not episodes. Dyn counts track items (1,600), split 1,400 Dynamics / 200 Affect (Appendix P.1). Mod. is the modality actually supplied to the model at evaluation, not the source's (T = text/transcript): every row is served as text, so no score in this paper came from a model that saw video or audio. Appendix P.1 carries the rest of this caption (§6). [§]reported per dataset but excluded from the ToM capability mean for want of a human anchor (Appendix P.1, §2.4); [†]non-commercial; its README also forbids training use, which the OpenToM adapter enforces by serving no training split at all (Ethics Statement); [*]covers SOTOPIA's data; its harness code is MIT; [‡]covers the annotations only, granting no rights in the video; [‖]source modality V+A+T (video, audio and transcript), served as text.

2 SocialWorldBench

SocialWorldBench puts six licence-cleared corpora behind one interface. An Episode is one conversation, stopped at the observation cut t_{obs} of §1. A Query is one question about one episode; it declares which capability it tests and which modalities and span it may see. An episode carries several queries, so every count below counts questions, which we call items, not conversations. Audits and repairs act on a whole head at once.

A capability is the kind of judgement a query asks for, fixed by where the evidence for its answer lies relative to the cut. State, Intent and ToM are static: their evidence is at or before the cut. Dynamics lies strictly after it. One cut therefore carries both kinds, which lets us pair them. A fifth, Interactive, is tagged on SOTOPIA and never scored. A Target is the gold answer, the answer the source corpus records as correct, kept in its native form and never rewritten as multiple choice. A Prediction is the model's answer, parsed back into that native form; one we cannot parse is scored as a failure. Label maps, versions and licences are in Appendices D and E; Appendix A tables every term of art.

2.1 What it contains

Table 1 is the inventory. The six corpora are cut into four tracks, each a named subset of the items released as one evaluation set. Core (7,500: OpenToM 4,432, MELD 2,601, CaSiNo 467) is the evaluation set. It is stratified on capability, modality and difficulty, not balanced on them and not by source: the pools cannot supply balance at this size, and the realized capability shares are 6.2%, 34.7% and 59.1%. The balanced artifact is the 1,401-item sample scored in §5 (467 per capability), the ceiling the manifest records for a capability-balanced Core. Each capability still rests on one corpus, a limit we report as one (§6). All (19,107) is everything we built, for coverage and sensitivity work, and the only track that keeps the 3,064 items in quarantined heads, each flagged there and barred from every scored set (§3.3). Gem (500) is small, balanced and deliberately hard. Dynamics

(1,600) is every post-cut item, admitted by target function and not by any history-based test (§6).

2.2 The paired-episode design

The unit of measurement is the paired episode of §1: both questions over byte-identical context, with the same conversation, observation cut and annotation source. Domain, register and annotation practice are therefore common to both scores. In v1 the axis is the CaSiNo negotiation corpus, 396 independent pairs, one per dialogue: the static half asks which strategies the last turn used, and the forecasting half is scored against the satisfaction a named participant went on to record.

Ground truth differs in kind between the two sides and between sources: a participant's own report is not an annotator's label. Scores are therefore reported per source, never pooled and never subtracted. The static half's majority-class baseline is 0.5152, from the input-ignoring policy that always answers non-strategic, and the best copy-from-the-prefix policy reaches 0.0535, far below it. The labels are annotation jargon (uv-part, vouch-fair) that no negotiator types, so this half cannot be answered by reading the input back, which qualifies it to carry the pairing. The two halves' referents, the last speaker and the surveyed participant, coincide in 191 of the 396 pairs (48.2%), and that share ships as a covariate. How this head was chosen is §3.2; per-source composition, the kind of gold on each side and the survey-field leak audit (0 of 792) are in Appendices P and O.

2.3 The action-conditioned forecasting track

One track asks a conditional question the paired design makes available: if this person makes this move next, what happens to the other? The move comes from a fixed library of 14 moves, not free text. Three controls each remove one kind of information: the input, the move, or the match between the move and the episode. A forecaster whose score is unchanged by the last is not using the move (all three: Appendix O.1). The track is instantiated observationally, on cleared human negotiations with annotated moves and self-reported outcomes (Chawla et al., 2021). It therefore measures conditional prediction, not causal effect: the speaker chose the move, and whatever made them choose it also shapes the outcome. The interventional counterpart, a randomized dyad pilot with four arms per decision point, is specified and released uncollected (Appendices B, O, C and O.1).

2.4 How a score is reported, and what v1 cannot report

Datasets are scored on different metrics against different floors, so two raw scores cannot be averaged. Normalization rescales each dataset's score into an N_d where "no better than ignoring the input" is 0 and "as good as the human reference" is 100, from that dataset's chance floor and its human anchor. A human anchor is human performance on the task as we pose it and on the same metric; it is not the inter-annotator agreement a source paper reports. The formula is in Appendix P.1.

A dataset with no human anchor has no m_{human}, so N_d is undefined and that dataset is left out of the mean, not entered as zero. In v1 that is every dataset, so we report no normalized aggregate. Five of our seven adapters declare no anchor, and the two that do cannot supply one: FANToM's items are entirely quarantined and SOTOPIA was never scored. Two of the five declare none only because their anchors were withdrawn for this release, neither being substantiable as human performance on our adapted tasks (Appendix T.1 names both and says what each was). Every capability on Core, Dynamics and Gem is therefore absent from the macro, the mean of the per-capability means of N_d, at 100% of each track by item count; the gap is in our evidence rather than in N_d, and Appendix P.1 says what would close it. Hence the rule this paper obeys: results are primary per dataset, on its native metric, against its own floor, beside its human anchor where one exists, and with its n (Table 2); the macro is secondary and unreported in v1.

3 How it was built

3.1 Adapters and admission

Sources are admitted on four criteria. First, a licence we could read from a primary source. Second, annotations that preserve the capability we want to test. Third, native modality: each corpus is used as recorded. Fourth, pairability: enough episodes that carry both halves of a pair. One adapter per corpus maps it onto the interface of §2: a small program that turns the corpus's own file format into our episodes, queries and targets. It declares each head's answer space, converts nothing to multiple choice, and refuses what the licence refuses.

3.2 Choosing the paired head, and what it was chosen over

The paired head of §2.2 beat an alternative, and its audit trail ships with it. It realizes 792 raw pairs, one per participant, and 396 independent pairs, one per dialogue. It therefore fails, by 604, the 1,000-pair yield threshold this design was first written against. We report that failure and retire the threshold, not the head. The threshold was never a power criterion, for the reason §5.2 gives: the statistic is computed at the model level.

The head that does clear it is disqualified by our own quarantine rule and ships as a secondary head; the switch is why 48.2% referent coincidence ships as a covariate (Appendix P). The manifest, the checksummed file recording what a build contains and how each head was tested, returns null for the string battery on our own headline head; that battery does not run on multi-label targets. That is a declined measurement, not a clean one. It is the failure mode §3.3 describes.

3.3 Quality control

The procedure is borrowed. A partial-input baseline answers an item with part of its input deleted; if the score stays high, the item did not need what was deleted. Such baselines diagnosed natural-language inference and reading comprehension years ago (Poliak et al., 2018; Gururangan et al., 2018; Kaushik & Lipton, 2018); later work generalized them into ablation-based dataset auditing, attacked single-token cues in options, licensed treating any single-feature correlation with the label as spurious, and supplied the caveat the practice turns on. Social and mental-state benchmarks inherited the concern more than the procedure: they name shortcuts and leakage as standing problems and stress-test by asking the same reasoning several ways and by hand-built minimal pairs. Those are careful checks, not input ablations. Appendix R.8 gives that lineage attack by attack, naming every paper and its contribution. What is ours is only that the audit runs on every head, on every build, into the released manifest. A head recoverable from its own prompt is therefore a number a user reads, rather than a judgement we made on their behalf.

What the audit does, and what it quarantined. We attack every head with the trivial policies of §1 and compare each policy's score with the head's floor, measured on that head's own items. That gap is the margin. A margin above 0.05 sends the head to item-level repair, which deletes offending items until the trivial policy is back at the floor; above 0.15 it is also marked loud in the manifest; above 0.25 the head is quarantined. These three tolerances are operating conventions we chose, not significance tests. None is a critical value, none has a null hypothesis behind it, and crossing one is not statistical evidence of anything.

What that removed from this build. A Dynamics item answerable without its history would manufacture the very pairing this benchmark exists to support. Six filters (Appendices P.2 and H) attack shortcut-solvable items, but only two test history and they removed 113, none from a Dynamics source. Where the artifact is distributional instead, meaning a property of the head rather than of one item, the whole head is quarantined: 3,064 items over four heads in this build, retained in All and flagged, barred from every scored set. Two of those four crossed the 0.25 margin. The other two are the OpenToM fine-grained location heads,

which carry 2,845 of the 3,064 items; we quarantined them by explicit decision, not by the threshold, and the manifest records which basis applied to each. A score must also be read against something, and our own sources do not supply human agreement statistics (Appendix P.1). Following Nangia & Bowman (2019) and Tedeschi et al. (2023) we therefore read every score against a measured floor rather than $1/k$, and we leave an unanchored dataset out of a normalized mean rather than entering it as zero (§2.4).

The audit's own blind spot. An audit bounds only the policies it runs, and ours recorded an absence of measurement as a clean result. Two OpenToM location heads scored exactly 0.0000 under every policy of the previous family. That reads like a clean head. It was not: every policy in that family needed a list of answer choices, and both heads are open-ended, so none could run at all. A copy-from-prefix family, which needs no choice list, then beat both floors by +0.2108 and +0.1972, and both are now quarantined, part of the 3,064. The family that caught this did not exist two builds ago. That is what any such audit is worth: a lower bound on triviality, refreshed per build, with the manifest recording which battery tested each head, so a clean head and an untested one can be told apart. Exposure is measured, and it is total: every item was public before every model we evaluated. The model-level contamination probes were never scored (Appendix H), so contamination in those models is unmeasured, and this benchmark is never contamination-free.

3.4 Why the item list is frozen and versioned

Which items a track contains changes both the ordering of models and their absolute scores, and we measured that on our own stored outputs. One set of responses, bought once, was replayed through the same prompt, parser and scorer against two builds, each a frozen version of the item lists. Nothing about the models changed between the scorings. On CaSiNo and MELD the ordering of the five models is identical across builds, and no pair reverses.

On OpenToM the re-freeze alone reverses one pair: Claude Sonnet 4.6 leads Claude Opus 5 by +0.0729 native macro-F_1 on the earlier build's overlap, and Claude Opus 5 leads by +0.0091 on this one's. The absolute scores move much further than the reversal does, because this build quarantines the two open-ended OpenToM location heads; the per-model before-and-after pairs are in Appendix T.13. The largest such move on this dataset is +0.2388, for Claude Opus 5, from 0.5731 to 0.8119. A swing that size in a native macro-F_1 with the model held fixed is what a quarantine costs. All of this is native per-dataset scoring: no aggregate is involved and none exists (§2.4).

Which build a number belongs to is part of the number. The earlier build fails prompt verification, the check that the stored prompt still matches the one the adapter builds today, on 622 of its 1,333 overlapping items, all MELD items predating the context fix of §6; this build verifies clean on all 1,390. So every item list is fixed by a checksummed manifest, and every baseline in §5 names its build (Appendix T.13).

4 Experimental setup

Models. We run five models of several sizes, all served by one provider. Each is named by its served version string: claude-haiku-4-5, claude-sonnet-4-6, claude-sonnet-5, claude-opus-4-8 and claude-opus-5 (Appendix Q lists run dates and cost). Every completed run is the static overlay: the static half of every paired episode. The Dynamics overlay is a second pass over the same episodes, run for zero models.

Protocol. We freeze one prompt template per target type, one decoding configuration and one deterministic parser before any full-scale run. Nothing is tuned per model (Appendix F). Each run names the frozen build it scored, for the reason in §3.4.

Scoring. Each kind of target has one metric. Targets with a fixed label set are scored by accuracy and macro-F1 over the source's own labels. Macro-F1 averages the labels' F1 scores with equal weight, so a rare label counts as much as a common one. Free-text targets

Dataset	Native metric	Floor	Human	n All	n Core	Core range	In macro?
CaSiNo (static)	strategy macro-F1	0.2148	—	667	467	0.3695–0.6495	no anchor†
CaSiNo (ACSF)	Likert CCC, mean	0.0000	—	804	0	—‡	no anchor
FANToM	accuracy	0.5000	0.93	219	0	—‡	all quarantined
MELD	weighted F1	0.3100	—	2,601	2,601	0.5389–0.6581	no anchor†
OpenToM	macro-F1	0.2868	—	13,573	4,432	0.4192–0.8482	no anchor
SOTOPIA	interactive reward	2.0	8.0	450	0	—‡	unscored*
Social-IQ 2.0	accuracy	0.2500	—	793	0	—‡	no anchor

Table 2: Primary reporting is per dataset, on that dataset's own metric, read against the floor in the same row, never against zero, and never pooled down the column: column two lists six incommensurable metrics, and their mean is not a quantity. Core range: the five models' lowest and highest native scores on the 1,401-item balanced Core sample (467 items per capability, drawn from the 7,500-item frozen Core track of build 139722f4), scored on each dataset's own native metric against that dataset's own floor and never pooled (grid: Table 10). ‡Row with no items on that track: not a pending number. The last column reports eligibility for the capability normalization (§2.4), not that a score exists: five of seven adapters, 18,438 of the 19,107 All items are absent from the macro rather than entered as zero. †anchor withdrawn as unsourced (§6); *anchored yet never scored; no dataset here can enter a capability mean and v1 reports no normalized aggregate. The footnotes in full, and what the frozen build predates: Appendix T.1.

are scored by match against a reference set. Distributional Dynamics targets are scored by a proper scoring rule: a model maximises it by reporting true probabilities, not by guessing confidently.

Every number in §5 is a native score read against a measured floor, which needs no anchor (§2.4). No capability mean is reported in v1, because normalization needs a measured human score and no dataset has one. Metrics, floors and weights: Appendix G.

Floors and controls. Ten degenerate baselines ship, each the protocol with one kind of information removed (Appendix Q). Random and majority are deterministic policies; they were run and fix the chance floor every score is read against. Answer-only and text-only withhold the episode and the nonverbal channels respectively. Null-history deletes the history entirely, the last utterance included, which last-utterance-only keeps. Shuffled-history permutes the history, separating use of the history from use of its order. Transcript-only, name-removed and temporal-shuffle remove side channels, speaker names and timestamps. No model has been run under any of the eight context controls: the harness implements them, the paid sweeps did not run them, and no claim here needs one.

Expert reference and statistics. Our expert reference is two human raters on 139 items stratified from Gem and Dynamics. We report it with its counts and the raters' agreement, never as human performance (Appendix I). It is not drawn from the shipped benchmark: most of it predates the stimulus fix of §6, its 62 MELD items included, and anchors nothing. Intervals come from a bootstrap, which resamples the scored data many times to see how far a number moves. Questions from one conversation are not independent, so the paired and action-conditioned intervals resample whole conversations, not single items: a cluster bootstrap. The item-set intervals resample items, leaving that dependence uncontrolled. The Holm correction raises the bar each comparison must clear, so one lucky comparison among many does not count as a result.

5 Baseline results

5.1 Per-dataset baselines

Every model clears every floor and no dataset is finished. Nothing in Table 2, the primary result table, is averaged down the column, and there is no capability macro to average into

(§2.4, Appendices T.1 and T.3). All 15 cells sit above their own floor. Each floor is a constant of its source corpus, not of the scored items. The best CaSiNo and MELD scores are 0.6495 and 0.6581, on floors of 0.2148 and 0.3100, while OpenToM's 0.8482 hides a finished head (Appendix T.2). The three datasets disagree about which model is better. 6 of 10 model pairs are ordered differently by at least two of them. Claude Haiku 4.5 is best of five on MELD and worst on both CaSiNo and OpenToM. A mean down the column would report one ordering and erase the others. MELD is the flat one, with 0.1192 between best and worst against 0.2799 and 0.4289. Its re-rating found the same: two experts with the full dialogue still find about three quarters of its items underdetermined from the transcript (§6, Appendix T.4.1).

5.2 The paired set

The paired set puts both questions to the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale. That is one corpus, so the axis is not cross-source. Each half is read against its own floor, and neither half is ever subtracted from the other.

The static head is not saturated, and not passed by answering the commonest label. It is a multi-label strategy head, scored by Jaccard overlap of predicted and gold label sets. Its baseline is a majority-class rate, not a guessing rate. The gold label is non-strategic on 51.52% of the 396 items. A copy policy scores 0.0535 on this head and surface copy exactly 0.0000 (§2.2). Of the five models, one is above that majority rate, one is indistinguishable from it, and three are below. Two of the three are below even an oracle over their own emitted labels, so this is not an artifact of how many labels they name. That count is specific to Jaccard. It holds under Dice and micro-F_1. Four of the five fall below under exact set match, but under macro-F_1 the baseline is a far weaker 0.1666 and every model clears it (Appendix T.8 gives the five metrics and the oracle audit).

The axis is computable on the released artifact. The halves share the stimulus, so the two orderings of models can be set side by side. Over $n = $ five models the two orders agree at $\rho = 0.8000$. Resampling the conversations gives $[0.600, 0.900]$; resampling the models, $[0.111, 1.000]$. 2 of the 10 model pairs reverse, and zero survive the Holm correction for testing every pair at once. That is all this establishes. It is not a relationship between social understanding and social forecasting. What is paired is the stimulus; what is correlated is a model score, one (x, y) per model, so the 396 pairs set the precision of each coordinate and not its degrees of freedom. With five models from one provider the model-resampled interval is uninformative. Pairing equalizes the input material and nothing else. It does not equalize the constructs, their difficulty, their label reliability or either question's attainable score, and the halves' referents coincide in only 191 of the 396 pairs (Appendix T.11 re-aggregates on that split; the relationship is unchanged). Table 12, Figures 3 and 5 give both halves, the expert reference and the per-item view (Appendices T and T.12).

5.3 Action-conditioned forecasting

On the next move there is no statistically detectable difference between the full forecaster and a separately trained action-only predictor: −0.0064 Brier, 95% CI $[-0.0229, +0.0116]$. That interval spans zero, so it is consistent with the full forecaster being somewhat better and with its being somewhat worse. It does not show the two are equivalent. We ran no equivalence test and set no margin of practical equivalence, so no equivalence claim is available anywhere in this paper. Two point estimates, carrying no interval and reported descriptively only (Appendix T.14), say the same. Read together, these say that on this head, on this held-out split, and among the predictors we fitted, access to the conversation bought no measurable advantage over the move alone. They do not say the conversation is uninformative about the next move: a restricted model class failing to extract information is not evidence that the information is absent. The outcome head runs the other way, history buying +0.0987 CCC $[+0.0080, +0.1958]$, though over both heads it does not survive correction for testing two heads at once ($p_{\text{Holm}} = 0.062$). The move adds nothing to the outcome additively (+0.0036

CCC, spanning zero). An exploratory arm, outside the corrected family and not part of any claim this paper makes, fits the history–action interaction and beats the additive one, which is what the additive null is a null about: it constrains additive models, not actions. No language-model forecaster was ever run on this track: no API key was configured for it. So every number here compares learned (ridge) and trivial (constant, action-marginal) predictors only. The track is observational (§2.3): the speaker chose the move, so a score is prediction given a move, never its effect. The comparison that decides the track is the full forecaster against that separately trained action-only predictor, not the same model with its history masked. Table 15 (Appendix T.14) reports every forecaster and control on it, says why, and says how the intervals are clustered.

Two qualifications travel with that table, neither optional: the next-move gain is specific to the action that actually occurred, not to having any action, and every interval is clustered on the dialogue (§4), which moved two contrasts off significance (Appendix T.14 carries both in full).

Where prediction fails. We do not report the split. A score gap does not say what breaks, and no artifact in this release labels a Dynamics error by the paired static answer: the partition is a property of the paired design, not a measurement on it, and the taxonomy figure it would populate is withdrawn from v1. Even labelled, neither label locates a deficient transition model: the split would describe which errors co-occur with a correct reading, not attribute them (Figure 1; Appendix N derives it and why). Episodes read correctly and still mispredicted do not occur on this build (Appendix K).

6 Intended use and limits

SocialWorldBench compares systems per dataset against a measured floor (§3.4); v1 reports no normalized aggregate, so it is not for a headline number. One source reports outcomes from real people: CaSiNo, one domain, one MTurk population (§T.6), with both halves of a pair from one session, 191 of 396 from the same person (§2.2). The balanced Core sweep is the only on-track run: 1,401 of 7,500 items, five models from one provider (§5.1). No human ceiling is measured: two authors rated a packet of which 16 of 140 items reach the shipped build (Appendix T.4.1). The two paired halves have incomparable scales (§5.2), so no static-minus-forecasting number is reportable, and the Dynamics track has been run for zero models. No model has been run under a control condition (§4), we report association, not cause, and no collection assigned a move, so every action-conditioned result is observational (Appendices S, S.18 and S.17).

Reproducibility Statement

Every number in the tables and macros is written by a generator reading a frozen run artifact; every other number is typed beside a comment naming its artifact. We release the evaluation code, checksummed manifests, exact prompts and decoding settings, and raw model outputs, each score with its provenance tuple (Appendix L). The dataset is deposited on Zenodo, reachable through an anonymous preview until the DOI is minted at publication; each record holds the scored prompt, source text and reference answer with SHA-256 hashes and its source's licence, and is for evaluation, not training. The code is not posted to an anonymous mirror: a scrubbed snapshot was verified but not uploaded before submission, and we do not cite a URL that does not resolve; the repository is released on acceptance.

Use of Large Language Models

Every rating reported as an expert reference was produced by a person, two authors working independently; no model answered or was averaged into it (Appendix T.4). Two model raters checked the packet for wording defects, which we repaired; no quantity they produced is reported (Appendix M). OpenToM, 71.0% of the 19,107 All items, states that its story drafts "are composed using LLMs" (Xu et al., 2024), so most item text here was model-

drafted upstream; FANToM's dialogues are model-generated. We generated none of it: our adapters are deterministic code, and no generative model wrote, judged or selected any item, option or gold label. LLM coding agents wrote the harness and analysis scripts under author review, the survey and licence audit were run the same way (one claim about that process is withdrawn in Appendix M), and this manuscript was drafted and edited with LLM assistance. The authors are accountable for every claim, citation and number.

Ethics Statement

The benchmark specifies an interventional probe with consenting dyads and releases its protocol (Appendix C), but no such collection was run: every result comes from existing public corpora under their own consent terms, predominantly English-language and Western-participant. Forecasting how a move changes another person's state is dual-use, supporting coaching and manipulation alike; the action library holds only verbal moves a person could openly be told to make, the forecast never conditions on private information about the partner, and we measure prediction, we do not deploy a recommender. OpenToM (CC BY-NC 4.0) forbids training and FANToM restricts its samples to evaluation; we fit no model on either, the deposit carries each record under its source's licence, and the only fitted forecasters use CaSiNo's own training split (CC BY 4.0). One use is not plain evaluation: the shortcut audit fits two one-parameter rules on the restricted sources that learn no weights and only remove items, which we read as curation and state so a reader can disagree.

References

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. ToM-SSI: Evaluating theory of mind in situated social interactions. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. URL https://aclanthology.org/2025.emnlp-main.1642/.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2917–2927, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Buch_Revisiting_the_Video_in_Video-Language_Understanding_CVPR_2022_paper.html.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021. URL https://aclanthology.org/2021.naacl-main.254.pdf.

Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2358–2367, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1223. URL https://aclanthology.org/P16-1223/.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, You Li, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, and Timothy T. Rogers. DEBATE: A large-scale benchmark for evaluating opinion dynamics in role-playing LLM agents, 2025. URL https://arxiv.org/abs/2510.25110.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5533–5538, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1554. URL https://aclanthology.org/P19-1554/.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. Competency problems: On finding and removing artifacts in language

data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1801–1813, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.135. URL https://aclanthology.org/2021.emnlp-main.135/.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107. URL https://aclanthology.org/D19-1107/.

Nanxu Gong, Zixin Chen, Haotian Li, Zishu Zhao, Jianxun Lian, Huamin Qu, Yanjie Fu, and Xing Xie. Does theory of mind improvement really benefit human-AI interactions? empirical findings from interactive evaluations. In Findings of the Association for Computational Linguistics: ACL 2026. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.2125/.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In The Fourteenth International Conference on Learning Representations (ICLR), 2026. URL https://openreview.net/forum?id=iE2JmbRJow.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

David Ha and Jürgen Schmidhuber. World models, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018. URL https://arxiv.org/abs/1808.09637.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems 28 (NIPS 2015), 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In The Fourteenth International Conference on Learning Representations (ICLR), 2026. URL https://iclr.cc/virtual/2026/poster/10009688.

Hume AI and the DaiKon organisers. The 2026 ACII dyadic conversations (DaiKon) workshop & challenge, 2026. URL https://arxiv.org/abs/2605.02672.

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5010–5015, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1546. URL https://aclanthology.org/D18-1546/.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. When choosing plausible alternatives, clever hans can be clever. In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pp. 33–42, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6004. URL https://aclanthology.org/D19-6004/.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 14397–14413, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. SIV-Bench: A video benchmark for social interaction understanding and reasoning, 2026. URL https://arxiv.org/abs/2506.05425.

Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 487–507, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29/.

Mengfan Li, Xuanhua Shi, and Yang Deng. RecToM: A benchmark for evaluating machine theory of mind in LLM-based conversational recommender systems, 2025. URL https://arxiv.org/abs/2511.22275.

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1):255–268, 1989. doi: 10.2307/2532051.

Kate M. Lubrano, Faisal Sayed, Ankita Rathod, Akshansh, Craver Corbyn Thomas-Smith, Mark E. Whiting, and Karina Nguyen. AttuneBench: A conversation-based benchmark for LLM emotional intelligence, 2026. URL https://arxiv.org/abs/2605.21739.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. Towards a holistic landscape of situated theory of mind in large language models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1011–1031. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.72/.

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. The MSP-Conversation corpus. In Proc. Interspeech 2020, pp. 1823–1827, 2020. doi: 10.21437/Interspeech.2020-2444.

Luz Martinez-Lucas et al. MSP-Conversation release 2.0, 2026. URL https://arxiv.org/html/2603.22536.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models, 2025.

Filippo Momentè, Alessandro Suglia, Mario Giulianelli, Ambra Ferrari, Alexander Koller, Oliver Lemon, David Schlangen, Raquel Fernández, and Raffaella Bernardi. Triangulating LLM progress through benchmarks, games, and cognitive tests. In Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 20051–20072, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1092.

Jared Moore, Ned Cooper, Rasmus Overmark, Beba Cibralic, Nick Haber, and Cameron R. Jones. Do large language models have a planning theory of mind? evidence from MindGames: a multi-step persuasion task, 2025. URL https://arxiv.org/abs/2507.16196.

Nikita Nangia and Samuel R. Bowman. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4566–4575, Florence, Italy, July

2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1449. URL https://aclanthology.org/P19-1449/.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://aclanthology.org/P19-1459/.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, 2020. URL https://arxiv.org/abs/2005.04120.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1145–1160, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.98. URL https://aclanthology.org/2021.findings-acl.98/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019. URL https://arxiv.org/abs/1810.02508.

Yunjin Qi, Zhaojun Jiang, Xuan Wu, Hanxi Pan, Yixuan Wang, Yanfang Liu, Xiang Ji, Churu Yu, Chunyuan Zheng, Yingze Chen, Jie He, Liuqing Chen, and Zaifeng Gao. NICE: A theory-grounded diagnostic benchmark for social intelligence of LLMs, 2026.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. In Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2412.19726.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition, 2019. URL https://arxiv.org/abs/1907.11510.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. Do neural dialog systems use the conversation history effectively? an empirical study. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 32–37, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1004. URL https://aclanthology.org/P19-1004/.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2257–2273, St. Julian's, Malta, March 2024. Association for Computational

Linguistics. doi: 10.18653/v1/2024.eacl-long.138. URL https://aclanthology.org/2024.eacl-long.138/.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. "you are grounded!": Latent name artifacts in pre-trained language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6850–6861, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.556. URL https://aclanthology.org/2020.emnlp-main.556/.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. UnNatural Language Inference. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7329–7346, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.569. URL https://aclanthology.org/2021.acl-long.569/.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. Assessing the benchmarking capacity of machine reading comprehension datasets. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 8918–8927, 2020. doi: 10.1609/aaai.v34i05.6422.

Weidong Tang, Jierui Li, Yueling Hou, Zihan Mei, Can Zhang, Xinyan Wan, Zhiyuan Liang, Pengfei Zhou, Yang You, and Wangbo Zhao. GroupToM-Bench: Benchmarking group theory of mind and nonlinear social emergence in MLLMs. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026. URL https://aclanthology.org/2026.acl-long.1859/.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. What's the meaning of superhuman performance in today's NLU? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12471–12491, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.697. URL https://aclanthology.org/2023.acl-long.697/.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In CVPR 2011, pp. 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

Polina Tsvilodub, Jan-Felix Klumpp, Amir Mohammadpour, Jennifer Hu, and Michael Franke. On emergent social world models — evidence for functional integration of theory of mind and pragmatic reasoning in language models. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 37382–37420, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1735.

Emilio Villa-Cueva, S M Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In Findings of the Association for Computational Linguistics: EMNLP 2025, 2025. URL https://aclanthology.org/2025.findings-emnlp.1230/.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. SOTOPIA-π: Interactive learning of socially intelligent language agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12912–12940, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.698.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. https://github.com/abwilf/Social-IQ-2.0-Challenge, 2023.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024. URL https://arxiv.org/abs/2402.06044.

Neemesh Yadav, Palakorn Achananuparp, Jing Jiang, and Ee-Peng Lim. DialToM: A theory of mind benchmark for forecasting state-driven dialogue trajectories. In Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2026. URL https://arxiv.org/abs/2604.20443.

Yuwei Yin, Chuyuan Li, and Giuseppe Carenini. Intentgrasp: A comprehensive benchmark for intent understanding, May 2026.

Yashwanth YS, Ruichen Wang, Shihua Zeng, Xuhui Zhou, Koichi Onoue, Vasudha Varadarajan, and Maarten Sap. SOTOPIA-TOM: Evaluating information management in multi-agent interaction with theory of mind. In Conference on Language Modeling (COLM), 2026. URL https://arxiv.org/abs/2605.02307.

Yu, Zhao, Lin, and You. Building social world models with LLMs, 2026a. URL https://arxiv.org/abs/2606.11482.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position, 2026b. URL https://arxiv.org/abs/2606.15032.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8807–8817, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009/.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zhang et al. See, infer, intervene: Proactive world modeling for goal-oriented social intelligence, 2026. URL https://arxiv.org/abs/2606.03371.

Bo Zhao, Fanghua Ye, Yixin Ji, Sicheng Zhao, Xiaojiang Peng, and Zitong Yu. AffectVerse: Emotional world models for multimodal affective computing, 2026. URL https://arxiv.org/abs/2605.19950.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. SocialEval: Evaluating social intelligence of large language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 30958–31012, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1496.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In The Twelfth International Conference on Learning Representations (ICLR), 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social world models, 2025b. URL https://arxiv.org/abs/2509.00559.

Caleb Ziems, William Held, Su Doga Karaca, David Grusky, Tatsunori Hashimoto, and Diyi Yang. Will scaling improve social simulation with LLMs?, 2026. URL https://arxiv.org/abs/2607.02464.

A Glossary

Table 3 defines every term of art this paper relies on, in the sense in which this paper uses it. Each entry stands on its own, so a term can be looked up without having read the rest of the paper; each is also glossed in one clause where it first appears in the body.

Table 3: Glossary. Definitions are this paper's usage. For head, human anchor and chance floor that usage is narrower than the general one. "First used" names where the term first carries weight, not every place it appears.

Term	What it means here	First used
Words about measurement in general		
Construct	The ability a score is meant to measure, as opposed to the score itself.	§1
Confound	Anything that varies alongside what is being compared and could explain the difference by itself. Here, chiefly which corpus each score came from.	§1
The material		
Corpus	One original collection of conversations that we reuse. v1 draws on six of them.	§1
Episode	One conversation segment, cut at a single moment t_{obs}. Everything before the cut is what a model is allowed to see.	Abstract, §2.2
Cut	The moment in a conversation where we stop showing the model, written t_{obs}. The model sees everything before it and nothing after it, so it cannot read the answer off the transcript. Every episode declares its own cut.	§2
Adapter	The code mapping one source corpus onto the common interface every item shares. It declares each head's answer space, converts nothing to multiple choice, and refuses what the licence refuses. A corpus can carry more than one.	§3
Native form	The answer format the source corpus already used. A Target keeps it and is never rewritten as multiple choice; a model's Prediction is parsed back into it, a parse failure scoring as a wrong answer. Scores are likewise reported on each dataset's own native metric.	§2, §4
The questions		
Item	One question put to one episode, with one gold answer. Every score is an average over items.	§3.3
Gold	The answer an item is scored against. On the static side it comes from an annotation the source corpus already carried. On the forecasting side of the paired set it is the number the participant filled in after the conversation.	§2, §2.2
Head	Every item from one source that shares a question form and an answer space — for instance all MELD emotion questions. Audits and quarantine act on a whole head, not on single items.	§3.3
Capability	What an item asks about: State, Intent or ToM (answerable from before the cut) or Dynamics (answerable only after it).	§2
Track	A named subset of the items that a model is run on: All, Core, Gem or Dynamics.	§2.4

Term	What it means here	First used
Paired design	Attaching a static question and a Dynamics question to the same episode and the same cut, so the two scores differ in what is asked and not in where it came from.	§2.2
Paired set	The episodes that carry both questions at one cut, over byte-identical context. The forecasting half of it is what the Cross-source entry below records the coverage limit for.	§2.2
Within-item	A property of the design, not of any statistic we report: both questions are put to the same episode, so the material does not differ between them. The relationship itself is estimated across models, one score pair per model, so no number in this paper is a within-item correlation. Listed because the term appears in the prior literature; the body never uses it.	not used
Cross-source	Repeating the paired comparison separately inside each of several independently collected corpora. v1 does not achieve this on the forecasting axis and no longer claims it: that axis carries a single corpus, CaSiNo, also the only one of our sources reporting outcomes from the people who took part.	§5.2
Checking that the items are worth scoring		
Trivial policy	A rule that answers without understanding: always pick the same option, or copy a word out of the prompt. Every build audits them.	§3.3
Shortcut	A regularity that lets a trivial policy beat the chance floor. Where it is a property of single items, those items are removed.	§3.3
Quarantine	What we do when the shortcut belongs to a whole head instead: the head stays in All with a flag on every item and is barred from Core and Gem.	§2.4
Degenerate control	The same evaluation rerun with one kind of information removed — the history, the video and audio, or the episode itself — to show what a score actually requires.	§4
Chance floor	What a model scores on a head without reading anything, by guessing or by always answering the commonest label. Measured per head by the random and majority controls; the mark every score is read against, and never zero.	§2.2, §4
Manifest	The released file listing exactly which items a set contains, with a checksum, together with each head's audit verdict — something a user reads rather than takes on trust.	§3.4
Frozen build	A numbered, checksummed snapshot of the item list. Which items are counted moves both the scores and the ordering of models, so every run names the build it scored: which build a number belongs to is part of the number.	§3.4
Item-set sensitivity	How much a score and a ranking depend on which items were counted, holding the model outputs, the prompt, the parser and the scorer fixed. Measured here by scoring one set of stored answers against two builds (§3.4, Table 14).	§3.4
Turning answers into reported numbers		
Macro-F1	F1 combines how often a predicted label is correct with how many true labels were found. Macro-F1 averages the per-label F1 equally over the source's own label set, so a rare label counts as much as a common one; plain accuracy does not.	§4
Human anchor	A published score for what people achieve on a dataset. Three of our seven adapters ship none, and two more declared a value we withdrew as unsubstantiable (MELD's 0.82, a placeholder never replaced; CaSiNo's 0.78, an undocumented constant), so five of seven cannot supply one — the count §2.4 uses. Nor can the remaining two: FANToM's items are entirely quarantined and SOTOPIA was never scored. Our own expert reference, two human raters on a superseded packet, is not a published anchor.	§2.4, §4

Term	What it means here	First used
Normalization (N_d)	Putting datasets on one scale as $N_d = 100 \times (\text{raw} - \text{floor})/(\text{anchor} - \text{floor})$, so 0 is the floor and 100 is the anchor. With no anchor this is undefined, and the dataset leaves the mean rather than entering it as a zero.	§2.4
Macro mean	The unweighted mean through the per-capability scores, taken only over capabilities that have at least one dataset with an anchor. v1 reports none on any track: no dataset it ships carries a usable anchor, so no capability score forms and the mean has nothing to average.	§2.4, §5.1
Never pooled	Raw scores from different datasets are never averaged into one number, their native metrics not measuring the same quantity. The macro mean above would not be an exception — it averages human-normalized scores — but v1 forms none, no dataset on any track having an anchor, so every score this paper reports is a native per-dataset one.	Abstract, §2.4
Rank reversal	A pair of models ordered one way by one measurement and the other way by another. Two distinct instances, and they are not the same finding: the item list reverses a pair at fixed model outputs (§3.4), and the static score orders a pair differently from the forecasting score (§5.2).	§3.4
Statistics		
Cluster bootstrap	A confidence interval built by resampling whole conversations rather than single questions. Questions about one conversation are not independent, so counting them as independent would make an interval look narrower than it is. Wherever a source asks many questions about one context, the resampling unit is the context.	§4
Holm correction	An adjustment applied when several comparisons are made at once, because some will look significant by luck alone. The step-down procedure raises the bar for each comparison in the set, controlling the family-wise error rate — the chance of even one false positive anywhere in it.	§4
Krippendorff's α	How far two or more raters agree, corrected for chance agreement: 1 is perfect, 0 is chance. Reported rather than Cohen's κ wherever raters vary per unit or ratings go missing at different rates; with exactly two raters and plain categories the paper reports κ instead.	Appendix I

B The action library and per-corpus label maps

The library has 14 moves in five families. Regulation (four): no-op, pause, delay the decision, suggest a break. Elicitation (one): ask a clarifying question. Alignment (five): acknowledge the other person's concern, reflect or summarise, offer two options, return to the shared objective, small talk. Disclosure (two): disclose one's own constraint, concede no need. Pressure (two): appeal to fairness, undervalue the partner's need. Two further partitions cut across those families. By origin, ten moves are the authored set the pilot arms would be drawn from and four are added because the negotiation corpus observes them (small talk, concede no need, appeal to fairness, undervalue the partner's need). By recommendability, thirteen of the fourteen may be suggested to a person and exactly one — undervaluing the partner's need — is non-recommendable while remaining scoreable, because the corpus contains it and a forecaster must be able to name it.

The negotiation corpus's strategy labels, and how each one maps. The map is a total function on CaSiNo's own ten-value strategy vocabulary, where every label maps, none is dropped, and an unrecognised label raises rather than being silently discarded. It is many-to-one at one place only — self-need and other-need both become disclose_constraint. Four library moves are reached only through this map and are the four of casino origin; the moves CaSiNo never expresses (pause, summarize_reflect, offer_two_options, delay_decision, suggest_break) are in the next-move head's label space but carry no observational mass.

Collapsing a multi-label turn to one move. CaSiNo tags a turn with one or more strategies: 4,615 annotated turns carry 5,858 label occurrences. The task needs one move per turn, so the mapped set is collapsed by a fixed rarest-first priority, the order the paragraph below lists the labels in. The rule is deterministic, so no sampling seed enters the target, and it deliberately does not favour the majority label, which would concentrate the next-move target on no_op even further beyond the 0.5152 floor of §2.2. One rule applies first: non-strategic maps to no_op, and where a turn is tagged both strategic and non-strategic the no_op is dropped, that annotation marking the absence of a strategy rather than a move. Nothing is lost: the full mapped set is retained on the episode record, and only the scoring target is single-valued.

Rarest first, which is also the collapse order: uv-part → undervalue_partner (131 turns), no-need → concede_no_need (196), showing-empathy → acknowledge (254), elicit-pref → ask_clarifying_question (377), other-need (409) and self-need (964) → disclose_constraint, vouch-fair → appeal_to_fairness (439), promote-coordination → return_to_shared_objective (579), small-talk → rapport_small_talk (1,054), non-strategic → no_op (1,455). Counts are label occurrences and sum to 5,858, not 4,615.

Because disclose_constraint pools two labels it takes the rarer one's rank, and no_op takes the last rank, 14, behind the five library moves at ranks 9–13 that this corpus never expresses and that carry 0 observational mass. Those five stay in the next-move head's answer space, which is why that head's chance floor is set by no_op and not by 1/14.

C Pilot protocol

Not collected. What follows specifies the randomized counterpart of §2.3 and is released as a protocol; no session has been recorded and no number in this paper comes from it. It is stated here because a fully specified uncollected design can be criticised and can be run by someone else, and because §6 and the Ethics Statement both scope their claims against it.

Scenarios, arms and session structure. Five scenarios are authored, one per row of Table 4, each with a written setup, private per-role goal cards, a shared prefix both participants read before recording, a single decision point with its cue, and exactly four branches drawn from the recommendable part of the library. The prefix is what makes the arms comparable: every session in a scenario reaches the same interactional position before the assigned move is enacted. A session runs as spoken consent on camera, a sync clap, a baseline segment, the scenario segment — the intervener reaches the decision point and enacts that session's arm while the counterpart, who has not seen the card, continues naturally for at least sixty seconds — and a wind-down, with a stop signal available to either participant throughout. A session is logged and excluded if the arm was not enacted at the cue, if either participant stopped, or if the recording fails the sync check; exclusions are recorded with their reason rather than dropped silently, so the realized assignment can be compared against the intended one.

Assignment. Arms are assigned by permuted-block randomization within strata (domain × culture), exactly balanced within every completed block rather than in expectation — the property that protects a feasibility-scale cell from drifting. The allocation is reproducible from the unit list and seed alone, and a keyed-hash path re-derives any one assignment from $(\text{seed}, \text{dyad}, \text{scenario})$ without the enrolment ledger, so an auditor can verify a single session in isolation. The manipulation check is whether the assigned move was in fact enacted at the cue, coded from the recording.

Instruments. The outcome vector is $Y = [\,\text{engagement, trust, tension, goal progress, next intent, relationship perception}\,]$. Each entry is measured in one of two rater frames, kept distinct: self-report, primary for trust and relationship perception, which have no reliable behavioural signature in a single episode; and observer-report, the partner's or a trained coder's rating of the recording, primary for engagement and tension, which are behaviourally expressed and are what a model watching the interaction could in principle infer. No published instrument is cited, because none is used: the scales are defined operationally in the released code, each stating its construct, response format, anchors, range and valence direction, and inventing a citation to a validated inventory would be worse than describing

the item honestly. The continuous dial supplies the tension channel over the thirty seconds after the decision point; the post-segment form supplies the other five. Agreement is reported as Krippendorff's α rather than Cohen's κ, because raters vary per unit and ratings would be missing at different rates across sites.

Scenario	Domain	The four arms
Founder–investor term sheet	negotiation	acknowledge, ask clarifying question, disclose constraint, return to shared objective
Compensation negotiation	workplace	acknowledge, offer two options, delay decision, disclose constraint
Weekend-plans conflict	planning	pause, acknowledge, reflect/summarize, ask clarifying question
Recurring chore disagreement	disagreement	acknowledge, reflect/summarize, offer two options, no-op
Investor update after missed targets	workplace	disclose constraint, reflect/summarize, ask clarifying question, return to shared objective

Table 4: The five pilot scenarios and their arms, from the released scenario file. Every arm is drawn from the thirteen recommendable moves of §B; no_op appears as an arm because doing nothing is a move a copilot must be able to be compared against. No session has been collected.

D Capability ontology: full label mappings

A head is every item from one source that shares a question form and an answer space, and it is the unit the audit and the quarantine act on. The capability tag is a property of the head, not of the item, and is fixed by where the answer's evidence lies relative to the observation cut. The mapping in full for the All track is below. Three things in it are meant to be checkable. One source does not equal one capability: CaSiNo contributes to Intent through its strategy head and to State through its satisfaction head, OpenToM to both ToM and Dynamics. ToM is 73% of All by item count (13,989 of 19,107), almost all of it one source — the composition imbalance §2.4 and Table 1 report and Core exists to correct. And the four heads barred by the quarantine rule are named in place rather than deleted from the accounting.

Labels that were not mapped. Two admitted sources reach this table with heads carrying no capability the ontology can equate. SOTOPIA's reward vector is a GPT-4 judgement over LLM-agent episodes; it is tagged Interactive and is never scored in v1, because agreement with an LLM evaluator is not the construct. Social-IQ 2.0's questions are tagged ToM as a whole rather than per question: a census of all 8,817 found zero future-prediction items, so none is equated to Dynamics. FANToM ships four retrospective question types and no next-action, next-state or outcome field, so its Dynamics yield is zero by construction rather than by our choice. In each case the label exists, and we decline the mapping rather than stretch it.

The All track's 19,107 items divide into 17 heads. Intent: casino/strategy (467, the paired static half). State: casino/satisfaction (200, the paired dynamics half) and meld/- (2,601, a scripted source). Dynamics: casino_acsf/outcome (419), casino_acsf/next_action (385) and opentom/- (596). ToM: opentom/multihop_fo and opentom/multihop_so (3,576 each), opentom/location_fo_fine (1,972, quarantined), opentom/location_fo_coarse and opentom/location_so_coarse (1,192 each), opentom/location_so_fine (873, quarantined), opentom/attitude (596), social_iq_2/- (793, no per-question tag), fantom/tom:belief:inaccessible (193, quarantined) and fantom/tom:belief:accessible (26, quarantined). Interactive: sotopia/- (450), never scored, its gold being an LLM judge's reward. The counts sum to 19,107.

State is this paper's name for the affect family of the released ontology. The four quarantined heads stay in All with a flag on every item and are barred from Core and Gem.

E Source datasets: licences, versions and manifests

The canonical sources are github.com/kushalchawla/CaSiNo (CaSiNo), affective-meld.github.io (MELD), github.com/seacowx/OpenToM (OpenToM), hyunw.kim/fantom/

(FANToM), github.com/abwilf/Social-IQ-2.0-Challenge (Social-IQ 2.0) and huggingface.co/datasets/cmu-lti/sotopia (SOTOPIA), all retrieved 2026-08-31. The list below the licence paragraphs records, per source, the file the adapter loads and that file's sha256 recomputed at build time. Every row records matches_declared = true, so a reproducer who obtains a different digest knows immediately that they hold a different release rather than discovering it as an unexplained score difference.

Discrepancies against the sources' own reported counts. Three are on the record and none is silent. MELD's released split sizes reproduce exactly (8,951 train, 995 dev, 2,330 test adjacent-utterance pairs), but 88 of its 1,432 dialogues are a single utterance and cannot form an $(i, i{+}1)$ pair, so the independent-dialogue count is 1,344 and not 1,432. CaSiNo's strategy annotations cover 396 of its 1,030 dialogues, the binding constraint on the paired yield of §2.2: outcomes are present on 2,060 of 2,060 participant records, so the annotation and not the outcome limits it. OpenToM ships 596 narratives in the plain configuration used here (the Long configuration, 2,300 questions over 100 narratives, is not used), of which 54 are dropped for having no locatable pre-move cut. Two sources carry no media in this build: MELD's test audio and video archives are absent from the cache, so every MELD item is served as transcript, and Social-IQ 2.0's QA release ships no transcript, so every Social-IQ 2.0 item is served as the question and its four options alone. That is the modality caveat §6 carries, and restoring the archives would not change it: no adapter passes media bytes to a model — every Episode.media value is a manifest key the prompt builder renders as a string — so the two audiovisual sources are text-input tasks by construction (Table 1, footnote ‖).

Licence consequences that bind the release. OpenToM is CC BY-NC 4.0 and its README additionally forbids training or fine-tuning on the data: it is evaluation-only, and the harness refuses to emit a training split for it in code, not merely by convention. SOTOPIA's data licence is CC BY-SA 4.0, whose share-alike obligation propagates to derivatives. MELD's annotations are released by its authors but the underlying audiovisual material is copyrighted television footage in which we hold no rights and do not redistribute. Social-IQ 2.0's MIT licence covers the QA annotations only and conveys no rights in the linked video. This is why the deposit carries item text with the source licence written into each record, and why the audio and video behind MELD and Social-IQ 2.0 are not redistributed at all.

Two licences the paragraph above does not restate: CaSiNo is CC BY 4.0 by its repository LICENSE, read 2026-08-31; FANToM is MIT, its data archive carrying no separate licence file and its README restricting the samples to evaluation use. Loaded file and the first twelve hex characters of its sha256 recomputed at build time: casino.json 4f2c4560a007, test_sent_emo.csv 8d37103938f7, opentom.json 11de58a9950e, fantom_v1.json 6a898e95df9f, qa_test.json afa3c25ad898, sotopia_episodes_v1.jsonl 100a4ab47e05. The full digests, the remaining files of each source, their byte sizes, the retrieval dates, the split sizes and the row counts are in the released manifest, whose own sha256 is 139722f4.

F Prompt templates and decoding configuration

One template per interface — not per model — is frozen before any full-scale run, and the frozen protocol file records a sha256 of each; the harness recomputes those digests from the code at start-up and refuses to run on a mismatch, so a template edited after the freeze cannot silently produce results labelled with the old protocol id. The three are reproduced verbatim below, at protocol eval_protocol_v1, content hash 99477804, prompt version swb-prompts-v1.

Infer (swb-infer-v1, template sha256 de442237).

```
You are evaluating a social situation. Infer the CURRENT social state from the
evidence available; do not speculate beyond what the material supports.

Episode: {episode_id}
Capability: {capability}
Modalities available: {modalities}
Participants: {agents}
{media}
Context:
{history}
```

```
Question: {query}
{target_hint}
{response_contract}
```

Forecast (swb-forecast-v1, template sha256 fa633244).

```
You are evaluating a social world model. Predict what happens NEXT, and, when
a candidate action is given, what happens under do(action) -- the counterfactual
outcome of taking that action rather than the observed one.

Episode: {episode_id}
Capability: {capability}
Modalities available: {modalities}
Participants: {agents}
{media}
Context so far:
{history}

Prediction horizon: {horizon}
Candidate action: {candidate_action}
Interventions to compare:
{interventions}

Question: {query}
{target_hint}
{response_contract}
```

Act (swb-act-v1, template sha256 21153243).

```
You are acting inside a live social environment. Choose your next action to
pursue your goal while remaining socially appropriate, and adapt to what this
partner has actually shown you so far.

Episode: {episode_id}
Capability: {capability}
Modalities available: {modalities}
Participants: {agents}
{media}
Interaction so far:
{history}

Goal / instruction: {query}
Candidate action under consideration: {candidate_action}
{target_hint}
{response_contract}
```

The output-format specification (response_contract, sha256 dfef61e7). The same block closes all three templates. Its keys mirror the harness's Prediction record exactly, which is what makes the parser model-independent. One line — the modality enumeration — is wrapped here for the page width and is a single line in the file; nothing else is altered.

```
Respond with ONE JSON object and nothing else. Use exactly these keys:

{
  "value":        <your answer in the task's native form, or null>,
  "distribution": <{"label": probability, ...} summing to 1, or null>,
  "outcomes":      <{"outcome_name": numeric_value, ...}, or null>,
  "ranking":       <["best_action_id", ..., "worst_action_id"], or null>,
  "evidence":      [{"modality": "video|audio|transcript|text|image|
                                  gaze|mocap|env_state",
                    "start": <seconds or null>, "end": <seconds or null>,
                    "turn_id": <turn id or null>, "text": <quoted span or null>}],
  "rationale":     "<short chain of reasoning>",
  "confidence":    <0.0-1.0>
}

Rules:
- "evidence" must contain ONLY observable spans you are pointing at. Never put
  your answer, your inference, or your reasoning inside an evidence span; the
  answer belongs in "value"/"distribution"/"outcomes"/"ranking" and the
  reasoning belongs in "rationale".
  Evidence is scored separately from the answer.
- Cite at least one span per modality you actually relied on.
- Omit no key; use null when a field does not apply.
```

The target hint (target_hint, fingerprint ef20a53c). The only part of a prompt that varies with the item is one appended line naming the answer shape the source's own target expects, so the model answers in the dataset's native space and never in multiple choice we invented. It is generated, not authored per item: categorical appends Answer format: put exactly one choice in "value".; multi-label asks for a JSON list; distributional for "distribution" with a probability per label summing to 1; continuous for a single number, with the target's unit inlined where declared; ranking for "ranking", best first; pairwise for the sign of the effect in $\{-1, 0, 1\}$; action for the action taken; and span states that the cited evidence is the answer. Where the target declares a closed choice set the choices are listed on a preceding line; where it declares outcome variables they are named for the "outcomes" field.

The observation cut, as rendered. The cut is enforced by construction, not by instruction: {history} is the episode text up to t_{obs} and nothing after it ever enters the prompt, so a model cannot ignore the cut the way it could ignore a sentence telling it not to look ahead. A Forecast prompt additionally states the prediction horizon on its own line. An episode with no prior context renders as (no prior context provided) rather than as an empty field, which is what makes the null-history control a legible manipulation of the same template rather than a differently shaped prompt.

Decoding configuration, frozen and identical across models. max_tokens 16000; up to 3 attempts per item with exponential backoff (base delay 1.0 s, factor 2.0, capped at 30.0 s); no system prompt; one sample per item; parser swb-parser-v1, fingerprint 06069fcc. Temperature, top-p and top-k are not set, and this is a limitation rather than a choice: every model in the sweep rejects those parameters with an HTTP 400, verified rather than assumed, so temperature=0 is not available as a determinism lever and a configuration naming it would be unrunnable rather than reproducible; the harness rejects such a configuration up front. Reasoning effort is unset for the same class of reason — one model in the set rejects it, so sending it would make decoding per-model. Determinism therefore rests on four recorded things rather than a sampler seed: the content-hashed templates and parser; a fixed seed for episode sampling; the provider-reported model version string captured per run, because a hosted alias moves without notice; and the complete published per-item outputs, the only artifact that makes a non-deterministic sampler auditable after the fact. Changing any of the above is a new protocol version, not a re-run.

G Scoring rules

Per target type. A categorical target is scored by exact match against the source's own label set, with macro-F1 alongside accuracy where the source's own metric is macro-F1. A multi-label target — the paired static half is one — is scored by Jaccard overlap between predicted and gold label sets. A continuous target is scored by $1 - |e|/\text{span}$ on the source's own scale, and additionally by mean absolute error and by the concordance correlation coefficient where the source reports one. A distributional target is scored by a proper scoring rule, below. A ranking target is scored by pairwise ranking accuracy over every unordered pair present in both orderings, pooled across items. Evidence is scored on its own axis and never mixed into answer correctness: the grounding module never reads the answer field, because a correct answer justified by nonexistent evidence must not share a diagnostic profile with a grounded one.

The proper scoring rule, exactly. For a distributional target over the item's candidate set L, with predicted probabilities p_l and gold g_l,

$$\text{BS} \;=\; \sum_{l \in L} (p_l - g_l)^2 \;\in\; [0, 2], \qquad \text{LL} \;=\; -\sum_{l} g_l^{*} \ln \text{clip}(p_l,\, 10^{-9},\, 1 - 10^{-9}),$$

with $g^{*} = g / \sum_j g_j$, so a soft annotator distribution is handled natively rather than collapsed to its mode. This is the textbook sum form, range $[0, 2]$, which is what every number in this paper reports; the per-label mean form $\frac{1}{|L|} \sum_l (p_l - g_l)^2 \in [0, 1]$ exists in the code for internal reporting only, and a reproducer who computes it will be low by a factor of $|L|$ — on the 14-move next-action head, by fourteen. Labels the model omitted count as $p_l = 0$.

L is the item's fixed candidate set, never the observed union of the predicted and gold labels. An earlier implementation divided by $|\text{set}(p) \cup \text{set}(g)|$, so the divisor depended on how many labels the model emitted: two models with identical belief ($p = 0.6$) scored 0.16 and 0.043 according to whether they named two labels or five, a 3.7× advantage for padding and two scores on the same item that were not comparable. Under a fixed L, padding with zero-probability labels cannot change a score, while spreading residual mass across wrong labels scores better than dumping it on one. Clipping keeps a confidently wrong prediction finite, so one item cannot define a dataset's mean.

Chance floors. A floor is measured per head, never assumed to be $1/k$. It is the best score reached by the strongest input-independent policy on that head's own item set: in closed form where the label space permits, by simulation otherwise, and in practice almost always the majority-label policy rather than the uniform one. Both are reported, because on a skewed head they differ by a great deal: the paired static half's floor is 0.5152 from always answering non-strategic, against 0.1056 for a per-label coin flip, and its dynamics half's floor is 0.8087 from always answering the median of a five-point scale, against 0.5979 for uniform sampling. The majority policy is fitted on the evaluation labels themselves, the heads it is measured on shipping no training partition, so it is an upper bound on what a label-prior predictor could achieve and is read as such wherever it appears — it is not a chance level.

Normalization and the aggregation weights. $N_d = 100\,(m_{\text{model}} - m_{\text{floor}})/(m_{\text{human}} - m_{\text{floor}})$, with no clipping in either direction: $N_d > 100$ and $N_d < 0$ both survive into the table, and an inverted, lower-is-better metric works unchanged by passing the floor above the human value. A capability score is the unweighted mean of the N_d of its datasets — the denominator is the number of datasets, not of items — so that a large corpus cannot decide the benchmark while a small, load-bearing one moves it by less than a rounding error; item counts are carried on every score for reporting and interval width, never as a weight. A dataset with no computable N_d is excluded from the mean and still reported on its native scale with its reason recorded; a capability all of whose datasets are unnormalizable is absent, not zero, because zero would claim a failure nobody measured. In v1 that is every capability, so no aggregate is reported, the weights above act on nothing, and the scalar summaries the code derives from the capability vector — macro mean, geometric mean, minimum, P_{10} — are not reported either. The old behaviour, in which an unnormalizable dataset entered its capability mean at its raw native value, remains reachable behind an explicit flag so that the 2026-09-01 decision to reverse it is auditable, and every stored report records which policy produced its numbers.

H Shortcut and contamination audit

The six filters, in order, with what each removed. Filters run in a fixed order and the first to catch an item owns it, so no item is counted twice and the columns of Table 5 sum. The ladder removed 2,267 of 21,374 candidate items, 10.6%, leaving 19,107. Two removal mechanisms are in play and they are not interchangeable. Three filters act item-wise (null_history, source_prediction, near_duplicate): the artifact is a property of the single item, so the item goes. Three act cue-balanced (answer_only, shuffled_history, name_lexical_cue): the artifact is a property of the head, and deleting every item the artifact answers correctly would invert it rather than remove it, leaving exactly the items where the shortcut points at a distractor. Cue-balanced removal instead deletes correct-side items until the trivial policy is back at the head's floor, which makes the shortcut uninformative. It is the expensive option and its cost is reported rather than absorbed.

When the head goes instead of the item. Where the pre-filter margin over the floor exceeds the quarantine threshold of 0.25 the head is quarantined rather than filtered: it stays in All with a flag on every item and is barred from Core and Gem. Four heads are quarantined in this build, for 3,064 items in total: two OpenToM fine-grained location heads (location_fo_fine, 1,972 items; location_so_fine, 873 items), quarantined manually, and two FANToM belief heads (tom:belief:accessible, 26 items; tom:belief:inaccessible, 193

items), quarantined on the threshold. The filter threshold itself is 0.05 over the head's floor, and a margin above 0.15 is additionally marked loud in the manifest.

What the contamination audit establishes — and what it does not. Two things were run and one was not, and conflating them would be the whole error. Run: a canary GUID was minted for each of the 19,107 shipped items, and 52 prediction files across five run roots were scanned for leaks; none was found. This bounds nothing about the models already evaluated, because no canary was embedded in the prompts of those runs — the zero establishes the detector and the GUID set for the next sweep, and the manifest says so in the same field as the number. Run: a seeded, sampled cross-source exact-string audit, 300 needles and 150 sampled episodes per ordered source pair; hit rate 0.0 everywhere. Absence of a hit bounds nothing about corpora this build does not hold. Not run: the per-model, per-source probe battery — canary recall, paraphrase recognition, source recognition — has not been executed against any evaluated model, so this paper reports no per-model contamination estimate and none should be inferred from the two audits above. The exposure ladder, the per-source provenance dates and the sensitivity of the paired estimate to exclusion are in Appendix S. Exposure is measured, and it is total: every item here was public before every model we evaluated. The model-level probes were never scored, so contamination in those models is unmeasured, and this benchmark is never contamination-free.

#	Filter	Examined	Removed	What it catches
1	answer_only	21,374	985	answerable from question and options alone
2	null_history	20,389	1	the question already carries its own answer
3	shuffled_history	20,388	112	indifferent to the order of the history
4	name_lexical_cue	20,276	1,090	a mined option token or a speaker name carries the answer
5	source_prediction	19,186	0	the source corpus is identifiable from the item
6	near_duplicate	19,186	79	the same prompt appears elsewhere ($\geq$ 0.90 shingle Jaccard)
	total	21,374	2,267	19,107 items kept

Table 5: The filter ladder on the frozen build. Order is fixed and the first filter to catch an item owns it, so no item is double-counted and Examined falls by the previous row's removals. Filters 1, 3 and 4 are cue-balanced. A zero is a measurement, not an omission — source_prediction alone found nothing above threshold on this build.

I Human baseline protocol

This appendix is the instrument: the sample, the interface, the instructions, the agreement statistic and the compensation. It was executed by two human raters, whose returned sheets and scoring run are reported in Appendix T.4; no model rating enters it at any point. What the resulting reference is and is not — a rating of a superseded packet, only 16 of whose 140 items carry over to the shipped build unchanged — is stated in §6 and audited in Appendix S.15, and no claim here rests on it. The replacement packet, rebuilt from the shipped build with every item checked for membership and quarantine, uses this same instrument and is unrated.

Sample and stratification. 140 items are drawn from Gem and Dynamics, stratified so that each capability and each contributing source carries a usable cell rather than a token one, and assigned across two raters at two hours each — 240 person-minutes, which sets the item count. Sixty items are double-rated and are the only basis on which agreement can be computed. The design resolves to 110 assigned rows per rater, 220 ratings in all, of which 217 were returned and 197 fall on the 139 items that scored (one item and three ratings are blank; none was unparsable). A further ten calibration items are excluded from the headline, because a calibration block measures whether the instructions are clear and folding it in would report instruction ambiguity as item difficulty. The achieved 95% interval is ± 7.9 points and is quoted wherever the reference is: at this n the reference is a coarse instrument by construction, and a 160-item target was rejected as not fitting the available rater time.

Interface and instructions. Each rater receives two matched files: a rendered document with one section per item — stimulus, question, options, and a per-item "How to answer" note keyed to the target type — and a CSV with one row per item in the same order, matched by item id and position. The rater fills three cells per row: the answer in the item's native form, a confidence rating on a five-point scale, and a flag for items judged ambiguous or mis-golded. Items are worked in order and raters are instructed not to revise an earlier answer after seeing later items, so the rating of an item does not depend on the sample that follows it. The walkthrough shipped to the rater is generated from the packet with its counts substituted in, so the instructions cannot drift from the packet they accompany.

Agreement. Agreement is computed on the double-rated items only — 58 of the 60 assigned as such returned both ratings — and is reported with that n attached. With two raters and nominal categories the statistic is Cohen's κ; ordinal and continuous item types use Krippendorff's α with the matching difference function, which is also the default should the rater set exceed two or ratings go missing at different rates. Both are reported pooled and per stratum, and the items the two raters disagree on are listed individually rather than summarised away: on a construct like this, disagreement is data. A separate output of the same pass is the mis-gold candidate list — items either rater flagged as ambiguous or wrongly golded — which is an audit of the benchmark, not of the raters, and is released with the reference.

Compensation and what the reference would and would not be. The raters are two of the authors, working unpaid; no external rater was recruited and no compensation was paid. That is a limit on the sample's independence and it is recorded rather than omitted — an author rating their own benchmark knows what the item is for. The reference is not human performance (two expert raters are not a population), is not a ceiling, and is not measured on the released artifact: the rated packet was frozen against the v1 adapters, and of its 140 items only 61 are served by the shipped build, a further 6 sit in quarantined heads, and 16 carry over unchanged. A rating of a stimulus the models never see is not a comparison, which is why the number is used as an agreement and mis-gold study and not as a human column.

J Full per-dataset results

The per-model, per-source grid is Table 10, read in Appendix T.2: 5 models against each of 3 Core datasets on that dataset's own native metric and against its own floor, on the 1,401-item balanced Core sample (467 items per capability, drawn from the 7,500-item frozen Core track of build 139722f4), scored on each dataset's own native metric against that dataset's own floor and never pooled. There is no per-capability column and no pooled column anywhere in it. A capability score is N_d, which needs one human anchor per dataset; none of the shipped datasets carries one, so the grid stops at the dataset (§2.4, Appendix T). Nothing in §5 is an aggregate of this table, because §5 forms no aggregate.

K Qualitative cases

No qualitative case is shown, and the reason is a result rather than an omission. The selection rule — pairs whose static half a model read correctly and whose forecast it then missed — was applied to the paired census and returns nothing: all five models answer the static half exactly right on 7 of the 396 pairs, and on every one of those 7 the weakest forecast any model produces is 0.75, or on a single pair 0.50, on a five-point scale whose floor is 0. The configuration this section set out to illustrate does not occur on this build, and none is constructed by relaxing the rule, which would be selection on the outcome (Appendix N). A second obstacle would bind even if a case existed: the pair records store the episode prefix and the gold target as SHA-256 checksums (prefix_sha256, target_sha256), not as text, the source dialogue being licensed corpus material held out of version control. What the paired design supports at the level of counts is in §5.2 and Table 12.

L Reproducibility

The command sequence. One released script runs the pipeline end to end — preprocess → sample → infer → score → tables — from a fresh checkout and an empty results directory, invoked as python3 -m swb.cli and python3 scripts/... from the benchmark directory, so the package need not be installed. It requires Python 3.9+ and three runtime dependencies. Corpus data is read from a cache root each adapter's manifest specifies; an adapter with no data on disk falls back to its CI fixtures and the first stage prints which ones did, because a fixture score is not a dataset score. Model inference requires an API key; with none, the identical code path runs an offline dummy model and every summary says so, so placeholder output can never be presented as a model result.

Why it is checked in two halves against two different things. The first half re-runs the whole harness twice from an empty directory and requires byte-determinism. It needs nothing but the repository, and it is deliberately not diffed against the committed tables, because it reproduces none of them: the committed artifacts are written by different generators from artifacts under runs/, so such a diff read one generator's output against another's and reported the difference as a reproduction failure. The second half re-runs exactly the two generators that write the committed tables and figure data, into a scratch directory, from the inputs the frozen manifest records, and diffs byte for byte. That half cannot run from a bare checkout, because the run artifacts it reads are licensed corpus material and model outputs held out of version control; a checkout without them exits with a distinct code rather than a pass.

The executed record. The procedure was run, and this paragraph is its result. The environment was macOS 15.7.7 on arm64 (Darwin 24.6.0) with CPython 3.9.6, pydantic 2.12.5, PyYAML 6.0.3 and NumPy 2.0.2. It ran offline with no API key, so the dummy model served every inference and half A's own scores are placeholder output, compared only against the second pass. The two passes of half A were byte-identical, so the pipeline is deterministic at a fixed seed. Half B re-ran the two generators of record into a scratch directory and compared every published artifact byte for byte: all five figure input files and all ten generated table files matched, with one file excluded by name — the figure-data provenance sidecar, whose generator still records absolute paths — and with two of the five CSVs matching as header-only files, which is a correct reproduction of a blocked figure and not evidence that a figure reproduces. The verdict was reproduced.

The checksum comparison. Recomputing manifest_sha256 from the item files returns 139722f4…4649131, the value every generated table header quotes, and the four tracks hold 19,107, 7,500, 1,600 and 500 items as declared. Of the 33 source files the manifest's source_file_checksums block covers, 31 re-hash to their declared SHA-256 with no mismatches; the other two are MELD's audio and video archives, which the manifest records as missing because that material is copyrighted television footage we do not redistribute.

What this record does not establish. The run was performed on a tree holding the complete run artifacts. It is therefore evidence that the generators, the committed artifacts and the frozen build agree with each other — not that a reader who downloads only the released archive can rebuild them. That second claim requires the archive to present the run directories under the names the generators are pinned to, and at this build it does not: the action-conditioned runs are staged into one merged directory, so the rebuild of Table 15 reports itself as not attemptable from the archive alone.

Generators of record, and the design constants that are not generated. Each generated table names its generator and the artifact it read in its own header, and the figure input files do so in a provenance sidecar; benchmark/scripts/clean_room.sh rebuilds all of them into a scratch directory and compares byte for byte against the committed ones. The exceptions are a short list of design constants (the size of the action library, the number of raters) defined once in a single macro file, each on a line carrying the date it was verified and the artifact it was read from; the manuscript checker refuses any that does not. The frozen

build every number in this paper is scored against is benchmark version 0.9.0, manifest 139722f4, built 2026-09-13 under seed 20260911 with no network and no API call. Every model run in this paper falls in the two-day range 2026-09-13 to 2026-09-14: the paired census on 2026-09-13, the balanced Core sample, the action-conditioned re-analysis and the scoring of the expert packet on 2026-09-14. There is no v1.0 build and no later sweep; a number dated outside that range is not from this release. The mirror is a snapshot of the repository with its history removed and no author-identifying material; the de-anonymised repository is released, with history, on acceptance. The mirror does not carry the source corpora or the runs/ directory.

What rebuilds from the repository alone, and what does not. The runs/ tree holds licensed corpus material and raw model outputs and is excluded from version control, so no table and no figure in this paper rebuilds from a bare checkout. What a bare checkout reproduces is the pipeline: clean_room.sh –repo-only runs the five stages twice from an empty results directory, on packaged fixtures and an offline model, and requires the two passes to be byte-identical. That establishes that the code runs end to end and is deterministic at a fixed seed; it reproduces no published number, and the script says so rather than exiting 0 as though it had.

Rebuilding the published numbers needs the artifact archive, and with it the same script regenerates and byte-compares every one of them. Table 1 and the track sizes (19,107, 500, 1,600) come from the frozen build runs/e3_benchmark_v3, recomputed from its manifest and item files rather than read off a summary; Table 12 and every \SWBpaired quantity from the five paired runs in runs/e4_paired_strategy; the figure inputs data/*.csv from those same two run sets, fig3_error_taxonomy.csv shipping header-only because the artifact behind it does not exist — empty for want of a measurement, not because the measurement came out null. One rebuild carries a caveat, and it is not Table 14. Table 14 and the partial-subset macros come from runs/sweep_v2/reaggregated_frozen_core_v3, aggregated against the same e3_benchmark_v3 the composition above is counted from: the sweep's frozen-manifest checksum in _build_tables_manifest.json (inputs.sweep_frozen_manifest_sha256) equals the build's own, and the file records sweep_frozen_build_matches_composition as true. The first frozen build, runs/e3_benchmark, enters that file only as itemset_sensitivity.prior_frozen_build, which is the prior column Table 14 exists to compare against. What those numbers are not is a benchmark result: they re-score responses bought by the off-track sweep, they cover 1,390 of the 7,500 frozen Core items, and they are not capability-balanced, which the manifest labels on every row and §6 states. Table 15 comes from one dated head artifact, recorded in the same file under inputs.acsf_heads because that directory accumulates runs and the newest wins by default; the clean-room script pins the rebuild to the recorded path, so a later run is reported as new evidence and not as a failed reproduction. This is the one rebuild that needs a manual step from the archive as staged: the packaging script merges the four action-conditioned run directories into a single artifacts/acsf/, so runs/e8_acsf_action_only/2026-09-14/ must be restored from artifacts/acsf/2026-09-14/ first. The artifact itself is present and unmodified; only the directory name is not preserved.

The remaining tables are scaffolds whose cells are still placeholders, and no artifact is withheld for them because none exists: each table's blocker, with what would unblock it, is recorded in tables/generated/_build_tables_manifest.json. Subject to the one directory-naming caveat, the archive is complete with respect to every number the paper prints. It is a single tar.gz with a per-file SHA-256 manifest (SHA256SUMS.txt), released alongside the code under the same licence (MIT).

What is in the archive, and what is deliberately not. benchmark/scripts/freeze_bundle.py stages the archive from a declared component list, and nothing outside that list can enter it. In roughly 75 MB it holds the evaluation code, the frozen build e3_benchmark_v3 and the paired build e1_paired_strategy, the prompts and the protocol file, the raw model outputs behind every published number, and the manuscript's own generated tables, figure inputs and structural checker, so the manuscript gate runs without a LaTeX toolchain. The

component list is the script's, and the archive's per-file SHA256SUMS.txt enumerates what it produced.

Three things are withheld, each for a stated reason. The corpora do not ship. The deposit holds each item's text with the licence of its source written into the record, beside prompt_sha256 and target_sha256; the corpora are fetched by the adapters from the original provider. The raw outputs are redacted in exactly two fields: a model's answer quotes the source span it relied on, so predictions.jsonl carries verbatim corpus text in evidence[].text, repeated inside raw. Each evidence span is replaced by its SHA-256 and character count — so a holder of the licensed corpus can still verify that the model cited the span the record claims — and raw is dropped. Every other field survives untouched and every score in this paper recomputes from what remains. The spans are corpus text, and the deposit already carries that text under its source licence; the outputs point at it by hash instead of carrying a second copy. The expert packet's stimulus prose is withheld. It is corpus text laid out for a person to read, and most of it predates the MELD fix, so it no longer matches the build. The answer sheets, composition and gold labels ship.

Released tracks, licences, and the observational track. We ship items, not corpora: the deposit holds each item's text under the licence of its source. Each adapter fetches material from the original provider under that provider's terms and verifies it against a checksum manifest; datasets whose licence we could not confirm from a primary source are excluded from the executable suite rather than included with an assumed licence. Licence status per dataset is in Appendix E.

All four released tracks — All, Core, Dynamics and Gem — ship on identical terms, and none of them ships a label in plaintext. Each released item record carries a prompt_sha256 and a target_sha256 and no prompt or target text, for the licensing reason above; a user regenerates both through the adapter from the source corpus and verifies them against those hashes. No track's labels are held back behind a submission server and there is no private evaluation split in v1: the holdback policy is implemented and documented, but what it withholds is the ground truth of the Live and Causal overlays, and this release ships neither.

The observational track is rebuilt from the source corpus by one adapter (casino_acsf) whose manifest pins the SHA-256 of each released file; the action library, the strategy-to-action map and the multi-label collapse rule are code, not documentation. The learned forecasters, the controls and the bootstrap are run by a single script from a seed, and they are fit on CaSiNo's own released train split, the only training partition any source in the suite both ships and permits us to use (Ethics Statement). The randomized pilot of Appendix C has not been collected, so nothing is released from it but its specification: the arm-assignment procedure is deterministic given a seed, a dyad id and a scenario id, and that procedure and its seed of record ship as code, against no session data. The table that would have reported it is cut from this version, and the table generator writes no body for a cut table: the release contains no file that stands in for a result the paper has withdrawn.

Known limits on exact reproduction. Hosted models change without notice, so scores are reproducible only against the recorded model version strings. SOTOPIA's interactive episodes would carry partner and judge stochasticity, but this package ships no judge and reports no SOTOPIA score (Table 2), so no result here depends on that source of variance.

M Use of large language models: the full disclosure

This appendix is the long form of the required AI-use statement. The statement itself carries the attestation, the scoped model-rater disclosure and the summary of use; what follows is the same account category by category.

We used generative AI tools extensively in this work. Because this is a benchmark paper, we answer first the question a reader of a benchmark should ask, and in three parts, because a single sentence about it would be false.

Benchmark construction, and three kinds of provenance. A benchmark can acquire model-authored content in three ways, and they need separate answers. Local construction. Our build turns source records into items by deterministic transformation: it calls no generative model to write an item, an option, a distractor or a gold label, and none to judge which items to keep. Answer spaces are fixed label vocabularies taken from the source — the seven MELD emotion labels, the CaSiNo strategy labels, the action inventory — or the source's own answer/distractor pair, permuted under a recorded seed. This is a claim about our pipeline, not about the origin of the material it consumes: the templates, option strings and gold labels we inherit were authored upstream, and deterministic execution on our side says nothing about how they were authored there. Inherited provenance. Two sources ship material their own authors produced with models, and we redistribute it; that is described below. Indirect model-assisted selection and repair. No model chose or rewrote an item, but model output did reach curation by a longer route: a two-sheet model-rater diagnostic — distinct from the human expert rating, and described two paragraphs below — worked a sample of the packet and flagged defective stimuli, and we acted on those flags. The items were then removed or repaired by us, not by a model, but the decision to look was prompted by a model reading. We record this as influence rather than authorship, and we do not claim the curation is model-free.

What the build actually checks. The build that freezes the benchmark makes no network call and no API call, and asserts at run time that the anthropic module is absent from sys.modules; the frozen manifest records the no-network and no-API properties of the run. That assertion is provider-specific: it establishes that this project's own model client was never loaded in the freezing process, and — with the absence of any HTTP or model-client import in the adapter package and the socket-blocking adapter tests — it is evidence that the freeze itself was offline. It is not a proof that no model of any provider touched any upstream input. Item selection and deduplication are algorithmic: the near-duplicate filter is banded MinHash over word shingles, and the Gem set is selected by baseline difficulty and shortcut margin, not by asking a model which items are good.

Difficulty scores. Difficulty is one minus a mean reference score, taken from frozen baseline runs where they cover the item and from the audited trivial-policy ensemble otherwise. The baselines are random, majority-class, constant-outcome and a hashed-token ridge regressor. None is a language model and no language-model response enters the computation — but the ridge regressor and two of the trivial policies are fitted predictors, so difficulty is not a model-free quantity in the strict sense, only a generative-model-free one.

Shortcut and trivial-policy audits, and the two parameters they fit. The audits that quarantine cue-solvable heads — the token-cue mining, its split-half cross-fitting, and the copy-from-prefix policy family — call no model and perform no model inference; the copy-policy audit asserts the absence of the anthropic module at run time as the build does. They are not, however, only string matching and statistics, and the earlier draft was wrong to say so. Two of them select a parameter on gold labels: mine_cue picks the single option token whose presence or absence best predicts the answer, and mine_answer_vocabulary collects the answer strings that recur in a head. Both are fitted on one deterministic half of a head — split by source narrative, not by item, so sibling questions cannot leak a gold string across the fold — and scored on the other, and their hit sets enter the difficulty ensemble and therefore the Core and Gem membership of the items they touch. They are one-parameter rules whose only purpose is to remove shortcut-solvable items, no weights are learned from them, and nothing they produce is a benchmark label — but they are fitted on source labels and we disclose them as such.

The expert reference is human. Every rating reported as an expert reference in this paper was produced by a person. No model answered, adjudicated, imputed or was averaged into any of it. Two human raters — the authors — independently worked the expert packet, answering 109 and 108 of their 110 assigned rows; the scored reference covers 139 items and 197 ratings, and agreement over the 58 double-rated items is Krippendorff $\alpha = 0.803$ and Cohen's $\kappa = 0.802$. The same two people later re-rated the repaired Meld state head —

33 items each, under written instructions that held the flagging threshold fixed — after the adapter's stripped-context defect was found and fixed. Its provenance is complete: both sets of answer sheets, the rating instructions, the packet manifests and the scoring output are released, and every figure quoted from them is read off those artifacts. LLM assistance was used to write the packet-assembly and scoring code, as it was for the rest of the harness (below); the ratings themselves were not. What this reference is and is not — an agreement study and a mis-gold audit on a packet the re-freeze superseded, never human performance, never a ceiling, and not an anchor — is stated in §6 and Appendix T.4, and no number in this paper treats it otherwise.

A separate model-rater diagnostic, which is not that reference and contributes no reported number. Independently of the human rating, two model raters were run over the expert packet as an instrument check on item wording and response formatting, answering 110 items each — 100 scored items plus a shared ten-item calibration block. This is not the expert reference, it is not a human baseline, and no quantity it produced is reported as a result anywhere in this paper; we disclose it because it is a model rating of benchmark items. The two sheets agree on $53/60 = 88.3\%$ of every scored item appearing in both, and on $51/58 = 87.9\%$ of the 58 scored items both sheets label overlap: two items (swb-e9-0077, OpenToM, and swb-e9-0015, MELD) carry inconsistent block labels across the sheets, and both were answered identically by both raters. The counts differ only in that bookkeeping and we report both with their denominators. Agreement between two model raters is not correctness, and is not evidence that a person can answer the items: nothing here was scored against gold, two raters can share an error, and a model rating cannot supply a human anchor, the quantity being anchored being what a person achieves. The exact API model identifier, run date, sampling temperature and verbatim prompt of these two diagnostic runs were not recorded, so the diagnostic cannot be reproduced from our release; the packet, the two answer sheets and the audit note are released as they stand. That gap is a gap in the model diagnostic alone — the human rating above carries its raters, instructions, sheets and scoring run in full. The diagnostic earned its place by finding defects rather than by producing a score: both sheets independently flagged MELD stimuli unanswerable once the adapter strips their context, and a FANToM option whose belief direction is inverted relative to its stem, and the human raters later flagged the same MELD rows. Those findings are reported as benchmark defects (§6); they are the indirect model-assisted curation declared above, and no model rating is pooled with, substituted for, or averaged into any human number.

Model-derived content inherited from source corpora. Three of our sources carry material their own authors produced with models, which we disclose because the benchmark uses those sources and redistributes their derivatives. The largest of the three is OpenToM, whose own README states that its story drafts are composed using LLMs and that some stories were then revised by people (Xu et al., 2024); it supplies most of the item text in both scored tracks, and the statement of record with the two proportions and the quotation is the AI-use statement's paragraph on inherited content. We generated none of it: our adapters map each source's own text and labels into our episode schema with deterministic code. SOTOPIA's native reward is a GPT-4 judgement over LLM-agent episodes (Zhou et al., 2024); we ship no judge. SOTOPIA appears in Table 2 as unscored. It is the one inherited source publishing both a floor and a human reference on its own scale, so the capability normalization is defined for it — that column reports whether a human anchor exists, not whether a score was produced — yet no SOTOPIA score is ever produced. Its native metric here is a deterministic rubric-term score, not the GPT-4 reward. Every real SOTOPIA configuration ships without a rubric, so the adapter sets those episodes aside as judge_required and contributes no score at all rather than a confident zero; SOTOPIA has no items on Core, and no number in this paper is scored against the GPT-4 reward, which is carried only as a labelled source reference field. FANToM's conversations are model-generated dialogue authored by that benchmark's creators (Kim et al., 2023); we neither generated nor filtered them, and its gold labels are the source's own. We did not use a model to produce, adjudicate or re-label any of this material.

Implementing methods; data cleaning and reformatting. The evaluation harness, the dataset adapters, the metric implementations, the contamination probe module, the paired-episode construction and the analysis scripts were written primarily by LLM coding agents working from author-specified interfaces and acceptance tests, under author review. The adapters normalize each source corpus into the common episode schema, so reformatting source data into benchmark items falls in this category — but, as above, that reformatting is executed by deterministic code, not by a model at build time. Every module carries unit tests, and the reported numbers are produced by executing that code rather than by a model reporting them. The contamination probes are a deliberate case in point: the paraphrase probe is a thirteen-rule regular-expression rewrite, not a paraphrase model, and its items are excluded from scoring.

Designing experiments, and refining the hypothesis. The experimental design — control conditions, ablations, the choice of proper scoring rules and the significance procedure — was drafted by LLM agents from an author-written research plan and then revised by the authors. Two design decisions changed as a result of agent-raised objections rather than author intent: the primary evidence moved from a single correlation coefficient to reversals between pairs of models, after an agent observed that a correlation over five models would be unstable; and the modality ablation changed from filtering to masking, after an agent identified that filtering confounds the modality effect with item-set composition. The research question is the authors' and predates any model assistance, but its refinement into the paired within-item form it now takes was shaped by these exchanges, and the policy requires that hypothesis refinement be disclosed as well as hypothesis proposal.

Literature summarization. The related-work survey, the novelty analysis and the licence audit were conducted by LLM agents working against primary papers, proceedings pages and repository LICENSE files, each bibliography entry recording the date it was fetched. Claims that could not be verified against a primary source were removed rather than softened. We withdraw the stronger claim made in an earlier draft, that every literature statement retained in the paper had been verified against a primary source: an independent primary-source check during review found a substantive scope error in our characterization of Momentè et al. (2025), whose reported -0.17 is SocialIQa against interactive games across nine models, not an aggregate over a social-benchmark suite. The affected passages have been corrected (§1, Appendix R). One verified error in a checked set is a reason to describe the process rather than to certify its output, and that is what this paragraph does.

Writing, and interpretation of results. Drafting and editing of this manuscript used LLM assistance throughout, including the prose that interprets the reported numbers. Every computed number in the text is produced by the analysis code and traced to the artifact that generated it, and every quoted literature number is traced to its primary source; the interpretation placed on those numbers was drafted with model assistance and is the authors' responsibility. The authors are accountable for every claim, citation and number in this paper.

Assistance during this review. The revision this statement belongs to followed an AI-assisted external review of the manuscript, and the policy's disclosure categories cover that work. The review provided methodology and experiment-design feedback, refinement of the central hypothesis and of what the paired comparison can identify, checking of quoted literature against primary sources, commentary on the interpretation of reported numbers, and direct edits to the abstract, the limitations section and this statement, all of which the authors then accepted, rejected or rewrote. The same review completed both model-rater diagnostic sheets described above. The exact served model identifiers and sampling temperatures for that review are not available to us. The second diagnostic sheet was completed in a separate context from the first and from the review notes; that is a procedural separation and does not establish statistically independent errors or identical prompting, and we do not treat the two sheets as independent raters.

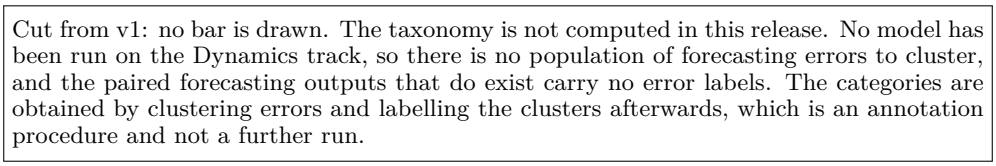


Figure 1: Distribution of Dynamics prediction errors over the failure taxonomy of §5.3, with bootstrap intervals, categories derived by clustering the errors and labelled afterwards rather than imposed in advance. Cut from v1 on 2026-09-15: the taxonomy is not computed in this release. No model has been run on the Dynamics track (zero), so no population of forecasting errors exists to cluster; the paired forecasting outputs that do exist are one head of one corpus and carry no error labels; and no clustering artifact exists anywhere under the released runs. What would fill the figure is that clustering and its after-the-fact labelling, which is an annotation procedure rather than a further sweep.

What we did not use them for. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, or to write or contribute to proofs; this paper contains none. The research question is the authors' own, though its refinement was model-assisted as stated above. No generative model wrote, rewrote, judged or selected any benchmark item, option, distractor or gold label, and no model score is reported as, or pooled with, a human score. Models are the object of study in this work. The three qualifications to that separation are stated above and nowhere else: the model-rater instrument check, which is not the expert reference, produces no number reported anywhere in this paper, and did influence curation; the model-derived material inherited from two source corpora, which we redistribute with its provenance labelled and, in SOTOPIA's case, never score against; and the two fitted, non-generative audit parameters that shape track membership.

N Extended analysis and discussion

N.1 Error taxonomy

The state-error / transition-error partition, and why it attributes nothing. Because both questions hang off one episode, every Dynamics error could be labelled by whether the model got the paired static question right: a state error where it did not, a transition error where it did and the continuation is still wrong. Neither label locates a deficient transition model. A wrong forecast can equally come from facts no question asked, continuations the cut left open, or label error. So the split describes which errors co-occur with a correct reading, rather than attributing them.

To characterize how prediction fails rather than only how often, we cluster incorrect Dynamics predictions and label the clusters into a taxonomy. The taxonomy is derived from the errors, not imposed on them: clusters are produced first, then labelled, then the labels are checked against a held-out sample of errors by a party that did not produce the clustering.

Figure 1 is where that distribution would be reported. It is cut from v1 and draws no bar; the paragraph after it says why, and the figure's own caption repeats it so that the two cannot drift apart.

The taxonomy is not computed in this release, so no category is defined here and no share is reported. Three things would have to exist and none does. No model has been run on the Dynamics track (zero of them), so there is no population of forecasting errors to cluster. The forecasting outputs that do exist are the paired casino/satisfaction halves — one head of one corpus, one capability — and they are unlabelled: each record carries a predicted value, a distribution and a rationale, and no error category. And the categories are not a quantity a further sweep would produce: they are obtained by clustering the errors and labelling the clusters afterwards, with the labels checked by a party that did not produce the clustering, which is an annotation procedure we did not run. We therefore report how often forecasting fails, throughout §5.2 and Table 12, and say nothing about how it fails.

The two paragraphs that follow set out what a taxonomy would have been used for; they are the design, not a result.

Two hypotheses about the failure mode imply different distributions over this taxonomy, and we state them in advance of the data. Under a state-error account, prediction failures accompany comprehension failures: the model misses the paired static question and the forecast is wrong too. Under a transition-error account, the model answers the static question and still forecasts wrongly. The second is compatible with a deficit in the transition and does not establish one: the static question tests one fact about the episode, so answering it leaves every untested fact open, and a forecast can also be wrong because the continuation was not determined at the cut or because the outcome label is wrong. Distinguishing those from a transition deficit needs what this design does not have — a specification of what a model would have had to know, and a measurement of the outcome's irreducible uncertainty. Neither account is tested here, and we state them in advance precisely so that the absence is legible. Choosing between them requires the distribution over error categories that does not exist, and even the weaker substitute — the model's accuracy on the static half of each pair whose forecast it missed — is computed by no script in this release. The two hypotheses are a design commitment about what the paired items make askable, not a finding, and nothing in §5 or Appendix T should be read as evidence for either.

The paired design is what makes the conditioning possible: because the static and Dynamics queries of a pair hang off the same episode, a model's prediction error can be conditioned on whether it answered the static question about that episode correctly. On the one paired head this release ships, that static question is a strategy label and not a state or belief query, so the conditioning is narrower than the design allows in principle. What it buys is a partition of the errors, not an attribution: the strongest form our claim can take is that a share of forecasting errors occurs on episodes the model demonstrably read correctly on the questions we asked, which narrows where to look and does not name a missing component. No episode is shown, because the case this design can show does not occur on this build. The rule below was applied to the paired census rather than left as an intention, and it selects nothing: all five models answer the static half exactly right on 7 of the 396 pairs, and on every one of those 7 the weakest forecast any model produces sits at 0.75 — on one pair, 0.50 — on a five-point scale. There is no pair on which the episode was read correctly and the forecast then missed, so there is no representative case to display, and none is manufactured by weakening the rule until one appears.

Cases are selected by an explicit rule stated with them — the pairs on which every model answered the static half correctly and every model missed the forecast — rather than chosen by hand for illustrative value. That rule is what returns nothing here, and reporting the empty yield is the honest outcome of having fixed it in advance. Appendix K records the further obstacle that would apply even to a pair that qualified.

What the observational–interventional contrast can and cannot say. The move added forecasting value on observational data, and that value lay in the partner's immediate reaction rather than in the final outcome — but only against the weaker of the two baselines we can form. Adding the observed move to an action-free forecaster improves the partner's next move by ΔBrier -0.0537 $[-0.0711, -0.0378]$ and the end-of-negotiation outcome by $+0.0036$ CCC $[-0.0034, +0.0115]$, which spans zero. Against a forecaster trained on the move alone the immediate-reaction gain disappears too (-0.0064 $[-0.0229, +0.0116]$), and what survives on this track is an outcome effect that belongs to the conversation rather than to the move: $+0.0987$ $[+0.0080, +0.1958]$ for the conversation over an action-only baseline, and $+0.0292$ $[+0.0055, +0.0514]$ for a history-by-action interaction the additive forecaster cannot represent. Every interval here resamples whole dialogues. Two readings are excluded regardless of those numbers. A forecaster that predicts observational outcomes well has not thereby been shown to predict effects, because the move it conditions on was chosen; and a pilot at the specified scale could not have shown that a forecaster's effect ranking is wrong in general, only whether on those scenarios it agreed with what happened. The interventional half of each is stated in the conditional because no such contrast is reported in v1: the randomized collection is specified and released but has not been run, no session exists, and nothing here licenses a causal claim. What the released protocol is intended to establish,

once collected, is that the interventional quantity is measurable at a scale two people can reach — and that the field's observational corpora, the only ones large enough to train on, answer a different question from the one a copilot asks.

What the paired correlation establishes, stated plainly. On the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap -0.2936) and no artifact in this release putting them on a common scale, the two halves move together. The model-level relationship is positive on all three coefficients — Pearson $r = 0.8431$, Spearman $\rho = 0.8000$, Kendall $\tau = 0.6000$ over $n =$ five models — and every one carries a model-resampled interval that reaches 1.000 and, on ρ and τ, falls to 0.111 and 0.000 ($[0.517, 1.000]$, $[0.111, 1.000]$, $[0.000, 1.000]$; resampling the conversations instead gives $[0.600, 0.900]$ on ρ). Five points are five points: the interval is as much the result as the coefficient is, the leave-one-model-out range on ρ is 0.6000–0.8000, and nothing here needs more precision. What it establishes is one thing and no more: on stimuli held fixed, over one head of one corpus, a model that scores higher on the static question also scores higher on the forecast. It is not evidence of a dissociation on this benchmark; it is not evidence against one on any other; it is not offered as an explanation of any published coefficient; and it does not replicate to a second corpus, because this release does not have one.

Why we do not reconcile the published estimates. Re-reading all five published estimates against their primary sources (Appendix R.10) showed that they do not relate the same pair of quantities to each other, that the one we had treated as the flagship negative result is a single cell of a matrix whose other social rows are positive, and that the one we had called an outlier is an unsigned r^2 against a loss and is substantively positive. Nothing among them is a clean report that social understanding anti-correlates with social forecasting. The reconciliation earlier drafts attempted is therefore withdrawn, together with the three competing accounts that existed only to explain why two published coefficients differed.

What survives is weaker and is the actual motivation. The five estimates differ in benchmark, construct, model pool, statistic and unit of analysis, and no two share items. A spread produced that way is evidence about measurement rather than about models: it does not tell us that the two abilities come apart, and it does not tell us that they do not. That is an absence of evidence, not a contradiction, and it is why an instrument that holds the conversation, the timestep and the annotation source fixed while varying only the question (§2.2) is worth building — on its own merits, not because anybody's numbers disagree.

What our estimate adds is correspondingly bounded. It is one model-level correlation, over five models, between two scores computed on the same conversations at the same observation cut. Holding the input material fixed removes exactly one explanation of a rank change, the corpus, and leaves the others — different constructs, different difficulty, different label reliability, different attainable score — standing. It cannot explain any published coefficient, because it does not vary what differs between those studies. Nor is its forecasting half plural in v1: the shipped Dynamics track is CaSiNo and OpenToM, and only CaSiNo supplies real human outcomes, so this is one behavioural source and one narrative source, not a replication across several. We pre-commit to reporting the value either way, and to describing it as one observation rather than as a verdict on the literature. The value itself is above and in Table 12, panel (b). We report no partial correlation controlling for a general-capability covariate, and none can be computed here: that covariate would be a normalized aggregate per model, and this release forms none on any track, every dataset it ships being unanchored (§2.4, Appendix T).

What each outcome would license. If the relationship is weak on these stimuli, then a static social score on this benchmark does not stand in for forecasting on this benchmark, and an improvement in one is not evidence of an improvement in the other. It would not show where the shortfall is: the state-error / transition-error partition above labels errors by what the same model answered about the same episode, which is a place to look and not an attribution. We do not report that split (§5.3). If the relationship is strong, the

conclusion is not that the paper failed but that on stimuli held fixed the two scores move together — one new observation, on the axis the published estimates never held fixed. It would be evidence bearing on the functional integration position of Tsvilodub et al. (2026), whose own experiments are static. Neither outcome identifies a mechanism, and neither explains anybody else's coefficient.

What neither outcome would imply. A gap between comprehension and prediction on this benchmark would not be evidence that models lack a world model in any general sense, nor that the gap is irreducible. Prediction targets are intrinsically uncertain and part of any gap is noise, and nothing in this release separates that part from the rest. What would is a reference collected on the same items the models were run on; our human expert reference is not that — it is a rating of a superseded packet with no item in common with the paired census (Appendix S.15) — so no irreducible component is estimated here and no claim in this paper is bounded by one. Nor would either outcome establish that prediction is uniformly the harder capability: pairing fixes the corpus, not the ceiling, and published per-track orderings already vary by model (Kong et al., 2026).

Relationship to interactive evaluation. Interactive social environments evaluate acting; this benchmark evaluates forecasting. Whether a model's predictive accuracy over an episode is informative about the outcome it would achieve as a participant in that episode requires the same episodes to be run in both modes, which the paired design makes possible for the sources that support live replay. We regard that as the natural next experiment and do not answer it here.

What remains unmeasured. Across the benchmarks we surveyed, none scores an interventional prediction — $P(s_{t+\Delta} \mid h_t, \mathrm{do}(a_t))$ against randomized, actually observed branches — and our own source review found no cleared corpus that could supply one. That is the sharpest remaining gap in social evaluation, and we flag it as future work rather than as a claim of this release.

Release and reproducibility. The benchmark, the paired-episode manifests, the harness, every per-item model prediction and the analysis code are released. Appendix L is where the clean-room reproduction will be recorded — the command sequence from raw source datasets to the final tables, run by a party that did not build the harness. That run has not happened; the appendix currently specifies it.

O Action-conditioned social-state forecasting: full definition

Definition. Let h_t be the observable history of a two-person interaction up to time t (transcript, and where recorded, prosody and facial behaviour), let c be the context the acting participant is entitled to know (their own goals and constraints, never the partner's private information), and let $a_t \in \mathcal{A}$ be a candidate move by that participant drawn from a fixed library. The task is to forecast the partner's social state and the interaction's outcome under that move,

$$P\big(s_{t+\Delta},\, o_{t+\Delta} \;\big|\; h_t,\, c,\, \mathrm{do}(a_t)\big), \tag{1}$$

for horizons Δ from the partner's next turn to the end of the interaction. s is a vector of continuous social-state channels (tension, engagement, trust, rapport) and o a vector of outcomes (goal progress, satisfaction, whether agreement was reached, what the partner does next). The form of (1) is not new: it is the transition model of a world model (Ha & Schmidhuber, 2018; Hafner et al., 2023) with a social state in place of a physical one, and it has been written down for social settings before — with an LLM prompted to fill the distribution (Zhou et al., 2025b), with market beliefs as the state and news events as the conditioning variable (Yu et al., 2026a), and with a learned intent-level model for a single customer in a retail setting (Zhang et al., 2026). What we add is the instantiation on real two-person recordings with continuous state, and the evaluation regime below.

Why the action must be constrained. If a_t ranges over free text, two forecasters are never conditioned on the same thing and their forecasts cannot be compared, and an unconstrained

recommender can invent moves no one should follow. We therefore fix a library of 14 verbal and self-regulation moves — pause, acknowledge the other person's concern, ask a clarifying question, reflect or summarise, disclose one's own constraint, offer two options, delay the decision, suggest a break, return to the shared objective, and a no-op — plus four moves that human negotiation corpora observe and a copilot would not recommend (small talk, an appeal to fairness, undervaluing the partner's need, conceding a need). Every move is something a person can enact on instruction; physical actions are excluded by design. Appendix B gives the library, its per-corpus label map, and the collapse rule for a multi-label turn.

Two kinds of evidence, and why the paper needs both. Equation (1) is written with $\mathrm{do}(\cdot)$ because the quantity a copilot needs is interventional: what happens if the user takes this move. Almost all available data is observational — the move was chosen by the speaker, and whatever made them choose it also shapes the outcome — so we instantiate the task on both kinds of data and keep them apart. Observational (§2.3): in a corpus of human negotiations whose turns carry strategy annotations and whose participants reported their outcomes, each annotated turn becomes an item — history, the mapped move (the utterance itself withheld), and the reported outcomes — and a forecaster is scored on how well it predicts outcomes given the move that was taken. That measures conditional prediction, not causal effect, and every table says so. Interventional (§2.3): specified, not collected. Under the released protocol two people would reach an identical decision point and one would enact a move drawn at random from the library, the other continuing naturally, both reporting their state; per-arm outcome means would estimate $E[Y \mid \mathrm{do}(a)]$ and their differences would estimate effects. No session has been recorded, rated or frozen, and this paper instantiates the task on observational data only. The contrast between the two would be the experiment — a forecaster fit on observational moves asked to rank interventions, its ranking compared with the randomized outcome — and it is the experiment the protocol exists to make runnable. It is not an experiment this paper reports, and no number below comes from one. What the paper says about the observational–interventional gap is therefore a statement about what the observational track cannot settle, never an estimate of the gap itself.

Evaluation. Continuous channels are scored with Lin's concordance correlation coefficient (Lin, 1989), the convention of the affective-computing challenges (Ringeval et al., 2019; Martinez-Lucas et al., 2026): a constant predictor scores exactly zero, and a predictor that tracks the trend at the wrong level is penalised, which is what a copilot's state estimate requires. MAE and Pearson r are reported beside it. Because a copilot acts on change, we also score the direction of change (down / flat / up) as a macro-F1. Categorical futures — the partner's next move — are scored with a proper scoring rule (Brier, log loss) and top-1 accuracy. Interventional items, were any collected, would be scored with effect-direction accuracy over $\Delta_Y(a, b) = E[Y \mid \mathrm{do}(a)] - E[Y \mid \mathrm{do}(b)]$, pairwise action-ranking agreement, and calibration, following the decision-centric evaluation ladder for world models (Yu et al., 2026b); that half of the scoring specification has never been exercised. Every forecaster is read against three controls that remove one kind of information: a constant predictor; the same model with the history removed; and the same model with the move replaced by a random other move. A forecaster whose score survives the last of these is not using the action.

O.1 Instantiating the task: observational and interventional tracks

Dynamics/observational: negotiation with annotated moves. CaSiNo (Chawla et al., 2021) is the one licence-cleared source with behavioural outcomes reported by real people after real interaction (CC BY 4.0; Appendix E). Its dialogues carry utterance-level strategy annotations on a subset, and every participant reported satisfaction, liking of the partner and points scored. We map the nine annotated strategies onto the action library (Appendix B; every label maps, none is dropped) and cut each annotated dialogue at every annotated turn $t \geq 1$. Each cut yields an outcome item — history to $t-1$, the acting participant's own priorities, the mapped move at t with its wording withheld, and both participants' end-of-negotiation outcomes as the target — and, where the partner replies, a next-move item whose target is the partner's mapped move at $t+1$ as a distribution. Items follow the

corpus's released train/validation/test split by dialogue, so no dialogue contributes to both training and evaluation. The train split loads 6,987 items from 345 annotated dialogues and the test split 835 items from 41, the split being the corpus's own and by dialogue. The 835 test items divide into 440 final-outcome items and 395 next-move items on disjoint item sets, scored by different metrics and never averaged: an outcome item exists at every annotated turn, a next-move item only where the partner in fact replies. Each of the 440 outcome items carries a six-outcome label vector, of which four Likert outcomes enter the head's score. Head A's concordance is therefore the mean of four per-outcome scores, each computed over all 440 items, so the head rests on $440 \times 4 = 1{,}760$ scored scalars and never on fewer than 440 items. The two outcomes left out of the mean are scored and reported separately (Appendix T.14). one in test — because the corpus's annotation list does not align position-for-position with its chat log, and a misaligned dialogue would attach a move to the wrong turn. They are skipped and counted rather than repaired by guessing the offset. These items sit on the Dynamics track and never on the Counterfactual track: the move was chosen, not assigned, and the harness refuses to score it as causal ground truth by construction.

Counterfactual/interventional: the randomized pilot. Not collected. No session has been recorded, rated or frozen, and no number in this paper comes from it. What follows is a released specification, written in the voice of a protocol rather than of a report, under the randomized-branching design of Appendix C: it would use five authored scenarios (a founder–investor term sheet, a compensation negotiation, a weekend-plans conflict, a recurring chore disagreement, an investor update after missed targets), each with a shared prefix, a decision point, and four arms drawn from the recommendable part of the library. One participant would reach the decision point and enact the arm assigned by a seeded permuted-block randomizer; the other would continue naturally for at least sixty seconds. Both would then rate their own state from the recording on a continuous dial and complete a post-segment form for engagement, trust, goal progress, intention for the next interaction and relationship perception, with the dial in the thirty seconds after the decision point supplying the tension channel. Sessions in which the assigned arm was not enacted, or in which a participant used the stop signal, would be logged and excluded. The protocol requires consent-based and visible collection (spoken consent on camera at the start of every session, a stop signal available to either participant, no covert capture), and its scale would be that of a feasibility study; the Ethics Statement and §6 say what that scale would and would not license. Releasing the protocol is the contribution here; running it is not claimed.

Forecasters. Two families are defined for both tracks, and only the observational track has been run. Prompted forecasters are language models given the history, the context and the candidate move under the frozen template of §4, which is the setting of the prompted social world model (Zhou et al., 2025b). Learned forecasters are small models fit on the observational training split — ridge regression over hashed history features and a one-hot move — whose only purpose is to isolate the value of the move: the same model with the move removed is the action-free control in the spirit of Zhao et al. (2026), and the same model with the history removed is the action-only control. No prompted forecaster has been run on this track as of Sep 14: every configuration in the frozen observational results is a learned or trivial one.

P Paired-episode construction: detail relocated from §2.2

The head that cleared the yield threshold, and why it was not used. The alternative paired head does clear the 1,000-pair threshold, at 1,030 pairs. It is solved 75.8% of the time by a string match against a 0.3340 floor — a +0.4243 margin, which our own quarantine rule (§3.3) disqualifies. It therefore ships as a secondary head with the margin printed beside it rather than as the paired axis. Choosing the strategy head instead moves the paired static half to Intent and separates the two halves' referents, which is why 48.2% referent coincidence ships as a covariate (§2.2).

Per-source composition, and the two gates that leave 396. Table 6 is the build in full. 13,948 raw pairs are constructible from the three admitted sources. Two gates then apply, and they are parallel criteria on that same raw pool rather than one ladder. Collapsing replays of a conversation to one independent configuration leaves 2,620. Requiring instead that the Dynamics gold be a human behavioural or self-reported outcome leaves 792, replays included. Applying both leaves the 396 pairs this paper scores. Each gate is a declared criterion, not a sample, so the drop is attributable rather than incidental.

Both gates decide the paper's evidence. The independence collapse removes 11,328 of the raw pairs, almost all MELD's: 12,276 adjacent-utterance pairs come from 1,344 dialogues, so treating them as independent would overstate the effective sample by roughly an order of magnitude. The human-gold gate then removes 2,224 of the 2,620 independent pairs, and it removes both large sources outright: MELD's gold is the emotion a screenwriter wrote next, and SOTOPIA's is a GPT-4 judgement over LLM-agent episodes, so neither is a behaviour a person produced. That is what leaves one corpus standing, and why the paired result is reported as one corpus and not as a cross-source finding. A fourth source, OpenToM, is excluded before either gate applies: its static half asks for an entity's original location, which the prefix names verbatim, so a static-versus-dynamics comparison against it would measure the copy and not the understanding.

Source	Static gold	Dynamics gold	Kind	Raw	Indep.	Human
CaSiNo	turn's multi-label strategy set	post-session satisfaction, 5-point self-report	human self-report	792	396	396
MELD	emotion of utterance i	emotion of utterance $i{+}1$	scripted television	12,276	1,344	0
SOTOPIA	social goal at the midpoint	goal reward, 0–10	GPT-4 judge	880	880	0
Total				13,948	2,620	396
admitted but yielding no pair, and why						
FANToM	all four question types are retrospective; no next-state field exists			0	0	0
Social-IQ 2.0	census of all 8,817 questions: zero future-prediction items			0	0	0
OpenToM	excluded: the static half is answerable by copying a string from the prefix			—	—	—
Declared yield threshold						1,000
Realized, against the threshold						**−604**

Table 6: Paired composition and the filter ladder, from the frozen paired manifest. Raw is every constructible pair; Indep. collapses replays of one conversation to one observation; Human additionally requires the Dynamics gold to be a behavioural or self-reported outcome from a real human interaction, which is the manifest's own gate criterion. The gate fails by 604 pairs and is recorded as failing rather than retired quietly: §3.2 says why the threshold does not bind what this paper reports, the statistic being model-level. The survey-field leak audit on the surviving head counts 0 of 792.

The paired episode as a tuple. Formally a paired episode is $(h_{\leq t_{\text{obs}}}, q_{\text{s}}, y_{\text{s}}, q_{\text{d}}, y_{\text{d}})$: one prefix from one conversation cut at one t_{obs}; a static query q_{s} tagged State, Intent or ToM whose target y_{s} is determined at or before the cut; and a Dynamics query q_{d} whose target y_{d} lies strictly after it. Both queries are issued over byte-identical context, and the construction is identical in every source.

The primary paired head, and what bounds each half. One paired task is primary in v1: casino/strategy, at 396 independent pairs (792 participant records, one pair per dialogue). Its static half asks which negotiation strategies the last speaker's turn enacts, over a ten-label multi-label set — uv-part, vouch-fair, non-strategic and seven more — and its gold is CaSiNo's own utterance-level strategy annotation, written by third-party annotators and by neither negotiator. Its dynamics half is the pair's participant's post-negotiation satisfaction, a five-point self-report collected after the session. The two golds are accordingly of different kinds, an annotator label against a participant survey response, and they concern parties that coincide in only 191 of the 396 pairs (48.2%); that share is reported as a covariate and not as a defect (§2.2).

Two measurements bound each half, both computed on the paired item set itself. The static half has a chance floor of 0.5152 — the best fixed label set, {non-strategic}, scored by Jaccard — against which the best trivial input-reading policy scores 0.0535, a margin of −0.4617: below the floor, and below even our SHORTCUT_MARGIN of 0.05. Surface copy scores 0.0000 exactly, because the labels are annotation jargon that no participant types. The dynamics half has a floor of 0.8087, always answer 4, the median, under $1-|e|/\text{span}$ on a five-point scale. The two floors differ by −0.2936 and the two metrics are not commensurable, so the difference of the two absolute scores is not interpretable and nothing in this paper subtracts them.

value2issue, and why it is secondary rather than primary. An earlier build put value2issue in the static slot: which of three issues — firewood, food, water — the named participant privately ranked highest before negotiating, the gold being that participant's own pre-negotiation survey response, never shown in the prompt or prefix and verified per item (0 of 2,060 leaks). It yields 2,060 raw and 1,030 independent pairs, so it clears the 1,000-pair feasibility precondition that strategy, at 396, fails by 604. It is nonetheless not primary, for two reasons both settled before any paid run.

First, it is largely solvable without reading anything socially. Against a 0.3340 floor, near-uniform over the three issues (688/687/685), the policy answer the first issue the participant mentions in their own turns scores 0.7583 — a margin of +0.4243 over the floor, the largest in the benchmark and above our own QUARANTINE_MARGIN of 0.25, the published rule under which we quarantined opentom/location_so_fine at +0.2039 and fantom/tom:belief:accessible at +0.3967. Negotiators say their priorities out loud. Applying our own threshold to our own headline would disqualify it, and a static-minus-dynamics comparison computed on it would partly measure copiability. Second, on value2issue both halves are the same participant's own survey responses from a single session, so the pair may measure within-person consistency as well as social inference. That caveat is a property of this retired head and is not true of strategy, whose static gold is an annotator's label about the last speaker rather than a survey response from the participant whose satisfaction forms the other half. Any statement that both paired labels are one participant's survey responses is a statement about value2issue alone.

The 1,000-pair precondition is therefore retired as superseded by the shortcut audit, and the retirement is recorded rather than dropped: the manifest still stores the gate as failing at a margin of −604. It does not bind because the headline statistic is model-level with $n = \text{five}$ — what is paired is the stimulus, what is correlated is a model score — so the pair count sets the precision of each model's coordinate and not the degrees of freedom of the correlation (§5.2). Passing a yield threshold on a head that a string match solves three times in four is not a qualification. One script builds both heads and reports both yields; every pair records which head it came from, and no yield, floor or accuracy from the manifest can be quoted without it. value2issue is reported only as a secondary head, always with its +0.4243 margin quoted in the same breath.

Provenance is recorded per side, never per pair. Each side of a pair separately records its source dataset, the annotation field the target was read from, and the kind of ground truth behind it: a human-reported outcome, a human-annotated label, a scripted continuation, a value derived from metadata, or a model judgement. Because these kinds are not interchangeable, the manifest never collapses them, and §2.4 reports per source and per kind rather than pooling into a single number.

What pairing buys beyond corpus control. Pairing puts the model-level comparison on a common item support, so a rank change between the two axes cannot be an artefact of changing corpus. It also makes a within-model conditional analysis available: a Dynamics error can be conditioned on the same model's static answer for the same prefix, which separates a comprehension failure from a transition-model failure. What pairing does not buy is a common ceiling, which is why the difference between the two absolute scores is not interpretable and the paper's claims are relative orderings, reversals in the ranking of one

model against another — not of one paired episode against another — and distance from the expert reference on the same items.

Pairability as a falsifiable precondition. Beyond licence, capability-preserving annotations and preserved native modality, a source is admitted on pairability, stated in advance as a falsifiable precondition: below the target paired yield, the within-item comparison this paper reports cannot be built from existing data at all, and the honest response is to say so rather than to report a weaker comparison under the same name.

Admission to Gem, and what v1 Gem is not. Gem is selected by measured difficulty against frozen baseline runs, not by human verification. An item is admitted only if (i) it survived all six shortcut filters and its head is not quarantined; (ii) it is measured hard — every frozen baseline run that covers the item scored zero on it, or, where no baseline run covers it, its difficulty is in the top decile of its own head by rank; (iii) its head carries no residual shortcut, its post-filter best trivial policy sitting within 0.05 of that head's floor; and (iv) the admitted pool is balanced to the Gem budget. Difficulty is $1 -$ (mean reference score on the item), the reference being the frozen baseline runs where they cover the item and the audited trivial-policy ensemble elsewhere; every item records which of the two it received. The cut is taken on rank rather than on a quantile value because difficulty is discrete: on a two-choice head a threshold at the 0.90 quantile admits every item tied at it, and an absolute cut at 1.0 would bar every binary head altogether, the always-first and always-last policies both being in the ensemble and unable to be both wrong. 2,588 items meet the admission criteria and 500 are admitted after balancing. The two difficulty routes contribute 1,368 and 1,220 candidates: the first is the measured route, every frozen baseline run covering the item scored zero on it; the second is the rank route, used where no baseline run covers the item at all. The frozen baseline runs cover two sources and 3,405 items, so the rank route carries the majority of the corpus and its difficulties are ordinal rather than calibrated — an ensemble mean over trivial policies of unequal strength. Both of OpenToM's fine-grained location heads are excluded here for residual shortcut, over and above the quarantine that already bars them.

Two things this leaves on the record. First, v1 Gem is not human-verified: the criterion we specified additionally requires a human other than the item's proposer to confirm the gold target, and the expert packet was rated — two human raters, 139 items, $\alpha = 0.803$ — but on a superseded stimulus set of which 16 of 140 items carry over to the shipped build, so no such confirmation exists for v1's items and v1 Gem is described throughout as shortcut-attacked and baseline-hard, never as human-verified. Second, Gem is anti-selected: it is chosen against the reference predictors, which therefore score near zero on it by construction. Its empirical floor is an artifact of that selection and is not a chance floor, so a normalized Gem score is read against the floor measured on the unselected head and never against the floor measured on Gem itself.

Why declared visibility matters. Every Query declares the modalities and the span it may see. That declaration is what makes the history and modality ablations of §4 controlled manipulations of a single item rather than comparisons across different item sets: masking a declared channel leaves the item population fixed, whereas dropping items that lack a channel confounds the modality effect with item-set composition.

P.1 Normalization coverage, per track

What would close the anchor gap. One measured reference per dataset, on the released stimulus and reported with its n, raters, agreement and modality, restores N_d. Nothing about the definition of N_d needs to change; the missing quantity is evidence we have not collected. Among our own sources, FANToM's human baseline is eleven volunteers, one rater per set, so no agreement statistic exists (Kim et al., 2023), and Social-IQ 2.0 publishes none (Wilf et al., 2023).

The normalization formula. Write m_{model} for the model's score on dataset d, m_{floor} for that metric's chance floor, and m_{human} for its human anchor, human performance on the task as

we pose it and on the same metric. Then $N_d = 100 \times (m_{\text{model}} - m_{\text{floor}})/(m_{\text{human}} - m_{\text{floor}})$. A dataset with no human anchor has no m_{human}, so N_d is undefined and that dataset is left out of the mean, not entered as zero (§2.4).

A capability mean admits only datasets with a computable human anchor, so a track's macro can rest on far fewer items than the track holds. Two counts of the Dynamics track are in circulation in this paper and both are correct. The track holds 1,600 items, the number Table 1 reports; of those, 1,400 carry the capability tag Dynamics and the remaining 200 carry Affect. By source the same 1,600 splits 804 casino_acsf, 596 OpenToM, 200 casino. The two partitions coincide arithmetically — the 1,400 capability-Dynamics items are precisely the 804 plus the 596 — which is why the two numbers look like a discrepancy and are not one. With that separated: the hole is larger on the flagship track than on Core, because neither casino_acsf nor OpenToM has a human anchor, so that same 87.5% (1,400 of 1,600) is absent from the mean and the macro there rests on the 200 anchored CaSiNo items. Gem is 60.4% absent. On Core the absent capability is ToM, 4,432 of the track's 7,500 items, every one OpenToM, which ships no anchor; and the only anchored ToM source anywhere in the benchmark is FANToM, all 219 of whose items are quarantined, which is also why the ToM row on All cannot be used. These are shares of items in a capability that is absent from the mean, not shares of items scored incorrectly, and they are read off the frozen build rather than estimated.

Table 1, in full. Counts in Table 1 are items, not episodes: one episode carries queries of several capabilities, which is what makes the paired analysis of §5.2 possible, and the overlap row prevents silent double-counting between Gem and Dynamics. Two asymmetries are visible and are discussed in §6: the multimodal sources are static-heavy — Social-IQ 2.0 contributes no Dynamics items, a census result rather than an omission — while Dynamics is carried by text, and only one source supplies human behavioural outcomes from real interaction. The excluded block records resources frequently assumed available: Social Genome's authors deliberately withhold the reasoning-trace ground truth to prevent contamination, so it cannot be scored locally at all. The § row, OpenToM, is reported per dataset but excluded from the ToM capability mean for want of a human anchor — and it is the whole of ToM on Core, so that exclusion removes 59.1% of the track from the macro — and with MELD's and CaSiNo's anchors withdrawn, the rest (§2.4). Every licence in that table was read from a primary source — the repository LICENSE file, the dataset card or the DUA — on 2026-08-31.

The $^{\|}$ rows: source modality against served modality. MELD and Social-IQ 2.0 are audiovisual corpora, V+A+T at source. Neither is served that way, and no model evaluated in this paper saw a frame or a waveform. A MELD item is served as the speaker-labelled dialogue transcript, and a Social-IQ 2.0 item as the question and its four options with no transcript at all. Both are text-input tasks by construction, not by accident of a local cache: every Episode.media value is a manifest key the prompt builder renders as a string. The scores reported for these two rows are therefore text-input scores, and how much a model with the missing channels would gain on them is unmeasured (Appendix T.10).

P.2 The six quality-control filters, by name

The filters of §3.3 are answer-only, null-history, shuffled-history, lexical-cue, source-prediction, and near-duplicate detection with contamination probes. The copy-from-prefix family added in this build — answer with a candidate the prompt already names — is the only one that reaches an open-answer head; the two OpenToM heads it caught are in Appendix S, with the cross-source exposure audit, whose exact-string matching returns 0 hits over all 49 (source, corpus) pairs. The filters do not all remove the same thing. Three of them — null-history, source-prediction and near-duplicate — delete the items they flag. The other three — answer-only, shuffled-history and lexical-cue — are cue-balanced: they delete items the trivial policy gets right, and keep deleting until that policy is back at the head's floor. The manifest gives the reason. Deleting every item a head-level artifact answers correctly would invert the artifact rather than remove it, because the survivors would be exactly the items where the shortcut points at a wrong answer. Cue-balanced removal makes the shortcut uninformative instead, and the manifest records that it is the expensive

option. Where the artifact is distributional — a property of a head's answer distribution rather than of any one item — the whole head is quarantined instead of either. Per-filter counts are in Appendix H and the four source-data defects they caught in Appendix S.

What the audit does not cover. An audit bounds only the policies it runs, and only the heads it can run them on. The copy-from-prefix family needs a discrete answer space for a copied span to land in, so on three of the benchmark's seventeen heads it declares itself inapplicable, with a recorded reason rather than a 0.0000 a reader would score as clean: casino/satisfaction and casino_acsf/outcome are continuous, sotopia/- is action-valued with a null target. Two of those are Dynamics heads, half the track's four, carrying 619 of its 1,600 items (38.7%), so no claim that every Dynamics head is clean is supportable: those items are untested, which is a different fact from tested and clean. On Gem the limit is sample size rather than applicability — its assessed heads hold between 9 and 146 items, and a margin estimated on nine items carries sampling error wider than any tolerance the audit acts on, so "nothing detected" there is a statement about those items and not a certificate of absence. Where the family does report, the supportable verdict is that no above-tolerance shortcut was detected by the tested family on the assessed heads; a longer-span extractor, a learned copier or a paraphrase-tolerant matcher could reach what these policies cannot. The tolerances themselves — 0.05, 0.15 and 0.25 — are operational curation conventions for when we act on a head: none is a critical value, none has a null hypothesis behind it, and crossing one is not statistical evidence of anything.

Q Evaluation protocol: full specification

This appendix carries the protocol detail relocated from §4: the model set with its version strings, the frozen prompt and decoding configuration, the retry and parsing policy, the full definitions of the eight control conditions and two deterministic baseline policies named in §4, and the scoring and statistical procedures. The limitations that follow from this protocol rather than from the data sources are collected in Appendix S.

Q.1 Model set, version strings and cost

A capability claim is reproducible only if the artifact evaluated is identifiable, so for every model we record the served version string the API returned at run time rather than the marketing name, with the run date, the context window, whether the model is multimodal, and the cost incurred. Every model in the v1 sweep is served through a provider API, so no open-weight checkpoint revision or quantization exists to record, and Table 7 states which fields a row could not carry rather than leaving them blank. No size ladder is listed, because the sweep contains none.

Q.2 Prompt templates

One template per target type is frozen before any full-scale run and not modified afterwards. It presents the episode up to the observation cut, states the query, and specifies the output format in the target's native space; for Dynamics queries it states explicitly that the answer concerns events after the presented context. No per-model prompt tuning is performed: tuning per model would make the cross-capability comparison uninterpretable. Prompt sensitivity is not measured in v1: no paraphrase sweep was run, so the scores here are conditional on one frozen wording per interface and the paper claims no robustness to rewording. The three templates are reproduced verbatim, with the output-format specification and the observation-cut rendering, in Appendix F.

Q.3 Decoding parameters

Decoding parameters are frozen alongside the prompts and held identical across models to the extent each API permits; every deviation forced by an API is listed explicitly rather than silently absorbed, since an unlisted deviation would be indistinguishable from a capability difference. The settings are written out in Appendix F. One deviation is forced on us and is

Model id	Served version string	Provider	Run date	Cost (USD)
claude-haiku-4-5	claude-haiku-4-5-20251001	Anthropic	2026-09-13	2.45
claude-sonnet-4-6	claude-sonnet-4-6	Anthropic	2026-09-13	9.09
claude-sonnet-5	claude-sonnet-5	Anthropic	2026-09-13	6.19
claude-opus-4-8	claude-opus-4-8	Anthropic	2026-09-13	10.38
claude-opus-5	claude-opus-5	Anthropic	2026-09-13	13.89
Total, paired census (792 items × 5 models, plus a 50-item probe)				42.16
Earlier off-track sweep, same five ids, 2026-09-03 and 2026-09-07				176.88
claude-opus-4-6, deleted for a 39.2% model-error rate, not re-run				—

Table 7: The evaluated model set. Cost is actual, read from each response's reported token usage rather than projected. Four fields this table cannot carry. (i) Served version string: only one endpoint returned a dated build identifier; the other four returned the alias we asked for, so on those rows the "served version" is the alias and is not evidence that the served weights were stable across the run — a reproducibility gap in the provider's API, and recording it is the point of the column. (ii) Release date and (iii) context window are not recorded by the harness and are not reconstructed here. (iv) Open weights: none, every model being API-served. claude-opus-4-6 was run and discarded: at a 39.2% model-error rate its scores are not a measurement, and the remaining $28.12 of the $205.00 cap did not cover the projected $35.16 re-run.

not a choice: no sampling parameter can be set at all, so every model runs at its provider default and determinism rests on the content-hashed templates, the fixed sampling seed, the captured version string and the published per-item outputs instead.

Q.4 Retry, parsing and failure policy

A deterministic parser shared across all models reads outputs into the target's native space, so no model benefits from parser leniency written for it. Transport-level failures (timeouts, rate limits, server errors) are retried under a fixed policy and are not model failures; a returned response that cannot be parsed is a model failure, and is counted, reported per model, and scored as incorrect rather than dropped. Dropping unparseable responses would advantage models whose formatting is unreliable, and on Dynamics it would preferentially drop the items on which a model was least willing to commit. On the paired census the parse-failure rate is 1 of 3,960 model-item pairs: a single Dynamics item on one model, named by episode id in that run's diagnostics, and scored as incorrect. Four of the five models produced none. On the earlier off-track sweep it is 7 of 12,975, all on one model and all on static items. The retry policy is three attempts with exponential backoff; the per-run retry count is not written to any artifact, so it is not reported — a number we did not record is not a number we will supply later.

Q.5 Control conditions: full definitions

Each control is identical to the main protocol except for the single kind of information it removes, so the difference between a control and the full condition is attributable to that information alone. Controls run on the same items, with the same template, decoding configuration and parser.

- Random. Uniform sampling over the target's label space, giving the chance floor in closed form where the label space permits and by simulation otherwise. Reported per target type, because the static and dynamics halves do not share a floor.
- Majority. The most frequent label in the split, far above the random floor on skewed targets and the stricter of the two. The heads it is measured on ship no training partition — the sources either have none or, for OpenToM, forbid training use and refuse the split in code (Ethics Statement) — so this predictor is fitted on the evaluation labels themselves: an upper bound on what a label-prior predictor could achieve, not a chance level, and read as such wherever it appears. CaSiNo is the one

source shipping a training partition, and only the learned forecasters of Table 15 use it.

- Answer-only. The query and answer options without the episode. It bounds how much of a score is recoverable from priors over the answer space, and it is the filter that removes items solvable without any social content.
- Text-only. Transcript without video or audio, on the multimodal subset only, isolating the nonverbal channels. Modalities are masked rather than removed, so the item set is held fixed and the modality effect is not confounded with item-set composition. In v1 the multimodal subset is entirely static — every Dynamics item is text — so this control is defined on the static half only and produces no Dynamics row.
- Transcript-only. Only the written dialogue survives. It is stricter than text-only: side channels that happen to be textual, such as a symbolic environment state, are masked as well, so what remains is what a human transcriber would have produced. An episode with no separate transcript keeps its plain text rather than being emptied, which holds the item set at the size of the full run.
- Null-history. The conversation history is withheld entirely, the final utterance included. It bounds what the query, the options and the model's priors supply on their own. It is not an admission test: Dynamics membership is fixed by target function, this filter removed a single MELD item out of 20,389 examined, and no Dynamics source was touched by it, so no item in that track was admitted by surviving it. Whether Dynamics items need the conversation is untested here — no language model has been run under this control (§6).
- Last-utterance-only. The query, the answer options and the final turn of the history; every earlier turn is dropped. The participants and the media stay, so the one thing withheld is the conversation before its last turn. It separates what the last thing said carries from what the conversation carries, and it sits between null-history and the full condition. Turns are split on lines only. A history that arrives as one paragraph is therefore a single utterance and its item is left unchanged, contributing nothing to the contrast; the harness counts those items rather than hiding them.
- Shuffled-history. The history is present but temporally permuted, separating "uses the history" from "uses its order": a forecaster indifferent to the order of the conversation is not modelling how the situation developed. The permutation leaves each turn's index attached to its text, so a model that reads the indices can partly undo it and the measured effect is a lower bound on the temporal shortcut.
- Name-removed. Every speaker name is replaced by a neutral token: "Speaker A", "Speaker B", and so on. Names carry priors about gender and culture, and in a contaminated model the memory of the source corpus itself, so a score that depends on them is not a measurement of the situation. The map is applied to the history, the query, the candidate action, the participant list, the evidence text and the gold target. Renaming the gold is what keeps an item whose answer is a speaker name answerable; it is a relabeling, and each item records it.
- Temporal-shuffle. The turn order is permuted and the explicit time markers are permuted too, independently of the turns, so neither the order nor the clock recovers the original sequence. It is the stricter version of shuffled-history, aimed at Dynamics, where the question is what happens next. The question's own timestamp is left alone, because moving it would change which question is asked. Turn ids are also left attached to their own text, since they are the anchors the gold evidence cites, so on a corpus that numbers its turns this control removes the clock and the order but not the index — again a lower bound on the temporal shortcut.

The identity-withholding control, Name-removed above, applies only where the source carries speaker names or role labels. Where it does, it would be reported alongside the others. None of these controls has been run on an evaluated model, and this paper reports no control result. The harness implements all of them; the floors quoted throughout come from a different thing, the degenerate baselines in runs/e5_controls (MELD) and runs/e5_controls_native (CaSiNo-ACSF) — random, majority-class, constant-outcome and a hashed-token ridge. A baseline bounds what the answer space alone achieves; a control

measures what a model loses when information is removed from its input, and no run on disk does the second. Appendix T.5 states what follows, including the one consequence the Dynamics track most depends on.

Q.6 Scoring implementation

Metric definitions, the proper scoring rule and its normalization, the chance floor per target type and the aggregation weights are in Appendix G. Two implementation choices matter for interpretation. A capability score is the unweighted mean of its datasets' normalized scores: the denominator is the number of datasets, not of items, so a large corpus cannot decide the benchmark while a small one moves it by a rounding error. Item counts ride along for reporting and interval width, never as a weight. No capability mean is reported in v1 at all: normalization needs a measured human score and no dataset has one, so the rule acts on nothing here. And the static and dynamics halves are never compared as raw scores, because they differ in chance floor and label cardinality; the quantities that carry claims are floor-normalised, $(v - \text{floor})/(1 - \text{floor})$ against each half's own measured floor. They are not control-adjusted: no control has been run. The rule is written out in Appendix G. One statement belongs here because it governs comparability: we report the textbook multiclass sum form over the item's fixed candidate set, range $[0, 2]$, so our numbers are directly comparable to published ones, and a reproducer who computes the per-label mean form instead will be low by a factor of the label count. The frozen All track holds 19,107 items over 2,043 distinct source contexts, 9.4 items per context on average and very unevenly distributed: OpenToM carries 13,573 items over 596 narratives, 22.8 per context, and the action-conditioned negotiation head 19.6, while Social-IQ 2.0 carries exactly one. An item-level bootstrap over OpenToM would therefore overstate its effective sample size by more than an order of magnitude, which is why the resampling unit is the context wherever a source replays one context into many questions. Every published interval records the unit it resampled and the n it was computed on, and the four differ. The per-model paired intervals resample pairs, $n = 396$, one index matrix shared by every model and both halves because the two halves of a pair share a conversation prefix and a timestep. The model-level relationship resamples models, $n =$ five — which is why its intervals are wide, and the width is the result rather than a defect — and the same coefficient resampled on pairs is reported beside it. The item-set comparison resamples items within each dataset, 492 CaSiNo and 700 each of FANToM, MELD and OpenToM. Table 14 carries two frozen-Core columns, one per build, and they overlap that draw by different amounts. The earlier build's column scores 467, 0, 622 and 244 of it, 1,333 items; this build's column scores 467, 0, 697 and 226, 1,390 items. The published intervals were computed on the earlier build's overlap. The action-conditioned track resamples whole dialogues, 41 of them behind 835 items. All use the percentile method at $\alpha = 0.05$ with 10,000 resamples and seed 20260913. Comparisons between models or between capabilities carry a significance test with a correction for the number of comparisons made. Correlations between the static composite and Dynamics are reported as both Pearson r and Spearman ρ over models, always with an interval and with leave-one-model-out behaviour, and never as a point estimate alone.

The procedures, exactly. Intervals are percentile bootstraps of the mean: 1,000 resamples of size N drawn with replacement, reported at the 2.5th and 97.5th percentiles of the resampled means. All randomness comes from a seeded local generator at seed 0 and never from global state, so identical inputs reproduce identical bounds from the run record alone. A single observation yields a degenerate interval rather than an error, so a small Gem slice still reports.

Comparisons between two systems evaluated on the same episodes are paired by item and never pooled by model: treating them as independent samples discards the pairing and inflates the variance. The test is a sign-flip permutation test, enumerated exactly where the item count lets the null be listed and estimated by Monte Carlo otherwise, under a $(\text{count} + 1)/(n + 1)$ convention so that a simulated p-value is never reported as 0 — it cannot be smaller than its own resolution. Effect size accompanies the test as a paired Cohen's d_z. Families of comparisons carry a Holm step-down correction controlling the family-wise error rate.

Model-level correlations between the static composite and Dynamics are reported as both Pearson r and Spearman ρ (Pearson of midranks, ties averaged), always with the model count and the model list attached, and always with the leave-one-model-out range: the statistic recomputed with each model held out in turn, the full range reported beside the point estimate. With five models that range is exact rather than estimated. An undefined statistic — fewer than two models, or a constant ranking on one side — returns not-a-number and is reported as undefined, never as zero, because "no correlation is defined" and "measured zero correlation" are different findings.

Q.7 Model set: what one provider does and does not give us

A controlled size ladder — the same family at more than one size, training data and post-training recipe held fixed — would be the one manipulation that varies scale alone, and we do not have one. All five evaluated models come from one provider, and a single provider's product tiers are not that ladder: tier and generation both change across them, so scale, training data and post-training recipe move together, and the run log records no parameter count for any of the five. We therefore report no scaling result and make no claim about families, within or across. A second provider was attempted and its account was exhausted before any usable run completed, so no between-provider quantity is computed or reported.

Q.8 Two scoring rationales

Distributional Dynamics targets are scored by a proper scoring rule rather than by top-1 accuracy, so that a model is rewarded for placing calibrated probability mass on the future that occurred rather than for a confident guess. Separately, an aggregation choice that changed the sign of a headline finding would not be a finding; §T.9 names the reanalyses of that kind this release cannot compute, and why.

Q.9 Expert reference: what it is not

The expert reference is collected, it was produced by people, and it is not a human baseline. It is two of the authors — human raters, not models — rating 139 items, reported with its item count, rater count and inter-rater agreement attached, and it is neither human performance — two expert raters are not a population — nor a ceiling. It is also not a reference on the released artifact: the packet was frozen against the v1 adapters, 61 of 140 items are served by the shipped build at all, a further 6 sit in heads the quarantine bars, and 16 carry over unchanged. The reference is therefore reported for what it is — a rating of a superseded packet, useful as an agreement study and a mis-gold audit — and claims in §5 are scoped accordingly. The replacement packet, rebuilt from the shipped build with every item checked for membership and quarantine, is unrated. The instrument is in Appendix I, the composition audit in Appendix S.15, and the results of the rating in Appendix T.4.

R Extended related work

This appendix carries the material that §1 compresses: the three published meanings of the phrase "social world model"; the closest structural precedents for action-conditioned forecasting; the standing benchmark traditions with their per-benchmark detail; and the long versions of three positioning arguments — the nearest prior measurement, the opposing hypothesis, and the benchmarks that carry both axes without relating them.

R.1 How to read the comparison table

Table 8 sets this benchmark beside the three prior benchmarks it is closest to. Each row is one question a reader can put to any of the four papers and then check against the paper itself. Episodes asks what material the questions are about; a conversation real people had is different evidence from a story a language model wrote. Forecast gold asks what a predicted answer is scored against: a label somebody wrote for the benchmark, or something a participant afterwards did or reported. Both questions share one context? asks whether

	Gu et al. (2026)	Yadav et al. (2026)	Lubrano et al. (2026)	This paper
Episodes	stories written by a language model	recorded sessions, from 3 corpora	recorded conversations	recorded human negotiations
Forecast gold	a label written for the benchmark, checked for plausibility	written trajectory options over mental-state profiles	the participant's own preferred-behaviour label	what the participant reported after the session
Both questions share one context?	yes — scenarios "identical" across the three question types	no — the transcript is withheld from the prospective task	not applicable	yes — byte-identical prefix at a declared cut
Observation cut declared?	implicit in the story	profile only	not applicable	declared per item; what the model may see is part of the Query
Static labels	written for the benchmark	written over corpus states	collected	audited reworkings of the corpus's own annotations
Shortcut / partial-input baselines on every head?	not reported	not reported	not reported	six filters plus a copy family, per head, into the manifest (§3.3)
Corpora on the paired axis	1 (generated)	3	1	1 — the v1 limit

Table 8: What distinguishes this benchmark, written as questions a reader can check. Read the third row first: putting several questions to one scenario held constant is Gu et al.'s design and is not claimed here. What is claimed is the combination in the last column — recorded human interaction, forecasts scored against what the participants afterwards did and reported, an explicit observation cut both questions share, static labels that are audited reworkings of annotations the corpus already carried, and a partial-input baseline on every head — each of which is individually precedented. The last row records the axis on which v1 is weakest, and it is not an axis on which we are ahead. Cells were read first-hand from each primary source on the dates in this file's header. "Not reported" means we did not find it in the paper, not that it was not done.

the comprehension question and the prediction question are put to the very same passage; if they are not, a difference between the two scores may be nothing but a difference in the material. Observation cut declared? asks whether the paper states where the cut is — the point in a conversation where the benchmark stops showing the model, so the model cannot see the answer. Static labels asks where the comprehension answers come from: written fresh for the benchmark, or derived from annotations the original corpus already carried. Shortcut / partial-input baselines on every head? asks whether the paper reports, for every head — a group of items sharing a question form and an answer space, such as all the emotion questions from one corpus — a baseline that answers with part of the input taken away, for instance without reading the conversation at all; if such a baseline scores well, the items can be solved without the reasoning they claim to test. Corpora on the paired axis asks how many separately collected sources supply the material on which both questions are asked. That is the row on which v1 is weakest.

R.2 Social world models: three published referents, none of them ours

The phrase already has three published meanings, and none is ours. For Zhou et al. (2025b) a social world model is a structured piece of text listing the agents, what they observe and what they believe: a language model reads a narrative, writes that structure out, and rolls it forward one step while it answers. The approach is text-only and prompted rather than

learned, collects no data from human subjects, and its authors state that their results do not establish generalization to real-world interaction. For Yu et al. (2026a) it is a learned transition over beliefs traded on prediction markets, $s_{t+1} \sim P_\theta(\cdot \mid s_t, e_t)$, with a mode for simulating counterfactuals; the variable it conditions on is an outside news event, not a move made by a participant, and the paper contains no dialogue and no affect. For Tsvilodub et al. (2026) the phrase names hidden mental-state representations a model reuses across static theory-of-mind and pragmatics tasks, and Tang et al. (2026) use their term in that sense as well. We use the phrase only here and in the word "toward" in our title, and claim neither the term nor the hypothesis of Tsvilodub et al. (2026); our task is what a system carrying such a model would have to be able to do.

R.3 Action-conditioned forecasting of social state

The structural precedent closest to §2.3 is the proactive intent world model of Zhang et al. (2026), which estimates a hidden customer state from video recorded before the interaction, predicts an outcome for each move in a fixed set (one of which is doing nothing), and selects the best move. It differs from our task in every dimension that matters for a copilot: it watches one customer rather than a pair of people, its state is a discrete intent phase rather than a continuous quantity, its counterfactuals are generated by a language model rather than observed, and by the authors' own report its world model is more useful as a training-time auxiliary than as a planner.

Continuous state within a pair of people is forecast, without conditioning on an action, by multimodal affect models (Zhao et al., 2026), and measured observationally by the dyadic-influence and rapport-trajectory challenges of Hume AI and the DaiKon organisers (2026), whose concordance metric and speaker-independent evaluation we adopt, as well as by the time-continuous emotion tradition running from AVEC to MSP-Conversation (Ringeval et al., 2019; Martinez-Lucas et al., 2020; 2026). Self-reported affect within a pair exists at small scale (Park et al., 2020). Yadav et al. (2026) benchmark the forecasting of dialogue trajectories from mental-state profiles as multiple choice, and their finding that models infer states they cannot then use predictively is one motivation for conditioning on the move explicitly. None of these conditions a continuous-state forecast on the move the acting participant makes, and none has an interventional check. Yu et al. (2026b) argue, for the embodied case, that world models should be evaluated for decision-making rather than for realism, and our metrics follow their evidence ladder.

R.4 The four standing traditions

Static social, mental-state and grounded benchmarks. The largest tradition scores how well a model describes a social situation it can see in full: recognizing affect in conversation, answering social commonsense questions, judging norms, and drawing social inferences from several modalities at once (Zadeh et al., 2019; Poria et al., 2019; Zhou et al., 2025a; Qi et al., 2026). A second probes how a model represents other minds, through false belief, information asymmetry and knowledge tracking, in text and in situated settings (Kim et al., 2023; Xu et al., 2024; YS et al., 2026; Villa-Cueva et al., 2025; Bortoletto et al., 2025). A third grounds social inference in nonverbal evidence and scores the reasoning trace as well as the answer (Mathur et al., 2025). These benchmarks are well built for what they measure, and SocialWorldBench's static side is largely a renormalization of their annotations. What they share is a property about time: the evidence for the correct answer sits inside the context the model is shown, and nothing is asked about the interval after it.

Our census quantifies how complete that property is in the flagship multimodal case. The method was a twelve-alternative future-orientation pattern (will, won't, going to, what happens next, next, after this/that, later, end up, likely to, predict, outcome, about to), run over all 8,817 Social-IQ 2.0 questions (Wilf et al., 2023). It returned 93 candidates, one author read every one, and every one is a false positive in two recurring shapes: spatial ("sits down next to the woman") and retrospective ("why did the man chase after the woman when she was about to walk away"). The screen is deliberately over-inclusive rather than precise — a recall device whose output was adjudicated by reading — so its false-positive rate is not a finding and its recall is bounded by the pattern list, which we publish for that

reason. What that licenses, and what it does not. A screen that reads its own 93 candidates establishes a fact about those 93. It establishes nothing about the 8,724 it excluded, and its recall over them is bounded by the pattern list. So the finding we state is the procedural one: the screening procedure found no future-prediction question. We then bound the rest directly. We drew 300 questions uniformly at random from the excluded remainder and read them, and none of the 300 is a future-prediction item either. With no hits in 300 draws, the one-sided 95% upper bound on prevalence in that remainder is 1.0%, at most about 88 questions. The nearest misses are worth naming, because each is still static under this paper's own definition. "Did the man think his team can win?" is a belief about a future event, but it is held and scored inside the clip. "Does the woman manage to convince the man of her theory?" is an outcome that resolves within the presented window. In total, 393 of the 8,817 have now been read. The supportable claim is that Social-IQ 2.0 asks for essentially no forecast about the interval after the observed context. It is not that a certified zero was established by regex.

Interactive and agentic environments. A fourth tradition places models inside social episodes as participants and scores the outcome they achieve (Zhou et al., 2024; Wang et al., 2024; Chawla et al., 2021; He et al., 2018; Moore et al., 2025). This tradition does involve the future, but it predominantly measures acting rather than forecasting: an agent that reaches its goal is rewarded whether or not it anticipated its partner, so an outcome score confounds predictive accuracy with policy quality. The exception is rare enough to name: one recent environment scores an explicit next-message forecast inside a live interaction (Chuang et al., 2025). That confound is why we score forecasts over recorded episodes rather than rewards over live ones.

R.5 The nearest prior measurement, in full

Momentè et al. (2025) is the closest existing measurement to ours, and the one we inherit the question from. They score nine open-weight models of at most 72B parameters on a battery of cognitive-ability tests, among them Social IQa, EQ-Bench, a SimpleToM split and LM-Pragmatics; they score the same models on clembench interactive dialogue games; and they relate each ability separately to the game ordering with Kendall τ-b. The per-ability structure matters: the negative coefficient is the Social IQa row on its own, and in the same figure the other social abilities correlate positively with the games (Appendix R.10).

The study is carefully executed and reports its own scope honestly. Our disagreement is with the unit of analysis, not the execution. Because the two sides are different benchmarks, a rank change between them admits two explanations the design cannot separate: the construct may differ, or the corpus may differ in domain, format, annotation practice and difficulty. Their coefficient is therefore a sound estimate of the relation between two benchmark suites; the relation between two capabilities on common items, estimated over more than one corpus family, remains open. SocialWorldBench changes the unit of analysis (§2.2). It does not extend the model pool, and we say so rather than imply otherwise: their nine open-weight models are a broader pool than the five same-provider models of our demonstration runs (§5), so the generalization constraint their limitations section identifies binds us harder than it binds them.

R.6 The opposing hypothesis, and whose term this is

An organized counter-position holds that social capabilities in language models are functionally integrated rather than dissociated. Tsvilodub et al. (2026) evaluate 48 models over 22 Theory-of-Mind and 16 pragmatics datasets, with Bayesian regression and causal subnetwork ablation, and read the result as evidence that models develop interconnected "social world models" rather than isolated competencies. That phrase is theirs, and is also used by Tang et al. (2026) and Zhou et al. (2025b); we use the name SocialWorldBench for a benchmark and claim no ownership of the term or of the hypothesis it names. Their evidence is drawn entirely from static tasks, which is exactly where the integration account and the dissociation account are hardest to tell apart. Our benchmark puts the question on a

predictive axis without presuming the answer, and a high model-level correlation on paired stimuli would be evidence for their hypothesis, which we commit to reporting as such.

R.7 Benchmarks that hold both axes without relating them

Several recent benchmarks do contain a comprehension track and a prediction track (Kong et al., 2026; Tang et al., 2026), and at least one study evaluates the same models on static Theory-of-Mind benchmarks and on an interactive setting (Zhou et al., 2025b). We are explicit that this region of the design space is occupied and that we are not the first into it. What these works do not report is a per-model relation between the two tracks on shared items: they print the two scores side by side and do not correlate them. The works that do correlate them (Yadav et al., 2026; Lubrano et al., 2026) do so within a single corpus family.

A claim withdrawn, 13 September 2026. Earlier drafts continued that sentence with "and replicated across independent sources", and called cross-source replication the gap the paired-episode design fills. We withdraw it. Our own paired forecasting axis carries a single corpus, CaSiNo, which is also the only one of our sources that supplies outcomes reported by the people who took part. So on the axis in question, v1 is itself a single corpus family. That is the exact deficiency the withdrawn sentence charged to Yadav et al. (2026) and Lubrano et al. (2026).

What survives is the design claim of §1. Both questions are put to one conversation at one cut, so a rank change between the two scores cannot also be a change of corpus. That is narrower than a cross-source claim and narrower than an empty-quadrant argument. Unlike either, it is true of what we shipped.

R.8 Auditing benchmark items: the procedure this paper inherits

§3.3 states, compactly, that the procedure SocialWorldBench audits with is inherited rather than invented. This subsection gives the lineage attack by attack, so a reader can check which parts are standard practice and which are ours.

The attacks, and where each comes from. Partial input is the hypothesis-only baseline of Poliak et al. (2018), the artifact analysis and hard-subset protocol of Gururangan et al. (2018), and the question-answering shape of Kaushik & Lipton (2018), which is the closest structural match to a social item made of a context plus a question. Our answer-without-reading-the-conversation policies are that family, and they inherit its limit from Feng et al. (2019): a successful partial-input baseline shows that a dataset is cheatable, while a failing one does not show that the dataset is clean. That asymmetry is why §3.3 says an audit bounds only the policies it runs, and why §1 reports a case where our own family silently could not run. Order and history: shuffling the utterance order of a dialogue history appears, as far as we found, once, in Sankar et al. (2019), which measures the change in a generative model's perplexity where ours measures a change in accuracy on scored items, an adaptation and not their protocol; word- and sentence-order permutation is better attested (Sinha et al., 2021; Pham et al., 2021), event-order and single-frame probes are the multimodal counterparts (Buch et al., 2022; Lei et al., 2023), and Sugawara et al. (2020) gives the framework the whole set instantiates: ablate a feature that a requisite skill depends on, then see whether the score survives. Names and entities: Hermann et al. (2015) introduced entity anonymization as a deliberate anti-shortcut device, Chen et al. (2016) record its cost — anonymization can make an item unanswerable, and so inflate the apparent effect — and Shwartz et al. (2020) show that name artifacts persist in pretrained language models. Lexical cues: our per-token cue statistic is the pointwise mutual information analysis of Gururangan et al. (2018); the remedies run from adversarial filtering (Zellers et al., 2018; 2019) to its embedding-level successor (Sakaguchi et al., 2020); Niven & Kao (2019) show that a headline result can be entirely a cue artifact; Kavumba et al. (2019) give the closest format match to our items, attacking single-token cues across answer options and rebuilding a balanced counter-set; and Gardner et al. (2021) supply the theoretical licence for treating any simple feature–label correlation on a complex task as spurious. Source prediction is the thinnest lineage: the only true instance we found is "Name That Dataset" (Torralba &

Efros, 2011), which is vision and operates on raw images, and Geva et al. (2019) recover the annotator rather than the dataset, a different quantity. Our text-item version is an adaptation, and we label it as one.

Where social and mental-state benchmarks sit, and what is ours. The concern is well established in this literature: Ma et al. (2023) write that situated theory-of-mind benchmarks are "prone to shortcuts and data leakage". The procedures differ from input ablation. Kim et al. (2023) detect illusory theory of mind by asking several question types that require the same underlying reasoning, and Shapira et al. (2024) stress-test social reasoning with hand-built adversarial minimal pairs; a search of their materials found no history shuffling, no name perturbation and no no-context baseline in either. Both are careful instruments, and neither substitutes for the other: a question-type consistency check catches an inconsistency an ablation cannot, and an ablation catches a solvable item a consistency check will pass. What we add is not the attacks. It is that the family is run over every head on every build; that its result is a released field of the manifest rather than a claim in a paper; and that the action taken on a head that fails, whether filtering the items or quarantining the head, is recorded with the verdict and the remedy. We also found no prior work running a comparable set as a single auditing suite over a social benchmark; Sugawara et al. (2020) is the closest, as a framework rather than a suite.

R.9 Benchmark construction methodology

Our construction follows the curate → normalize → broad-plus-hard-set → held-out-generalization recipe of IntentGrasp (Yin et al., 2026), which assembles a large All evaluation set and a small, balanced, difficult Gem set from many source corpora, then tests transfer with leave-one-domain-out experiments. We adopt that structure and add the paired temporal axis it does not have.

R.10 The five published estimates, and what each actually reports

§1 states that published associations between the static and dynamic axes vary across task definitions. This subsection gives the full inventory. Every entry was re-read first-hand against the primary source on 13 September 2026, each with one targeted retrieval and one deliberately non-leading one, and the verbatim quotations are recorded in the corresponding citations.bib entries. That pass corrected three of the five characterizations carried in earlier drafts, including one substantive scope error, and the corrections are stated below rather than made silently.

Momentè et al.: one cell of a matrix, not an aggregate. Momentè et al. (2025) score nine open-weight models (3.8B–72B) on a battery of cognitive-ability tests and relate each to performance on clembench interactive dialogue games and to large QA benchmarks with Kendall τ-b. Their Figure 3 is therefore a matrix, one row per ability and one column per evaluation family. The τ-b of -0.17 is the single Social-Reasoning row, that is, Social IQa, against games. It is not an aggregate of static social benchmarks, and an earlier version of this appendix described it as one, pooling Social IQa, EQ-Bench, a SimpleToM split and LM-Pragmatics against games. In the same figure the remaining social rows are positive against games: pragmatics (LM-Pragmatics) is the largest, theory of mind (the SimpleToM mental-state split) and emotional intelligence (EQ-Bench) are also positive, and attribution/judgment is approximately zero at -0.06. The paper's own conclusion runs the other way from a bare negative reading of the -0.17 cell: its abstract states that "differences emerge regarding core executive functions and social/emotional skills, which correlate more with games", and its body that "interactive tests correlate more strongly with socio-emotional skills compared to static tests".

One limit on our own check. Figure 3 is a raster heatmap whose cell numerals are not recoverable from the PDF text layer, so we confirmed the sign pattern independently by recomputing τ-b from the paper's own per-model score table (their Table 3, nine models) against the mean of its clembench game rows. That gave -0.28 for Social Reasoning, -0.06 for attribution/judgment, and positive values for theory of mind, emotional intelligence

and pragmatics, with pragmatics largest. Signs and ordering reproduce; magnitudes differ because the paper aggregates games by clemscore rather than by unweighted mean.

Hu et al. (SimBench): against MMLU-Pro specifically. Hu et al. (2026) report a Pearson r of 0.939, stated verbatim as "simulation ability correlates most strongly with knowledge-intensive reasoning (MMLU-Pro, $r = 0.939$)". Two things narrow it: the non-simulation side is one named benchmark, MMLU-Pro, not general capability at large, and "most strongly" means the value is the maximum over the several correlates they examine. Reporting it as this literature's general-capability association therefore overstates it in two separate ways. It remains an across-benchmark estimate.

Yadav et al. (DialToM): and where the p-value belongs. Yadav et al. (2026) report a Spearman ρ of 0.67 over fourteen models, computed inside a single benchmark, relating their retrospective task — mental-state inference from dialogue context, which they call "Literal ToM" — to their prospective task, forecasting state-consistent dialogue trajectories from isolated mental-state profiles without dialogue context, which they call "Functional ToM". Verbatim: "cross-task correlations are weak and statistically non-significant under a linear assumption ($p = 0.35$), though they display a moderate monotonic trend (Spearman $\rho = 0.67$)". The p of 0.35 therefore belongs to a separate linear analysis and is not a significance test of the rank coefficient. We report it as such, and did so before this check.

Lubrano et al. (AttuneBench): axes narrower than the headline. Lubrano et al. (2026) report a model-level Spearman ρ of -0.36 over their eleven models, computed inside a single benchmark. We state its axes precisely, because they are narrower than the paper's headline: the coefficient relates Pairwise Accuracy, which scores preference prediction across three response variants, to Binary HP Accuracy, which scores exact-match agreement with the participant's own preferred-behaviour labels. It is not emotion recognition against preference prediction, and it is distinct from the conversation-level Pearson r of -0.097 reported in the same paper. Their broader qualitative claim, that rankings across the emotion-recognition, behavioural and preference-prediction tracks are "largely independent" of one another, is consistent with it.

Ziems et al.: an unsigned r^2 against a loss, with a negative slope. Ziems et al. (2026) report an r^2 of 0.60 between general-benchmark PC1 — the first principal component of general benchmarking performance — and longitudinal social forecasting over thirty-five base models. Two properties of this statistic matter and were previously misstated here. First, the outcome variable is a loss and the fitted slope is negative (-0.04). Verbatim: "longitudinal forecasting improves with generally more capable models (slope $= -0.04$), and … performance on this task is correlated with general benchmarking performance ($r^2 = 0.60$)". More general capability therefore goes with lower loss and better forecasting, so in substance this is a positive association. Second, r^2 is unsigned and is not a Pearson r: we do not report it as a signed $r = +0.60$, and we do not place it on a common axis with the three signed coefficients above. An earlier version of this appendix called this "the one estimate that cuts against the rest"; that was wrong on both counts and is withdrawn, as is the contradictory framing that §1 carried before that pass, which simultaneously called the same paper the strongest case that the two axes are related.

What the inventory does and does not support. Two of the five estimates are negative and three are positive, so the signs are not uniform. But the five are four different statistics — Kendall τ-b, Pearson r, Spearman ρ and an unsigned r^2 — computed over disjoint model pools, on different units of analysis, and between different pairs of constructs; two use an across-benchmark design and three a within-benchmark one. The dynamic axis is defined variously as interactive gameplay, human-behaviour simulation, context-free trajectory forecasting, first-person behaviour classification and longitudinal forecasting loss. No two of the five share items, model pool, statistic or definition of the dynamic axis. We therefore do not claim that they are incompatible readings of one phenomenon, and we do not treat their numerical range as a spread on a single scale; both framings appeared in earlier drafts and both are withdrawn. The defensible claim is narrower: the published associations vary with

how the two axes are operationalized, and because nothing is held fixed across the five, that variation is uninformative about the underlying relation. It is a reason to build a controlled benchmark, not itself evidence of a conflict.

S Design-level limitations

The limitations in the main-text Limitations section are the ones that bear on how the central measurement should be read. The limitations below follow from the design and the frozen dataset clearance rather than from any result, and are recorded here in full; the evaluation protocol several of them refer to is specified in Appendix Q.

S.1 The interventional probe is deferred

No randomized collection is run for v1. The probe specified in Appendix C needs recruitment, consent and ethics review, so every action-conditioned result in this paper comes from the observational track and is read as conditional prediction, never as an effect. The track stays in the benchmark because the quantity it measures — what a move did when it was assigned rather than chosen — exists in no public dataset, and the protocol, seeded randomizer, instruments and power analysis are released so the probe can be run.

S.2 Observational action effects are confounded, and the moves are mapped

On the observational track the move at each turn was chosen by the speaker. A forecaster that conditions on it learns $P(o \mid h, a)$ for chosen a, which differs from the interventional quantity by whatever made the speaker choose; the paper never reads an observational difference between moves as an effect. The moves are the corpus's strategy annotations mapped onto our library by a fixed table; the mapping is released, and the five library moves the corpus does not observe (pause, summarise, offer two options, delay, suggest a break) appear only in the pilot. The shuffled-action control of §5.3 is what certifies that a forecaster's score on this track reflects the move at all.

S.3 The learned forecasters are deliberately weak

The learned models are linear over hashed text and a one-hot move. They exist to isolate the value of the move under matched capacity, not to set a state of the art, and no claim in this paper depends on their absolute level.

S.4 Prediction targets are inherently uncertain

A social episode has no unique correct continuation. We mitigate this with proper scoring rules, sets of acceptable continuations, and the expert reference, but any reported S_{dynamics} is bounded above by an unknown irreducible component. This is the deeper reason behind the main text's refusal to read a raw static–dynamics gap: comparisons of absolute scores are not comparisons of equally attainable quantities, so the claims are relative ordering across models, reversals between pairs of models, and distance from the expert reference on the same items.

S.5 Pairing removes a confound; it does not remove all of them

Holding conversation, timestep and annotation source fixed removes the corpus confound that motivates this paper. It does not equalize the difficulty of the two sides, does not make the two annotation fields equally reliable, and does not rule out that a source's static and dynamic annotations were produced by correlated annotator judgements. Where a source's two sides come from the same annotator pass, we flag it in the manifest and report the paired estimate with and without that source.

S.6 Scripted sources, derived targets, and what the released track contains

The static half includes scripted television, whose next line was written rather than chosen, and the released forecasting track carries no audio or video. On MELD we can now put a number on what that costs, and it is not small. Two expert raters, given the full speaker-labelled dialogue context this build's adapter supplies, still flag 72.7% and 78.8% of items as underdetermined from the text alone. They agree perfectly on which items those are: on the 19 items both rated, the ambiguity flag gives Cohen's $\kappa = 1.000$. On the separate question of which emotion the utterance carries they agree at Cohen's $\kappa = 0.936$, differing on one of the 19. Their answers match MELD's own label on 29 of 66 ratings (43.9%) of a sample whose seven labels are balanced to a 14.9% modal class.

Every one of their 50 flags names a competing emotion that stays defensible on the repaired stimulus rather than calling the item broken, and the reason is the same each time: the transcript does not carry the delivery. The residual is a reduction, not a clearance. Before the fix the same two raters flagged every MELD State row they were given, 35 of 35 each. They flagged every MELD row of any kind, 48 of 48 on each sheet. They then re-rated under a written instruction to flag exactly as freely as the first time, with both outcomes pre-committed; that instruction is what makes 100% and 72.7%/78.8% comparable at all. Their agreement on the emotion label for this head was already $\kappa = 0.914$ before the fix, so the $\kappa = 0.936$ after it is no evidence that the fix raised agreement; what certifies the residual flags is the different statistic above, $\kappa = 1.000$ on the flag itself. The two ratings are different draws of MELD items rather than the same items rated twice (Appendix T.4.1).

State is MELD, MELD is multimodal, and we serve it as transcript — so State is measured on a channel that does not determine its own label, and that is a property of the task we adapted, not of the raters (Appendix T.4.1, Table 11). MELD dominates the raw paired-construction yield yet contributes no item to the shipped Dynamics track, which is CaSiNo and OpenToM and is entirely text; raw construction yield and released coverage are therefore different populations and are reported separately, under the per-source, per-kind rule of §2.2. The two sources that do carry audio and video are static-only, so $\Delta_{\text{multimodal}}$ — the score difference between full and masked-modality input on the same items — is measurable on the static half and undefined on Dynamics: v1 licenses no claim about whether nonverbal channels help prediction (Table 1, Appendix T.10).

Several targets are also not records of human behaviour. SOTOPIA's rewards are GPT-4 judgements over model-agent episodes, and its 7,200 reward rows cover 450 scenario × agent-pair configurations: not 7,200 observations, and not automatically 450 independent ones, since scenarios and agents recur and the grouping keys are not verified in the released build. FANToM's conversations are model-generated. OpenToM's story drafts are composed with LLMs by its own authors, with an amount of human revision its README does not quantify (Xu et al., 2024), and it is the largest source in both scored tracks. That generation is upstream: our adapter maps the source's own text and labels into our schema with deterministic code and writes no new material. OpenToM also ships no future-prediction question type, so its Dynamics targets are derived by us from plot metadata under a released rule, and its CC BY-NC licence admits it for evaluation only. Two corpora additionally shipped defective answer keys, detailed below.

S.7 Reused data carries its sources' biases

Every item derives from an existing corpus. Scripted television dialogue, crowdworker-authored negotiations and procedurally generated narratives are not samples of naturalistic social behaviour, and their demographic, linguistic and cultural composition is that of their original collection. v1 is English-centric and does not support cross-cultural claims. Post-filter, the scripted source is confined to the static half and is absent from the forecasting one. The shipped Dynamics track is 1,600 items from three sources — the action-conditioned negotiation head 804, OpenToM 596, CaSiNo 200 — and carries no MELD item at all. The paired census is narrower still: one crowdworker-authored negotiation corpus and no scripted dialogue, being the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same

stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale. On the static side the position reverses: MELD is 2,601 of the 7,500 Core items, 34.7%, and it is the only source for the State capability. We report no static-half score with that source removed, because this paper forms no score across datasets at all: the three Core datasets are read on three different native metrics and are never pooled (Appendix A), so "the static half without MELD" is not a quantity this release defines.

S.8 Contamination cannot be excluded, and the exposure is total

The source corpora are public and predate the evaluated models. Two of our five probes ran and three were never scored, so contamination in the evaluated models is unmeasured and this benchmark is never contamination-free. What did run bounds exposure and cross-source overlap, not per-model memorization, and any such audit is weaker still for proprietary models whose training data is undisclosed. Two of the five probes have no published precedent, so they are described in full and released with the harness rather than cited.

The exposure is measured rather than assumed, and the measurement is not reassuring. As shipped, temporal-provenance coverage of the tracks is 0.0: only two adapters carry a first-public date at all and both carry it at month precision, which the ISO date parser rejects outright. The build normalizes every source to a full ISO date and records, per source, the date, its precision and its basis, which takes coverage to 1.000 on both the created-at and first-public fields; these are dataset-level release dates, not per-item authoring dates, and every item carries its precision and basis so that a manufactured day number is not read as a measurement. Because a single assumed training cutoff is a choice that moves the answer, exposure is reported as a ladder over cutoffs rather than as one number. The latest first-public date anywhere in the tracks is 2024-02-01, so the ladder reads 1.000 at a 2025-01-01 cutoff and at every later one: against every model evaluated in this paper, every item was already public. Quoting a lower rate at an earlier cutoff would mean choosing a cutoff no evaluated model has. What the audit does bound is cross-contamination between our own sources: sampled and seeded exact-string matching of each source's items against every other source's corpus files returns 0 hits across all 49 (source, corpus) pairs: seven sources searched against the other six and against the excluded craigslist_bargain corpus, 150 needles sampled per source, no source searched against itself. Absence of a hit there bounds nothing about corpora this build does not hold. Table 9 gives all three in full.

No model-level contamination probe was scored in this release, so no model–source pair is flagged and no exclusion sensitivity is reported. The paraphrase and source-recognition probes are generated — 200 items each — and unscored, because scoring them requires model calls this build does not make; Δ_{live} is undefined rather than zero, there being no Live overlay in v1; and the canary set was minted after the runs on disk were collected, so its zero leak count establishes the detector for the next sweep and bounds nothing about the models already evaluated. Every exposure number in this paper is therefore an audit of the items, not a test of a model.

S.9 Source-data defects, in full

Four defects were found by the shortcut and leakage audits and all four are corrected or quarantined rather than absorbed; a fifth, our own MELD adapter's stripped context, was flagged by the model-rater diagnostic, confirmed independently by both human raters, and is fixed in this build (§6). That fix is why MELD moves from 2,357 to 2,601 items: the near-duplicate filter had been deleting 253 MELD items whose prompts were byte-identical only because the context was stripped — "Push!" and "Hey!" are distinct questions once the dialogue prefix is restored — and it now deletes 8. The re-audit is what licenses the fix, since adding context can manufacture a shortcut as easily as remove one: the head's best trivial policy after the change sits exactly at its floor (0.4810 against 0.4810), with the copy-from-prefix family at 0.0069.

(i) FANToM shipped the correct option at choices[0] in 1,540 of 1,540 belief items; the options are now permuted per item from a recorded seed, verified on every build. (ii) OpenToM

(a) Exposure ladder: share of 19,107 All-track items already public at each cutoff

Assumed cutoff	Items at risk	Exposure rate
2023-01-01	4,072	0.213
2024-01-01	5,534	0.290
2025-01-01	19,107	1.000
2026-01-01	19,107	1.000

(b) Per-source provenance, and the self-audit's verbatim-match rate

Source	Items	First public	Precision / basis	Verbatim
meld	2,601	2018-10-01	month; MELD preprint, Oct 2018	0.850
casino	667	2021-03-01	month; CaSiNo preprint, Mar 2021	0.853
casino_acsf	804	2021-03-01	month; re-cut of CaSiNo, inherited	0.644
social_iq_2	793	2023-01-01	year; release day not verified	0.786
fantom	219	2023-10-01	month; FANToM preprint, Oct 2023	0.781
sotopia	450	2023-10-01	month; SOTOPIA preprint, Oct 2023	0.984
opentom	13,573	2024-02-01	month; OpenToM preprint, Feb 2024	0.980

Table 9: The exposure audit in full, from the frozen build's contamination block (manifest_sha256 139722f4). Panel (a): the ladder is reported over four cutoffs because a single assumed cutoff is a choice that moves the number, and the latest first-public date anywhere in the tracks is 2024-02-01, so every cutoff any evaluated model could have puts the rate at 1.000. Panel (b): no date in this build is an item-authoring date — every one is a dataset-level release date, and each carries the precision and the basis it was derived from, so that a manufactured day number cannot be read as a measurement. Social-IQ 2.0 is the one source dated only to the year, and it is marked. Verbatim is the share of a source's items whose text is found verbatim in that source's own public corpus file; a hit is expected there and the column is the honest statement of exposure, not a defect count. The cross-source version of the same audit returns 0 hits across all 49 (source, corpus) pairs.

ships two copies of its gold that disagree: in the distributed opentom.json the second-order fine-grained answer is the entity's final location in 1,192 of 1,192 items, so a reader that ignores the false belief entirely scores 1.000; we load the corrected release instead. A fourth provenance limitation belongs beside them: SOTOPIA's targets are GPT-4 judgements over LLM-agent episodes rather than human behaviour, and its 7,200 reward rows are replays of 450 scenario × agent-pair configurations, which is the count every interval over that source uses — 450 is an upper bound on the number of observations, not a demonstration that they are independent, scenarios and agents recurring across configurations. OpenToM ships no future-prediction question type at all, so its Dynamics items are derived by us from plot metadata under a released rule and are labelled derived rather than annotated everywhere they appear, and its CC BY-NC licence and no-training README admit it for evaluation only.

(iii) The lexical separation of FANToM's belief options described in the main-text Limitations is a third defect of the same class, distributional rather than per-item: the correct and incorrect options are produced by different generators, so the cue travels with the option text wherever the permutation of (i) puts it. Balanced removal — deleting, from the side the policy solves correctly, until its accuracy on the survivors is back at the head's floor — has a closed form: with policy accuracy a and floor b the fraction of the head that must go is $f = (a - b)/(1 - b)$, which equals twice the margin only when $b = 0.5$; FANToM's accessible-belief head happens to floor at 0.5082, which is why the doubling shorthand an earlier draft used looked right on the one head anyone checked. That identity over both belief heads predicts about 1,101 of their 1,540 items (71.5%). The six filters as actually run removed 1,321 (85.8%), and the two figures are deliberately kept apart: a one-pass closed form assumes a fixed floor, a single policy and deletion drawn only from the correct side, and a multi-pass cascade honours none of the three, so neither number forecasts the other. The remnant then carries a strong positional floor of its own — a new artifact manufactured by the repair rather than a repaired head. It is therefore quarantined: kept in All with a flag on every item, excluded from Core and Gem, and named in the released manifest with its remedy, which is upstream regeneration of the option strings and is not ours to apply. Two consequences are stated in the main text: Core contains no FANToM items, and v1

reports no human-normalized ToM capability mean. The quarantine figure the main text quotes, 3,064 items over four heads, supersedes the earlier audit's 219 over two.

(iv) Two OpenToM location heads carry a shortcut of the same distributional kind and were quarantined on the same rule, but they are worth separating because the audit was at fault as well as the data. The families this benchmark ran until this build — an input-independent floor, a history-free shape family, an order-invariant history family and a mined option-token cue — all operate over a head's choice set, and each of their members begins by skipping any item with fewer than two choices. Both fine-grained location heads are open-answer. They were therefore scored at 0.0000 by every policy, which the manifest recorded and which reads like a clean head; it was an absence of measurement. A copy-from-prefix family — answer with a candidate the prompt already names, selected by most recent mention, first mention or mention count — reaches an open head, and on the pre-filter pools it beats the floor by 0.2492 on location_so_fine and 0.1644 on location_fo_fine.

Both figures use cross-fitting folds grouped by source narrative. The item-wise split they replace (0.2039 and 0.1284) is not an independent split on these heads: OpenToM puts two location_so_fine and four location_fo_fine questions on each of 596 narratives, and sibling questions share the narrative's gold container, so the vocabulary mined on one fold could contain the exact string the policy was about to be scored on in the other. Grouping removes that leak and the measured shortcut went up, which should not be glossed: grouping also shrinks the mined vocabulary — the train fold's coverage of held-out golds falls from 84–90% to 68–82% — and a smaller candidate set is easier to hit, the larger of the two effects on these heads. We report the grouped figure because its folds are the ones that are actually independent, not because it is the larger number. One verdict moves with it: location_fo_fine sits above the 0.15 band under grouped folds and sat below it under the item-wise split. The head was flagged either way; what changed is which band the measurement falls in, and it changed because of the fold rule rather than because of anything about the head.

Both heads are excluded from Core and Gem after the stated repair attempt failed its tolerance within the documented stopping budget, and are retained only as flagged diagnostics in All: five passes of cue-balanced removal on location_so_fine gave up half its items and left the family 0.0750 above floor. That budget — five passes, or roughly half the head retained, whichever came first — was fixed post hoc rather than pre-registered: the probe, its five-pass default and the report that used it were written and committed together, and a later run of the same probe used seven passes, so we record the budget as what was spent and not as a criterion set in advance. What the probe establishes is that this procedure, run against this nine-policy family for this many passes, did not bring either head inside tolerance. It does not establish that the heads are irreparable, that a different procedure would fail, or that further passes would consume the rest of the items, and we claim none of the three. The remedy we would prefer is upstream regeneration of the answer space, so the gold container is not recoverable from mention order. The tolerances these decisions turn on — 0.05, 0.15 and 0.25 — are operational curation conventions for when we act on a head: none is a critical value, none has a null hypothesis behind it, and crossing one is not statistical evidence of anything. The general lesson is recorded here rather than softened: a shortcut audit bounds only the policies it runs, on the heads it can run them on (Appendix P.2), and a zero from a family that cannot express a policy for the head is not evidence of anything.

S.10 Aggregation choices, and how little the macro covers

"Static understanding" is an average over three capabilities that are not equally represented across sources. §T.9 reports what can and cannot be recomputed under alternative weightings, but no aggregation is neutral. The sharper problem is not the weights but the coverage, measured per track in the released manifest and stated in §2.4: a capability whose datasets all lack a human anchor is absent from the macro entirely, and since the MELD and CaSiNo anchors were withdrawn as unsourced that is now every capability on every flagship track — 100% of Core, of Dynamics and of Gem, where the same counts read 59.1%, 87.5% and 60.4% while those two anchors stood. This is why the paper's primary reporting is

per dataset on the native metric (Table 2), and it is now the only reporting there is. An alternative weighting cannot repair a capability that is not in the mean at all; only a human anchor can, and one cannot be supplied by a model rater, because the quantity being anchored is what a person achieves.

S.11 What the normalized macro would have covered, and why it covers nothing

The human-normalized macro of §6 covered exactly two capability–source pairs on Core — State on MELD, 2,601 items, and Intent on CaSiNo, 467, together 3,068 of 7,500, 40.9% by membership — and it now covers none of them. Both anchors were traced and neither is sourced. MELD's 0.82 shipped under its own comment calling it a "provisional expert reference; replaced by the human-calibration subset"; that subset never existed, and the harness had no human-baseline plumbing, no response schema and no scorer, so there was never a path by which measured data could have replaced it. CaSiNo's 0.78 has no provenance anywhere in the project history, and the same file had already withdrawn a sibling anchor on the stated rule that a human anchor the project has not measured must not define a scale; the strategy head was missed in that pass and survived only by not being numerically conspicuous, which is not a provenance. Both are now None.

With them gone, no dataset in the build can enter a capability mean. The two anchors that remain are unusable for independent reasons: FANToM's covers 219 items of which 219 are quarantined, and SOTOPIA's sits on a capability this package never scores. On All the mechanical macro-absent share is therefore 4,668 of 19,107, 24.4%, and the other 75.6% must not be read as normalizable. What survives untouched is the apparatus: the N_d definition, the rule that an unanchored capability is absent rather than zero, and the statement of what would restore a mean — a human number measured on our items, our observation cut and our metric. Every dataset is still reported on its native metric against its own floor. A capability missing from the macro is not a capability scored at zero, and the two statements must not be swapped.

For MELD the withdrawal rests on a measurement, not on an absence of evidence. With the dialogue context restored, two expert raters still find 72.7% and 78.8% of items underdetermined from the text, and they agree on the ambiguity flag itself at $\kappa = 1.000$ and, separately, on the emotion label at $\kappa = 0.936$ (Appendix T.4.1). A text-only MELD ceiling is not a number we failed to locate; it is a number the stimulus does not determine. A text-only MELD estimate could be collected — more raters would narrow its interval, as for any sample — but it would estimate how one population of readers resolves an ambiguity the transcript leaves open, and it would move with the population rather than converge on a ceiling. That is why this build ships no human anchor for MELD. The same caution applies in the other direction: a human score published for an original dataset anchors our adapted task only when splits, questions, observation cut, modalities, shortcut filtering and metric all match, which for these sources they generally do not.

The pattern is the finding, not the two cells. Two further adapters outside the frozen build carry the same defect: a forecasting anchor of 0.79 labelled a "human forecaster reference from the calibration subset (provisional)", and a coordination anchor of 34.0 labelled a "provisional human-human median". Four anchors, all provisional, all referring to a calibration subset that was never built and for which no harness existed. Nothing this paper reports depends on the latter two, and they are named here because a benchmark that hands a reader a normalized scale owes an account of where its denominators came from — which, in this project, was nowhere.

S.12 Single-turn forecasting horizon

Dynamics targets sit a short distance past the observation cut. We do not evaluate long-horizon social forecasting, multi-step rollouts, or the compounding of prediction error.

S.13 What the compute cap cost, and what the expert packet covers

The sixth model of §6 could not be repeated because $176.88 of a $205 compute cap was already spent and the re-run projected $35.16 against $28.12 remaining; those four figures are the state of that cap on 2026-09-07 and are not the current budget. A second provider was attempted and its account was exhausted mid-sweep, so nothing from it is reported. Excluding a model that returned 39.2% model errors may itself select for API reliability rather than for capability, and the attempted run is kept in the record for that reason. The on-track measurement that this sweep is not has since been made, and it is a different run: the 1,401-item balanced Core sample (467 items per capability, drawn from the 7,500-item frozen Core track of build 139722f4), scored on each dataset's own native metric against that dataset's own floor and never pooled, 15 native per-dataset scores over 5 models, each against its own floor (Table 10). It is still unanchored, so it yields native scores and floors and no capability score and no macro, and it is reported only per dataset. The sweep of §6 was off-track by construction, not by sampling: its runner built its evaluation items live from the adapters instead of reading the frozen build, and cut each adapter to its alphabetically first 700 episode ids, unseeded and unstratified. Of the 2,595 episodes it scored, 1,390 are in frozen Core on the shipped build, 538 are in All but not Core, and 667 are in no released track at all. Those three counts are a property of the build rather than of the sweep — on the previous build the same 2,595 episodes split 1,333 / 556 / 706 — which is the same fact Appendix T.13 reports and a second reason this run set is not a sample of anything. The cut is worse than a random sample and invisible in every summary statistic those runs carry.

On the expert packet, State rests on MELD alone and Intent on CaSiNo alone (supply, not sampling), and Social-IQ 2.0 is excluded at ∼170 s per item of video. Its nominal ±8.3 percentage points was a worst-case binomial planning figure for 139 independent binary responses, computed before collection. It is not a confidence interval for the packet as designed — a stratified mixture of accuracy, macro-F1 and a 1–5 continuous scale — and not an interval for agreement on the double-rated items either. The achieved figure is ±7.9 points on 139 items, and it is quoted as imprecise every time it appears. The model-rater diagnostics run against the packet establish that the items can be read and answered by some respondent under the rubric; they establish neither correctness, nor human answerability, nor a human anchor, and two diagnostic sheets agreeing with each other is not evidence about either. That has not changed now that people have rated it: a human number is comparable to a model number only if both are taken on the same items, stimuli and metric, which is exactly what 79 of 140 items the shipped build does not serve — 73 absent from it and 6 in quarantined heads — denies us here.

This is also the positive reason no text-only MELD human anchor appears anywhere in this paper. The absence is not a gap in our literature search; we measured it, and the measurement is in Appendix T.4.1. A text-only human number on MELD can be estimated and more raters would narrow its interval, but what it estimates is how one particular set of readers resolves an ambiguity the stimulus does not resolve, so more raters move it toward the mean of whatever population they were drawn from rather than toward a ceiling — what is missing is a channel, not a sample. An anchor of that kind would be neither a ceiling nor a like-for-like comparison, and quoting one beside a model score would misstate the headroom in whichever direction the readers happened to lean.

S.14 Scope deferred from v1

Interventional prediction against randomized observed branches, new multimodal recording, multilingual and cross-cultural collection, a living leaderboard, and longitudinal interaction are out of scope for this release.

S.15 The expert reference, in full

The expert reference is two authors, and off the shipped stimulus. The packet is rated by two of the authors; 16 of its 140 items carry over to the shipped build unchanged. 79 items

the build does not serve, and 45 whose MELD stimulus the fix below changed, are excluded, so no human ceiling on the released track is measured (Appendix T.4).

S.16 One human-behaviour source, in full

CaSiNo is the only licence-cleared source in this build that reports outcomes from real people in real interaction, and it contributes one negotiation domain and one MTurk population, with no second leg for the observation cut (§T.6). Its 1,030-dialogue yield clears a construction-feasibility gate on dialogue count and is not a power calculation for anything; the head that ships realizes 792 raw and 396 independent pairs and fails that gate, which §3.2 reports and which is why the threshold was retired rather than the head. Both halves of a pair come from one dialogue, between one pair of people, in a single HIT, so a session-level error term sits inside every pair. On the 191 of 396 pairs whose two halves are about the same person (48.2%, §2.2) a person-level error term sits inside them too. The released data separates neither from social inference.

S.17 Dependence, selection, and the absence of randomized data

Dependence, selection, and no randomized data. Our items are not independent observations. One conversation supplies several cuts, both participants and one shared outcome, so every resampling must keep those groups together — the item count is not the sample size. Filtering items the trivial policies solve also changes the population described, and no collection assigned a move (§2.3).

S.18 Why pairing does not identify a mechanism

Pairing does not identify a mechanism. A paired correlation cannot say why a model behaves as it does, nor why published estimates vary. A correct static answer establishes only the fact tested; a wrong forecast may reflect untested facts, an unpredictable outcome or label error, and does not locate a deficient transition model. We report an association for the models evaluated, and no cause for it.

T Additional results

T.1 The capability macro, and why no track carries it

The capability macro reads each capability against its chance floor rather than against zero — and, on Gem, against the unselected head's floor, Gem being anti-selected so that its own floor is an artifact of that selection. What §2.4 measures about it is now total: the macro cannot be formed on any track. Withdrawing the MELD and CaSiNo anchors as unsourced (§6) left Core at 7,500 of 7,500 items in macro-absent capabilities, Gem at 500 of 500 and Dynamics at 1,600 of 1,600 — previously 59.1%, 60.4% and 87.5%. On All the mechanical figure is 4,668 of 19,107 (24.4%), and the remainder is not reportable either: it is anchored only by FANToM, whose 219 anchored items are 219 of 219 quarantined, and by SOTOPIA, which this package never scores.

We therefore report no capability macro in v1, on any track, for any model, and the table that would have carried it is cut rather than printed empty. Two things are missing and only one is a run. No model has been evaluated on the All track or on the Gem track at all — those two tracks are the cut table's columns — and even a completed sweep on both would fill nothing, because N_d needs a human anchor per dataset and no dataset now has a usable one. We keep the definition rather than the grid: the N_d rule, the exclusion policy, and the statement that what would restore a column is one measured human reference per dataset on the released stimulus, reported with its n, raters, agreement and modality. What this paper does report is per dataset on each dataset's native metric (Table 2), which needs no anchor; the per-model, per-source breakdown is Table 10 in Appendix J, per source and not per capability, because the capability column is the half of that absence no sweep repairs.

Model	CaSiNo strategy macro-F_1	MELD weighted F_1	OpenToM macro-F_1
Chance floor	0.2148	0.3100	0.2868
Claude Haiku 4.5	0.3695 (+0.1547)	0.6581 (+0.3481)	0.4192 (+0.1324)
Claude Sonnet 4.6	0.4310 (+0.2162)	0.5542 (+0.2442)	0.7745 (+0.4877)
Claude Sonnet 5	0.4574 (+0.2426)	0.5700 (+0.2600)	0.8073 (+0.5205)
Claude Opus 4.8	0.5548 (+0.3400)	0.5389 (+0.2289)	0.7802 (+0.4934)
Claude Opus 5	0.6495 (+0.4347)	0.6166 (+0.3066)	0.8482 (+0.5614)
Spread (best − worst)	0.2799	0.1192	0.4289

Table 10: Every model on every Core dataset, on that dataset's own metric, with the margin over that dataset's own floor in parentheses — the grid behind Table 2's range column, on the 1,401-item balanced Core sample (467 items per capability, drawn from the 7,500-item frozen Core track of build 139722f4), scored on each dataset's own native metric against that dataset's own floor and never pooled. Read down a column, never across a row: the three columns are three metrics on three label spaces, there is no row mean, and no normalized aggregate exists (§2.4). Bold is the best score in its own column and carries no claim about any other. Rows are in release order, not rank order, because the columns disagree about rank: 6 of the 10 model pairs are ordered differently by at least two of the three datasets. Every cell clears its own floor and every cell is a point estimate — no interval over the 467 items per dataset is computed in this release.

Table 2, in full. MELD and CaSiNo were anchored at 0.82 and 0.78 in the previous version of Table 2; both are withdrawn as unsourced (§6), and that withdrawal is what the † cells record. MELD's 0.82 was a placeholder never replaced and CaSiNo's 0.78 an undocumented constant; neither is substantiable as human performance on our adapted tasks. *SOTOPIA is anchored yet never scored: its native metric here is a rubric-term score, not the GPT-4 reward (Zhou et al., 2024), and every real configuration ships without a rubric, so the adapter contributes no score rather than a zero. Floors and counts in that table are read off the frozen build, which predates the withdrawal. On the ‡ rows no model has been run on All or on Gem either, so those dashes are not pending numbers; and every cell of the Core range column is a point estimate, no interval over them being computed in this release.

T.2 The per-model grid, and where OpenToM is already finished

Table 10 is Table 2's range column written out: every model against every Core dataset on that dataset's own metric, with the margin over its own floor. It is the same 15 measurements, not a second sweep.

One qualification the dataset-level number hides. OpenToM's best macro-F1 is 0.8482 against a 0.2868 floor, which reads as headroom; but that macro-F1 is taken over five question heads, and on the coarse first-order location head (138 of the 467 items) four of the five models score $F_1 \geq 0.99$ — 0.9917, 0.9762, 0.9916 and 0.9916, with Claude Haiku 4.5 the lone outlier at 0.4052. That head is finished. What keeps the dataset score off the ceiling is the attitude head (0.4729–0.6396) and the second-order multi-hop head (0.3595–0.8049). A v2 wanting headroom on OpenToM has to retire the saturated head rather than re-weight it, and the head scores are reported here so that decision is auditable.

T.3 Capability profiles

Figure 2 is designed to give each model's score on the four capabilities Core and Dynamics carry — State, Intent, ToM and Dynamics — rather than the two-axis projection of Figure 3. That is the view that shows whether a deficit is specific to Dynamics or part of a flat profile. A fifth capability, Interactive, is tagged in the build but is SOTOPIA alone and lives only on the All track, so the figure does not draw it. In v1 none of the four bars exists. ToM has no normalizable source on frozen Core — OpenToM alone, 4,432 of the track's 7,500 items,

Cut from v1: no bar is drawn, and no bar is zero. A capability score is a mean of human-normalized dataset scores, and no dataset on any track this release ships carries a measured human reference, so N_d is undefined and all four capabilities are absent. Filling these bars needs a measured human reference per dataset, which is an addition to the benchmark and not a further sweep.

Figure 2: Capability profile per model: the view that would show whether a model's deficit is specific to Dynamics or part of a flat profile. Cut from v1 on 2026-09-15: all four bars are absent, and absent is not zero. A capability score is a mean of human-normalized dataset scores, and no dataset frozen Core carries has a usable human anchor — State is MELD alone and Intent CaSiNo alone, whose anchors were withdrawn as unsourced (§2.4); ToM is OpenToM alone, which published none; and no model has been run on the Dynamics overlay. So N_d is undefined everywhere and the figure is a designed view with nothing to draw. The scores this paper does report are native per-dataset ones on a partial, unbalanced subset of the frozen Core track — 1,390 of 7,500 items (affect 697, intent 467, tom 226), not the 1,401-item balanced evaluation sample, never capability scores, and what the item set alone can do to them is in Appendix T.13. Filling these bars needs a measured human reference per dataset, not merely another sweep.

none anchored — the Dynamics overlay has not been run, and State and Intent lost their scale on 2026-09-14 when the MELD and CaSiNo anchors were withdrawn as unsourced (§2.4, §6). Every bar is absent rather than zero, and 7,500 of 7,500 items are now in a macro-absent capability, so the figure is a designed view with nothing yet to draw and is labelled as one. The scores in Table 14 are native per-dataset scores on a partial, unbalanced subset of the frozen Core track — 1,390 of 7,500 items (affect 697, intent 467, tom 226), not the 1,401-item balanced evaluation sample, and are not Core scores (Appendix T.13). Table 2 and Table 10 are on the balanced Core sample instead. Neither kind is a capability score, because no capability score exists on any track.

T.4 Expert reference

The expert packet was rated, and the table that would have set it against the models is cut (end of this subsection, where the reason is given). Any such comparison must report inter-rater agreement alongside the gap, a gap measured on items the raters themselves disagree about not being headroom. What follows is a small expert reference with n stated, not a human-performance study.

Two of the authors rated the packet. Each was assigned 110 rows and answered 109 and 108; 58 double-rated items are usable, and agreement over them is Krippendorff $\alpha = 0.803$ and Cohen's $\kappa = 0.802$, the raters differing on 11 of the 58. The headline is a mean item score of 0.581 [0.501, 0.659] over 139 items and 197 ratings — an achieved half-width of ±7.9 points, which is not a precise number and is never quoted as one. The 10-item calibration block is scored separately and excluded, measuring whether the instructions are clear rather than whether the items are. That reference is not measured on the shipped stimulus. Checked item by item against the released build, 79 of its 140 scored items are not served by the shipped build — 73 absent outright, 6 on quarantined heads — and of the 61 that survive, 45 are MELD items whose stimulus changed under the context fix below, leaving 16 of 140 carried over unchanged. No human ceiling on the released track is measured in this paper, and with the comparison table cut this paragraph is where the paper says so.

T.4.1 A defect found by audit, confirmed by the raters, repaired, and re-rated

The MELD adapter's stripped context (§6) was found by the mechanical trivial-policy audit, then independently confirmed by both raters on the v1 packet, where each flagged every MELD row on their own sheet — 48 of 48, on both sheets — and MELD supplied 30 of the packet's 56 candidate mis-golds. After the adapter was repaired we put the repaired head back to the same two raters, on 33 rows each of the rebuilt packet's 47 MELD items, and instructed them in writing to hold their flagging threshold fixed — neither more generous

because they knew the stimulus had been repaired nor harsher to prove the point — with both outcomes pre-committed, a fix that did not work being stated as such. Both sheets came back complete: 33 of 33 answered, none blank.

The repair moved the raters, and did not finish the job (Table 11). Flagging fell from every item to roughly three in four, mean confidence rose about a point on a five-point scale, and agreement with MELD's own label rose from 25 of 67 ratings to 29 of 66 — on a sample whose seven labels are balanced to a modal class of about 15% in both samples. Inter-rater agreement is reported separately for the two things the raters recorded, because they are different variables. On the emotion label the 19 items both raters saw give Cohen's $\kappa = 0.936$, the two differing on one item of nineteen. On the ambiguity flag — the binary decision of whether an item is underdetermined by the transcript — the same 19 items give Cohen's $\kappa = 1.000$: identical flag-or-no-flag choices on every one, with 15 flagged by both. The flag κ is the statistic that establishes that the residual flags are reliable; the label κ is a different measurement and says nothing about the flags. Neither is an improvement and neither is ever reported as one: the label κ on this head before the fix was already 0.914.

Meld State, same two raters	pre-fix (v1 packet)	context-fixed
rows rated per rater	35	33
answered	34 / 33	33 / 33
flagged ambiguous, rater A	35/35 (100.0%)	24/33 (72.7%)
flagged ambiguous, rater B	35/35 (100.0%)	26/33 (78.8%)
mean confidence (1–5), A / B	2.43 / 2.66	3.45 / 3.42
answers matching Meld's label	25/67 (37.3%)	29/66 (43.9%)
inter-rater agreement, on each variable:		
emotion label, Cohen's κ	0.914 (n=17)	0.936 (n=19)
ambiguity flag, Cohen's κ	undefined	1.000 (n=19)

Table 11: The MELD context fix, checked by the raters who confirmed the defect. Both columns are the same head, task and two raters, instructed in writing to hold their flagging threshold fixed and told that a fix which did not work was as useful a finding as one that did. The columns are different samples of items, not the same items re-rated (§T.4.1). The two κ rows measure different variables and neither is an improvement: the flag κ is the reliability check on the residual flags, the raters having made the same flag-or-no-flag choice on all 19 shared items, while the label κ is a separate statistic on which they already agreed at 0.914 before the fix. Pre-fix flag κ is undefined because both raters flagged every row, leaving no variance to agree about. Across all MELD rows of the v1 packet — the 35 State rows here plus 13 next-utterance rows — the pre-fix flag rate is 48/48 for both raters and mean confidence 2.06 / 2.42.

The comparison is between samples, not within items, and the text must never say otherwise. The rebuilt packet re-drew its MELD items and re-used the item-id namespace, so swb-e9-0005 names a different utterance in the two packets. Of the 47 MELD items in the re-rate, 23 utterances also appear among the v1 packet's 48, and only 16 share both id and utterance. Restricting to the utterances a rater personally saw twice gives the same direction on much smaller n: rater A, 17 items, 17/17 flagged before and 13/17 after, mean confidence $2.59 \rightarrow 3.41$; rater B, 14 items, 14/14 before and 11/14 after, $2.79 \rightarrow 3.43$. What is held fixed across the two columns of Table 11 is the head, the task, the two raters and the written flagging threshold — not the items. Why three quarters of the items stay flagged: the text channel underdetermines the label. Every one of the 50 flags — all 24 of rater A's and all 26 of rater B's — names at least one competing emotion that remains defensible on the repaired stimulus. Not one says the item is broken, and nine name the missing channel outright, in the word delivery:

> "Surprise is strongest, but anger or disgust also fits the emphatic rejection; text does not establish delivery."
> "Neutral recognition, anger, and disgust remain defensible without Brenda's delivery."
> "A cheerful greeting and a neutral greeting are both defensible."

MELD is a multimodal corpus and we serve it as transcript. Restoring the dialogue prefix repaired what context could repair; what is left is the audio and the video, which the released item does not carry. The one item of nineteen the raters answered differently is the third quote's: one read the greeting as joy, the other as neutral, and both had independently flagged exactly that contrast.

A second finding, about the label inventory rather than the stimulus. Both raters independently reached for an emotion the seven-way choice does not offer. Rater A, on a defensive self-justification: "embarrassment is suggested but is not an offered label." Rater B, on an apology: "Embarrassment is not an option"; and, on a bafflement item, "confusion has no separate option." Three further flags name embarrassment, confusion or disappointment as the reading that fits. That is a defect in the adapted task — MELD's label set, carried into our item, does not span what its own utterances express — and no amount of restored context repairs it. What this licenses, and what it forbids. It licenses the claim that the benchmark's own audit machinery found a real defect, that independent human judgement confirmed it, and that the repair is verified rather than asserted. It forbids a text-only MELD human ceiling, and is the reason this paper publishes no text-only MELD human anchor at all: two experts with the full dialogue restored still find about three quarters of the items underdetermined, agreeing perfectly on which ones ($\kappa = 1.000$ on the ambiguity flag, 19 of 19 shared items), and they match MELD's label on 43.9% of a label-balanced sample while agreeing with each other on the emotion label at $\kappa = 0.936$. Such a number can be estimated, and adding raters would shrink its interval in the ordinary way; what it cannot be is a ceiling. It would measure how one population of readers resolves an ambiguity the transcript does not resolve, and that quantity stays wherever those readers sit however many are polled, because what is missing is a channel and not a sample size. Any such estimate is reportable only with its rater population and protocol attached, and never as the score a model would have to reach. What this subsection does not report is an expert–model gap, and the reason is not that the rating is missing. It was collected: 139 items, two raters, 197 ratings, Krippendorff $\alpha = 0.803$, frozen on 2026-09-14 and quoted above. What cannot be built from it is the like-for-like comparison it was collected for, for two reasons, neither fixable in this version. The reference is on the wrong stimulus: only the 16 items noted above carry over unchanged, so a re-scoring restricted to items the released build serves would rest on sixteen items, which is not a number we would print as a reference. And there is nothing on the other side: no model has been run on the Gem or Dynamics subsets the packet covers. We therefore report the expert rating as a measurement of the items the raters actually saw, and report no human–model gap at all, rather than typeset a comparison whose two columns would be measured on different item sets. The one part of the reference that is on the shipped stimulus, the re-rated Meld head, is Table 11 above.

T.5 Degenerate controls

Degenerate controls bound how much of a reported score is attributable to the interaction context rather than to priors over the answer space. The no-history condition is the one that bounds it: a model scoring near its full-context value with the history removed is reading the answer space, not the conversation. The converse does not follow, so a small no-history drop bounds the context's contribution without showing that the task is unforecastable.

No model was run under any control condition, so this paper reports no control result. What the release carries — runs/e5_controls and runs/e5_controls_native — are degenerate baselines: random, majority, constant and ridge, on MELD and CaSiNo respectively. A baseline is not an ablation: it bounds what the answer space alone achieves, and neither run holds a model evaluated under any of the five ablations. The consequence is the one the Dynamics track most depends on: this paper does not establish that its Dynamics items are unanswerable from the final utterance alone. That is the track's definitional check, it was not run, and nothing here should be read as having passed it (§6).

The central relationship of §5.2 is therefore untested against both of its standing robustness checks. It has not been recomputed under the no-history control, nor with each source held out in turn (Appendix T.6). Neither check was run, neither is reported as passed, and the

relationship is stated throughout this paper at the strength the evidence that does exist supports, with both checks outstanding.

T.6 Generalization across sources

Because the benchmark is assembled from heterogeneous sources, a capability score could reflect whichever source dominates that capability rather than the capability itself; leave-one-dataset-out evaluation recomputes every score with each source held out in turn, and a result that does not survive removal of a source depends on that source. Surviving every fold is weaker than it sounds: it shows the result is not carried by any one of the sources we happen to have, not that it would hold on sources we do not.

No leave-one-dataset-out evaluation was run, and none is reported here. There is no fold artifact in this release: every score this paper gives is computed over the full set of sources, and no result has been shown to survive the removal of any one of them — so which source any given number depends on is, for v1, unknown. The per-model version of the same folds is not reported either, and no appendix holds a place for it. We keep the design and its reading, because both constrain what a reader should expect of the folds when they exist. Six source rows would not be six replications of the static/dynamics relationship. Dynamics is carried by two sources, CaSiNo (804 items) and OpenToM (596, derived from plot metadata rather than reported by a participant), so only those two folds could move S_{dyn} or the rank correlation at all, and only CaSiNo's removal takes away outcomes reported by the people who had the conversation. The other four vary the static side alone and hold S_{dyn} fixed by construction, so a small range down that column would follow from the benchmark's composition rather than from corpus-independence. The companion check — the no-history control of Appendix T.5 — is absent for its own reason, stated there.

T.7 Correlation and rank reversals, tabulated

Every coefficient in Table 12 is model-level: the unit of analysis is a model, n is five, and the pairing is a property of the stimuli the two scores were computed on, not of the correlation. Five points is the whole sample, so the table gives Pearson r, Spearman ρ and Kendall τ only beside bootstrap intervals — over items and, separately, over models, which answer different questions and differ enormously — and a leave-one-model-out range. It reports both halves' floors in the same block, the two coordinates not being on one scale (§5.2), and the static column also carries the copy policy, so a static score is read as a margin over what copying from the prompt achieves.

Whether a reversal survives requires joint uncertainty over the five-vector of scores on each half and a multiplicity correction over the ten pairs. Both are now computed, and no reversal survives. Each replicate draws one resample of the 396 pairs, shared across every model and both halves, and evaluates both legs of a pair on it, so the covariance between the legs is carried by the resample rather than assumed; the family corrected over is all 10 pairs, not the 2 that reversed, since restricting it to the observed reversals would be selection on the outcome. 2 pairs reverse as point estimates and zero survive Holm, the smallest adjusted p being 0.309. A permutation cross-check under a more permissive family — Holm over only the two observed reversals — returns the same count. Seed 20260913, 10,000 percentile resamples.

Figure 3 is that table drawn: each model one point, the static half on x and the forecasting half on y, with the $y = x$ line as a drawing aid and no zero point, since the two axes have different floors.

Figure 4 draws the same ten pairs as a plane: each point is one model pair, its static advantage against its forecasting advantage on the same paired episodes, so a reversal is a point in an off-sign quadrant. No marker is filled, because filling one would assert a correction that this release does not compute.

2 of the 10 pairs reverse as point estimates and zero survive correction, on the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152

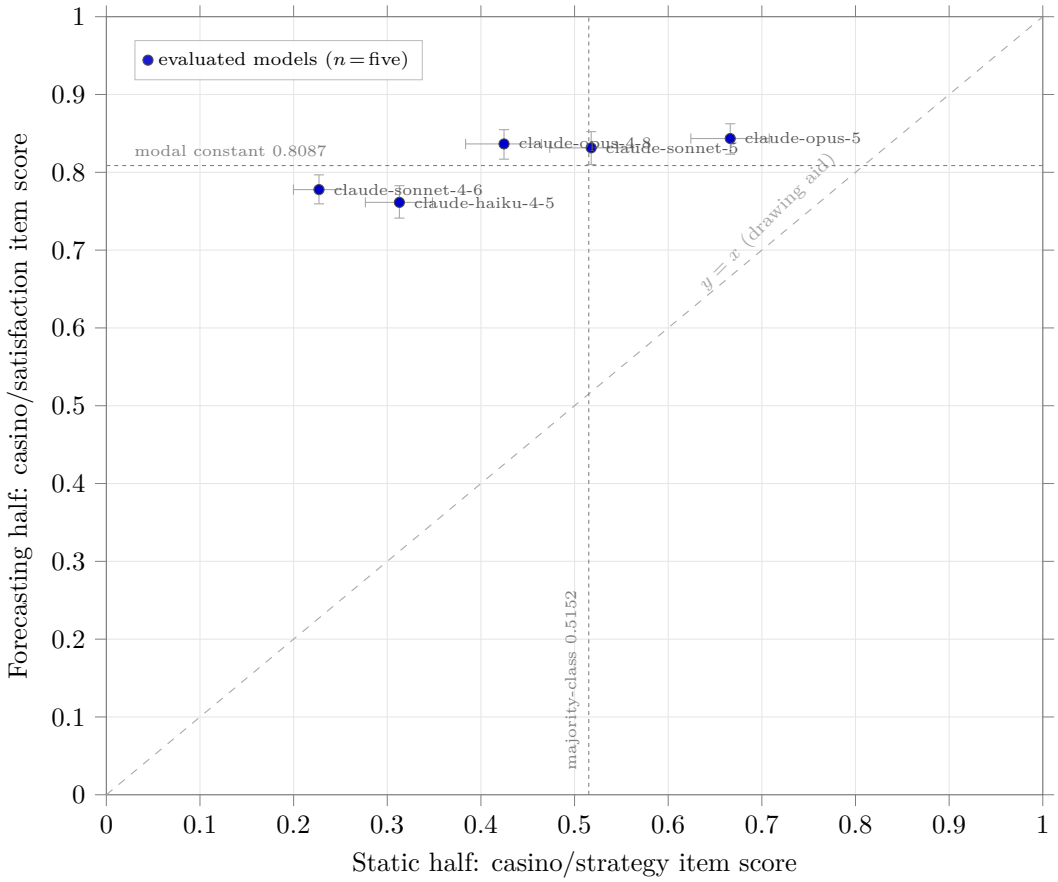


Figure 3: The two halves of the paired census, one point per model. The horizontal axis is the static casino/strategy half and the vertical axis the forecasting casino/satisfaction half — two heads of one corpus, not two capabilities — both scored on the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale. Bars are 95% percentile bootstrap intervals from 2,000 resamples over items (seed 0). The dotted lines mark each half's baseline — the best input-independent policy, which on the static half is the majority class non-strategic (51.52% of items) and on the forecasting half the modal answer, not a guessing rate in either case. Their separation is what makes the dashed $y = x$ line a drawing aid rather than a zero point: distance from the diagonal is not an amount of missing ability. The star marks the expert reference once that packet is rated. The plane is read for order along each axis, over $n = \text{five}$ points (§5.2).

and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale. Claude Haiku 4.5 is ahead of Claude Sonnet 4.6 on the static half and behind it on the forecasting half; Claude Opus 4.8 is behind Claude Sonnet 5 on the static half and ahead on the forecasting half. In both, the static leg is solid — $+0.0859$, CI $[+0.0499, +0.1239]$ and -0.0931, CI $[-0.1352, -0.0513]$ — and the forecasting leg is the one that fails: -0.0164, CI $[-0.0335, +0.0006]$ and $+0.0051$, CI $[-0.0076, +0.0177]$, both containing zero. Holm-adjusted p over the ten pairs is 0.309 and 1.000.

Over the same five models Spearman ρ is 0.8000, with $[0.600, 0.900]$ resampling the 396 conversations and $[0.111, 1.000]$ resampling the five models; Kendall τ over models is $[0.000, 1.000]$ and Pearson r $[0.517, 1.000]$. Refitting with each model held out in turn gives 0.6000–0.8000, exact over the five refits, with no fold changing the sign. The width of the model-level interval is the result, not a defect in it: the association is robust to which con-

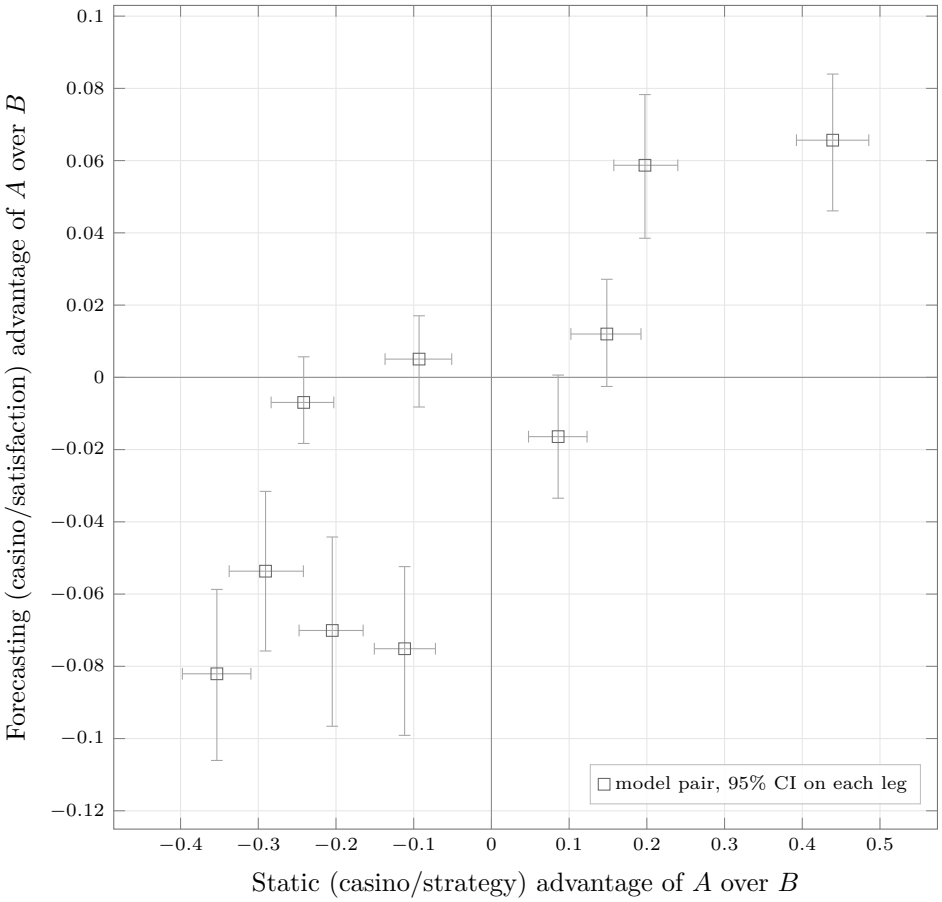


Figure 4: Why neither between-half reversal is supported. Each point is a pair of models (A, B), positioned by A's advantage over B on the static half and on the forecasting half, both computed on the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale, with a 95% interval on each leg. Off-sign quadrants are reversals: 2 of the 10 pairs lie in one, and zero survive Holm correction over all ten (smallest adjusted p 0.309; seed 20260913, 10,000 resamples of the 396 conversations, each replicate evaluating both legs of a pair on the same draw). The figure shows why: in both reversals the horizontal bar clears zero and the vertical bar crosses it, so it is the forecasting leg that fails. The eight remaining pairs show the same asymmetry in a milder form — separated on the static half, bunched on the forecasting one. No marker is filled, because none qualifies.

versations were sampled and is not determined by five models. Dropping Claude Opus 5 alone takes ρ from 0.8000 to 0.6000, which is the size of the whole effect.

T.8 Why the static half sits below its baseline, and what that is

The quantity three of five models score below on the static half is a majority-class baseline, not a chance floor. The gold is exactly non-strategic on 204 of the 396 items — 51.52% — and the baseline is the constant that answers it. "Chance floor" implies a guessing rate and invites the reading "worse than guessing", which is not what the number says.

It is not a set-size artifact. The suspicion is natural — gold sets average 1.09 labels while models emit 1.25 to 2.02, and the two lowest-scoring models are the two most verbose — and it does not survive three checks. A set-size-tolerant relaxation (Dice / example-based F1),

scored against its own exhaustively searched baseline, moves every model by +0.03 to +0.08 and flips no sign. Keeping only each model's first-listed label leaves the same three below. Decisively, an oracle ceiling that keeps the single best label out of the model's own emitted set — an unattainable upper bound, never a score — still leaves Claude Sonnet 4.6 at 0.4070 and Claude Haiku 4.5 at 0.4663, both under 0.5152. Claude Opus 4.8 is the one genuinely metric-sensitive case: 0.4247 as scored, 0.5400 under the oracle, so its below-baseline status is a set-size effect and is reported as such. Of the five set-overlap metrics we checked, only macro-F1 puts every model above baseline, and it reorders them: its own baseline is a far weaker 0.1666, and three labels with ≤1.5% support each carry a tenth of the score.

The behaviour behind it is nameable, and it is the same on both halves. Split the items by whether the gold is the majority set. On the 204 items where annotators recorded no negotiation strategy, Claude Sonnet 4.6 says so on 11.3% of them and Claude Opus 5 on 67.6%; on the other 192 every model scores between 0.357 and 0.661, so they are not confused about which strategies exist. Spearman between a model's static score and its majority-label recall across the five models is 1.000: the ordering on this half is essentially that one quantity. The over-attribution is not symmetric — the majority-label false-positive rate is 4–10% for every model, so they are not avoiding the label as a token but failing to select it when it is correct, which rules out a formatting explanation. Models over-attribute strategic intent, and the forecasting half is the same phenomenon: its gold is 4 or 5 on 338 of 396 items (85.4%), its baseline is the modal answer, and the two models below it under-predict satisfaction (mean 3.75 and 3.63 against a gold mean of 4.21). One behavioural claim, two measurements. Both strata correlate positively with the forecasting half across the five models, so the relationship of §5.2 is not an artifact of the non-strategic items.

Two caveats the elicitation earns. Our static query asks "which negotiation strategies does X's final turn use? List every one that applies." — an explicit invitation to a multi-label answer, against gold that is single-label on 92.7% of items, so some of the set-size inflation is the benchmark's own doing. And non-strategic is an absence marker rather than a strategy, marking an utterance the annotators gave no label, so a head whose majority class is "no annotation" measures annotation agreement as much as strategy recognition. Neither rescues the two models below their own oracle ceiling, and both belong in the record.

T.9 Sensitivity of the central result

Four reanalyses that would belong here are not computable on what this release ships, and we name them and their obstacles rather than list them as though they had been run. Alternative capability weightings need capability scores, and none exists on any track (§2.4). Aggregating within each source rather than across sources is not a reanalysis when the central result is computed on one source, the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale. Excluding contamination-flagged items presupposes a flag, and no evaluated model has one: the paraphrase and source-recognition probes are generated and unscored in this build, because scoring them needs model calls this freeze did not make. And excluding each Dynamics subtype in turn presupposes subtypes, where the forecasting half is a single head.

What is recomputed is the resampling unit, and it moves the answer: over the same five models, resampling the conversations gives $[0.600, 0.900]$ on ρ and resampling the models gives $[0.111, 1.000]$, with a leave-one-model-out range of 0.6000–0.8000. None of this addresses the two limits stated in §5.2 — the axes are not commensurable, and the ten model pairs are dependent — because a reanalysis of the same scores cannot supply a scale or an independent replicate.

T.10 Modality: where a masked-input comparison is defined

Modality. On the multimodal subset we compare full- with masked-modality input on the same items, so $\Delta_{\text{multimodal}}$ is not an item-set change. In v1 that subset is static-only: every item on the Dynamics track is text (§6), so $\Delta_{\text{multimodal}}$ is defined on the static half and is

not defined on Dynamics at all. No value is reported, because the comparison was never run. The text-only control is implemented and shipped in the harness, and no evaluated model has been run under it — nor under any other control condition (§4), which is why every floor in this paper comes from a deterministic policy rather than from a model with information removed. What would fill this is a paired run of the same items in full context and text-only, on the two sources that carry audio or video at all, MELD and Social-IQ 2.0. A second thing would have to change first, and it is not a budget problem. In this release the "full context" arm is itself text: no adapter passes media bytes to a model, every Episode.media value being a manifest key the prompt builder prints as a string, so MELD is served as its transcript and Social-IQ 2.0 as its question and four options with no transcript at all (Table 1, footnote ‖). Running the control today would compare text against text; measuring $\Delta_{\text{multimodal}}$ needs an operator's own copy of the footage, a model that can take it, and then the paired sweep. That is a further paid sweep, it is not scheduled for v1, and until it exists the size of $\Delta_{\text{multimodal}}$ is unknown on every dataset — so we cannot say how much of any score on the multimodal sources survives deleting the non-text channel. The masked-input comparison is part of what the benchmark is for (§6); it is not part of what this release measures.

T.11 Does the relationship depend on the two halves naming the same person?

The two halves of a CaSiNo pair name their participant independently: the static half asks about the speaker of the last strategy-annotated turn, the forecasting half about the surveyed participant. They coincide on 191 of the 396 pairs (48.2%), and a reviewer asked whether the relationship we report is an artefact of that overlap.

We re-aggregated the same stored predictions on the 191 same-referent pairs and, separately, on the other 205, using the same statistics, the same resample count and the same seed as the census fit. Two bootstraps run, as they do on the census. Each model's score on each half gets a percentile interval from resampling pairs; each subset's model-level coefficient gets one from resampling the five models, which is the interval reported below and the only one that carries the uncertainty of a five-point fit. The model-level relationship is unchanged: Spearman $\rho = 0.8000$ and Kendall $\tau_b = 0.6000$ on both subsets, against $\rho = 0.8000$ on all 396. Pearson r moves from 0.8637 on the same-referent pairs to 0.8206 on the different-referent pairs, inside intervals that overlap over nearly their whole length. Two of ten model pairs reverse between the halves in every subset, and none survives Holm in any of them. Table 13 gives the per-subset coefficients with their intervals.

This is a descriptive check, not a test. No statistic compares the two subsets and none was computed; two overlapping intervals are not evidence that two quantities are equal. Each subset's fit rests on the same five models from one provider as the census fit, so splitting 396 pairs in two adds no models and inherits every hedge §5.2 carries. What the check supports is the whole of what we claim from it: the relationship we report is not an artefact of whether the two halves are about the same person. It bounds that one confound and no other.

One thing the split exposes, which we report rather than absorb. The static measured floor is not the same on the two subsets: 0.4660 on the same-referent pairs against 0.5610 on the different-referent pairs, with the census floor at 0.5152. The label degeneracy the strategy head carries is therefore not spread evenly across the referent classes, so any floor-normalized version of these scores stands on a different scale in each subset. Table 13 prints the native scores only. They are never differenced across blocks here or anywhere else in this paper, and a reader should not difference them either.

T.12 The paired per-item difference, per model

Figure 5 reports the raw difference between the two halves' native scores, the offset of §5.2 one item at a time rather than one model at a time, with intervals from resampling the items. It differences two scores whose scales are not equated (§5.2), so it describes where the per-item offsets lie and does not measure how much less a model forecasts than it comprehends. The same caution applies as to the model-level displacement: no artifact in

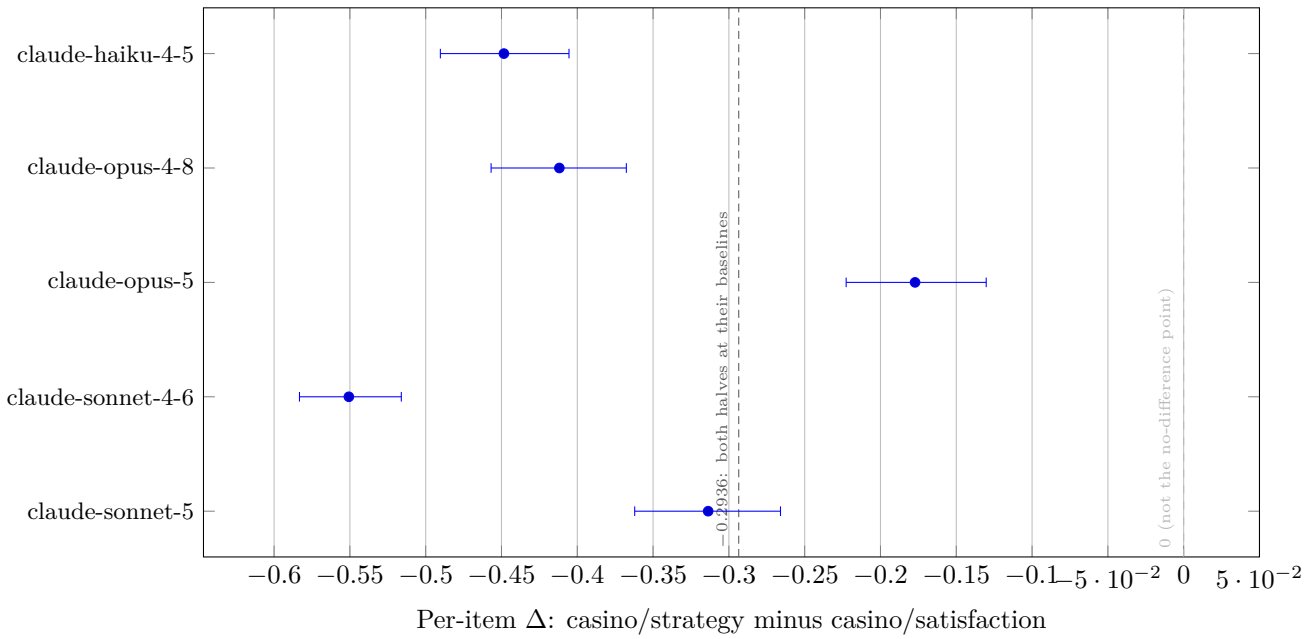


Figure 5: The per-item difference, per model — a description, not a deficit. Each point is a model's mean per-item difference between its static casino/strategy score and its forecasting casino/satisfaction score on the 396 independent CaSiNo pairs (792 episodes, one corpus): a static casino/strategy half and a forecasting casino/satisfaction half scored on the same stimuli, with chance floors 0.5152 and 0.8087 (floor gap −0.2936) and no artifact in this release putting them on a common scale; bars are 95% percentile bootstrap intervals from 2,000 resamples over items (seed 0), one index draw per resample shared across models. Zero is not the point of equal performance and Δ is not an amount of missing ability. The two halves have different chance baselines, so a model with no measured ability on either — sitting exactly at both baselines — already shows a Δ of −0.2936, drawn here as the second reference line. All five means are negative and every interval excludes zero, which the floors alone predict; neither fact is a comparison between the halves, and no artifact in this release puts them on a scale on which one could be made. The figure is read as a description of where each model's per-item offsets lie, and of how tightly they cluster.

this release puts the two halves on a common metric, so no version of the difference — raw, baseline-subtracted or control-adjusted — is interpretable as an amount of ability.

T.13 Item-set sensitivity: the same answers on two builds

Table 14 is the controlled comparison behind §3.4: nothing about the models varies in it. The responses were bought once, by a sweep that assembled its evaluation items live from the adapters instead of reading the frozen build and capped each adapter at its first 700 episodes by episode id — an unseeded, unstratified, alphabetical cut. Of the 2,595 episodes it scored, 667 are in no frozen track at all, its ToM score came entirely from a source the freeze quarantines (§6) and its Perception score from three synthetic fixtures; that sweep is reported nowhere. What is reported is its stored answers, replayed twice through the same prompt builder, parser and scorer, restricted each time to the episodes in frozen Core according to one build. No model was re-run and no scoring rule touched: the build is the only difference between the columns.

Every cell is native, and that is the point. No dataset on this track carries a usable human anchor, so no capability score and no macro mean exists on either build (§2.4) — not a low one, an absent one. What survives without an anchor is the native score, the adapter's own metric on its own items. Table 14 therefore holds three blocks on three metrics — weighted

F_1, strategy macro-F_1, macro-F_1 — and nothing in it may be read across the blocks. There is no aggregate row because there is no aggregate. This is a weaker object than a ranking of models and the only one the evidence supports, and it is sufficient for the claim being made, which is about the item set rather than about the models.

Two of the three orderings do not move at all. On CaSiNo and MELD the five models fall in the same order on both builds and not one of the ten ordered pairs reverses. On CaSiNo the scores do not move even in the fourth decimal — the build swap leaves its 467 items untouched, so the column is a literal identity and is printed as one, because a sensitivity analysis that reports only the case that moved is not a sensitivity analysis. On MELD all five scores rise slightly as the overlap grows from 622 items to 697, and the order is undisturbed. The reader who takes from this subsection that item lists always matter has taken the wrong lesson: on two of our three corpora, holding the build loose would have cost nothing.

On the third, the build swap alone reverses a pair. On OpenToM Claude Sonnet 4.6 leads Claude Opus 5 by 0.0729 native macro-F_1 on the earlier build's overlap and trails it by 0.0091 on this one. Nothing about either model changed between those two numbers; the second build quarantines opentom/location_fo_fine and opentom/location_so_fine (§3.3) and the pair's order follows the quarantine. We do not report that reversal as significant and no artifact in this release would let us. Both separations are small against the spread of the column, neither carries an interval, and the ordered pair is one of ten. What the comparison licenses is the weaker and more useful statement: which of these two models ranks higher was a property of which items were counted, not of the models — so a ranking published against an item list that anyone may re-derive is a ranking that can flip without notice.

Two per-model pairs, and the verification split. On the same items, the same stored answers score 0.3455 on the earlier build and 0.4595 on this one for Claude Haiku 4.5, and 0.6592 and 0.8481 for Claude Sonnet 5, because this build quarantines the two open-ended OpenToM location heads (§3.3). Separately, the 1,390 items on which this build passes prompt verification (§3.4) split affect 697, intent 467, tom 226.

The absolute scores move much further than the ordering does. On the same OpenToM items, the same stored answers score 0.5731 on the earlier build and 0.8119 on this one for Claude Opus 5 — +0.2388 absolute, with the model, the prompt, the parser and the scorer all held fixed. Every one of the five models moves in the same direction and by at least +0.11. That is what a head quarantine costs when it is applied to a set someone has already scored, and it is a larger effect than any difference this paper reports between models on that dataset. A native score is not interpretable without the build it was computed on.

The earlier column cannot be rebuilt, which is the sharpest form of the argument. Re-aggregating the same responses against the earlier build today fails prompt verification on 622 of its 1,333 overlapping items. All are MELD items — exactly the count of MELD items in that overlap — and they fail because the stimulus the models were shown predates the MELD context fix (§6), so the prompt today's adapter builds is not the prompt that was bought. This build verifies clean on all 1,390 of its items (affect 697, intent 467, tom 226). The left column of Table 14 is therefore a stored artifact the released code cannot regenerate, and is printed with its build's checksum in the column head for that reason: a number whose item list is not pinned is a number nobody can return to.

What none of this is. Neither column is a benchmark result and neither is reported as one anywhere else in this paper. Each is a partial, unbalanced subset of the frozen Core track — 1,390 of 7,500 items (affect 697, intent 467, tom 226), not the 1,401-item balanced evaluation sample; the earlier one restricts the same responses to 1,333 items. Both inherit the off-track sweep's alphabetical episode cut, so neither is a sample of anything, and both score MELD on the stripped-context stimulus the re-freeze replaced. Nothing in either column survives as a State, Intent or ToM number for the shipped benchmark. The measurement

that supplies those is a different run: the balanced evaluation sample, drawn per capability from the frozen build with its manifest checksum verified and its seed recorded. That run exists. On the 1,401-item balanced Core sample (467 items per capability, drawn from the 7,500-item frozen Core track of build 139722f4), scored on each dataset's own native metric against that dataset's own floor and never pooled, 5 models were scored across 3 datasets — 15 cells in all — and every cell clears its own floor: CaSiNo 0.3695–0.6495 on a 0.2148 floor, MELD 0.5389–0.6581 on 0.3100, OpenToM 0.4192–0.8482 on 0.2868. The grid is Table 10 and the reading of it is Appendix T.2. What the run does not yield is any aggregate: every dataset in it is unanchored, human is null on all 15 dataset scores, so no capability score and no macro mean follows from it and none is printed.

T.14 Action-conditioned track, full results

Table 15: Action-conditioned forecasting on the observational negotiation track. No language-model forecaster was ever run on this track, and none of this table is a measurement of one. No API key was configured for the action-conditioned experiment and the track carries no language-model configuration at all: every row below is a learned (ridge) or trivial (constant, action-marginal) predictor fit on the training split, so this table — and every action-conditioned result this paper reports from it — is a comparison among learned and trivial predictors only, and says nothing about the five models evaluated in §5.1. A Prompted block reserved on each head in earlier drafts is cut rather than printed empty (2026-09-15): no run behind it existed and none was possible before the freeze. The learned and control configurations had 0 parse failures in 8 of 8. Its item counts are a different census from Table 1's and are not a subset of them: the 440 and 395 items below are the casino_acsf adapter's own held-out split from the ACSF run, whereas Table 1 reports the assembled Dynamics track, in which casino_acsf contributes 804 of 1,600 items. Neither number is wrong and neither should be summed into the other. The two heads are reported in separate panels because they are separate measurements: Head A scores the end-of-negotiation Likert outcomes of one set of held-out items with Lin's concordance (a constant predictor scores 0), Head B scores the partner's next action on a disjoint set of items with the Brier score in sum form, $\mathrm{BS} = \sum_l (p_l - g_l)^2 \in [0, 2]$, over the fixed 14-label candidate set, where a uniform forecast scores $1 - 1/14$. The two are never averaged. In the contrast panel both interval endpoints are signed: an interval running from a negative to a positive endpoint is a null, and is labelled as one. Its third column reads the interval against zero in words and then gives the number of bootstrap resamples out of 10,000 in which the first arm of the contrast was ahead, so a null can be seen to be a null rather than inferred from a small point estimate. Every interval here resamples whole dialogues, not items, the 835 held-out items coming from only 41 of them; the item bootstrap is not reported, and understates Head A by 2.64×. Two rows exclude zero item-wise and span it here. The first contrast group decides the track: the full model against a separately trained action-only predictor, which asks how much of each head needs no conversation at all — a different question from the @ no history rows, which mask an input the model was trained with, and answered differently on Head B. Mean MAE is shown beside CCC and ranks the forecasters in the opposite order — shrinking a prediction toward the training mean lowers absolute error while destroying concordance. CCC is the head's pre-registered metric, but MAE is not decoration: it is the metric the learned arms lose on, and the two answer different questions. CCC asks whether a forecast moves with the item, MAE asks how close it lands on average, and Appendix T.14 states what each is worth here in Likert points. deal_accuracy is excluded: the held-out deal rate makes it return the label prior for every configuration, including the constant predictor.

Head A — final outcome, 440 held-out items: Lin's CCC over the four Likert outcomes (higher is better)

Forecaster	CCC [95% CI]	mean MAE
Controls		
Constant (training means)	0.000 [-0.000, 0.000]	0.824
Action only, $E[Y \mid a]$	0.014 [-0.016, 0.044]	0.835
Learned		
History only (action-free)	0.111 [0.006, 0.225]	0.861
History + action	0.115 [0.008, 0.230]	0.858
@ shuffled action	0.106 [0.006, 0.219]	0.860
@ no history	0.081 [0.006, 0.172]	0.797
Action only, ridge (no history)	0.016 [-0.016, 0.050]	0.834
+ history × action interaction	0.144 [0.049, 0.244]	0.845

(continued on the next page)

Table 15, continued from the previous page.

Head B — partner's next action, 395 held-out items disjoint from Head A: Brier over the 14-way action distribution, sum form, range $[0, 2]$ (lower is better)

Forecaster	Brier [95% CI]	Top-1 [95% CI]
Controls		
Constant (training means)	0.817 [0.793, 0.841]	0.334 [0.280, 0.389]
Action only, $E[Y \mid a]$	0.739 [0.707, 0.770]	0.425 [0.381, 0.472]
Learned		
History only (action-free)	0.786 [0.760, 0.813]	0.380 [0.332, 0.427]
History + action	0.733 [0.707, 0.757]	0.428 [0.389, 0.470]
@ shuffled action	0.819 [0.792, 0.848]	0.329 [0.281, 0.374]
@ no history	0.756 [0.728, 0.784]	0.425 [0.382, 0.472]
Action only, ridge (no history)	0.739 [0.707, 0.770]	0.425 [0.381, 0.472]
+ history × action interaction	0.730 [0.703, 0.758]	0.428 [0.388, 0.469]

Contrasts — paired bootstrap over whole dialogues (cluster bootstrap), 10,000 resamples, the same resampled dialogues scoring both arms

Contrast	Δ [95% CI]	vs. 0; resamples favouring
History + action − action only, ridge (how much is not solvable from the action alone?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0987 [+0.0080, +0.1958]	excludes 0; 9845/10000
Head B, next action (ΔBrier, better < 0)	-0.0064 [-0.0229, +0.0116]	spans 0 — null; 7704/10000
History + action − history only (does the action add information?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0036 [-0.0034, +0.0115]	spans 0 — null; 8297/10000
Head B, next action (ΔBrier, better < 0)	-0.0537 [-0.0711, -0.0378]	excludes 0; 10000/10000
History + action − @ shuffled action (does misaligning it remove the gain?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0090 [-0.0040, +0.0214]	spans 0 — null; 9067/10000
Head B, next action (ΔBrier, better < 0)	-0.0865 [-0.1111, -0.0636]	excludes 0; 10000/10000
History + action − @ no history (does the history matter?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0341 [-0.0189, +0.0922]	spans 0 — null; 8837/10000
Head B, next action (ΔBrier, better < 0)	-0.0237 [-0.0395, -0.0081]	excludes 0; 9988/10000
History + action − action only, $E[Y \mid a]$ (same question, non-parametric transition prior)		
Head A, final outcome (ΔCCC, better > 0)	+0.1008 [+0.0087, +0.1996]	excludes 0; 9846/10000
Head B, next action (ΔBrier, better < 0)	-0.0066 [-0.0235, +0.0118]	spans 0 — null; 7710/10000
History + action − constant (is anything learned at all?)		
Head A, final outcome (ΔCCC, better > 0)	+0.1150 [+0.0077, +0.2304]	excludes 0; 9835/10000
Head B, next action (ΔBrier, better < 0)	-0.0840 [-0.1041, -0.0649]	excludes 0; 10000/10000
History only − action only, ridge (is the conversation worth more than the transition prior?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0951 [+0.0053, +0.1919]	excludes 0; 9815/10000
Head B, next action (ΔBrier, better < 0)	+0.0473 [+0.0170, +0.0800]	excludes 0; 5/10000
@ no history − action only, ridge (masking an input is not the same as training without it)		
Head A, final outcome (ΔCCC, better > 0)	+0.0646 [+0.0024, +0.1381]	excludes 0; 9795/10000
Head B, next action (ΔBrier, better < 0)	+0.0173 [-0.0010, +0.0358]	spans 0 — null; 326/10000
History × action − action only, ridge		
Head A, final outcome (ΔCCC, better > 0)	+0.1280 [+0.0466, +0.2104]	excludes 0; 9993/10000
Head B, next action (ΔBrier, better < 0)	-0.0087 [-0.0272, +0.0107]	spans 0 — null; 8174/10000
History × action − history + action (can an interaction the additive model cannot represent be found?)		
Head A, final outcome (ΔCCC, better > 0)	+0.0292 [+0.0055, +0.0514]	excludes 0; 9952/10000
Head B, next action (ΔBrier, better < 0)	-0.0023 [-0.0100, +0.0054]	spans 0 — null; 7176/10000
History × action − constant		
Head A, final outcome (ΔCCC, better > 0)	+0.1443 [+0.0489, +0.2445]	excludes 0; 9987/10000
Head B, next action (ΔBrier, better < 0)	-0.0863 [-0.1069, -0.0660]	excludes 0; 10000/10000

Excluded as degenerate: deal_accuracy is 0.9750 for all 8 configurations, the constant predictor included, against a base rate of 0.9750. It returns the label prior, not a forecast, and is not an outcome of this experiment.

Why a trained baseline, and two comparisons that are routinely confused. The comparison that decides this track is the full forecaster against a separately trained action-only predictor, not the same model with its history masked. @ no history is the action-conditioned model evaluated with its history features zeroed, though it was trained with them; it asks how sensitive a trained model is to losing an input. Action only, ridge is a separate model fit

from scratch on the action alone, and asks how much of the head is solvable without the conversation. They do not agree: on Head B the masked model is +0.0173 Brier worse than the trained action-only baseline $[-0.0010, +0.0358]$, spanning zero, while the full model beats the masked one by -0.0237 $[-0.0395, -0.0081]$. Reading the second as evidence for the first is how a next-move headline survives that should not.

The resampling unit, and why it is the dialogue. The 835 held-out items come from 41 dialogues. Items inside one share a history, a pair of participants and — on Head A — a single post-negotiation survey repeated across every turn, so they are not independent draws. A paired item bootstrap removes the between-arm variance but not that dependence, and understates every interval it touches, so every interval in this paper resamples whole dialogues. The item intervals are retained in the artifact solely to quantify the understatement: on Head A the clustered interval is 2.64× wider on the headline ΔCCC, on Head B 1.04×, the per-item Brier being essentially uncorrelated within a dialogue. The effective sample behind Head A is nearer 41 than 440, which is the single largest reason the outcome results below are stated as directional rather than as estimates. Two of the 22 contrast rows the table renders change verdict with the unit: Head A's ΔCCC against @ no history ($[+0.0080, +0.0601]$ item, $[-0.0189, +0.0922]$ clustered) and Head B's ΔBrier for @ no history against the trained action-only baseline ($[+0.0009, +0.0336]$ item, $[-0.0010, +0.0358]$ clustered). Counting the MAE and RPS context metrics as well, 12 of 66 change, all on Head A.

Head B: the next-move result does not survive a trained baseline. Against history only, adding the action is worth ΔBrier -0.0537 $[-0.0711, -0.0378]$, 10,000/10,000 resamples favouring the full model, and against the shuffled-action arm -0.0865 $[-0.1111, -0.0636]$: the gain is specific to the action that actually occurred and is not a marginal effect of having any action in the feature vector. Misaligning the action therefore costs more than deleting it. Against a separately trained action-only predictor, however, the difference is -0.0064 $[-0.0229, +0.0116]$, 7,704/10,000 — no statistically detectable difference. The interval spans zero: it is consistent with the full forecaster being up to 0.0229 Brier better and with its being up to 0.0116 worse, and it does not establish that the two predictors are equivalent, no equivalence test having been run and no margin of practical equivalence justified. Descriptively, and as a ratio of point estimates carrying no interval of its own, a forecaster that never sees one word of the conversation reaches 92% of the full model's Brier gain over the constant predictor (Brier 0.8167 falling to 0.7391 against 0.7327) and 97% of its top-1 gain (0.3342 rising to 0.4253 against 0.4278). What we are entitled to say is narrow and we state it narrowly: among the predictors we fitted, on this held-out split, giving the forecaster the conversation bought no measurable advantage over giving it the move alone. We are not entitled to say that the conversation carries no information about the next move, nor that this head measures action-transition frequency instead of social forecasting: every arm here is one ridge family over hashed features, a restricted model class that fails to extract information is not evidence that the information is absent, and no language-model forecaster was ever run on this head. What the result does do is put the burden of proof where it belongs — a head on which a trained no-conversation baseline is not measurably worse than the full model should not be used to claim social prediction without first beating that baseline.

Head A: the outcome null is a null about additive models. The action-conditioned ridge is additive in hashed history features and a one-hot action, so by construction it cannot represent a history-by-action interaction; the artifact says so in its own null_scope_caveat. The additive null is real — adding the action to a history-only model is worth +0.0036 $[-0.0034, +0.0115]$, 8,297/10,000 — but it is a statement about that model class and must not be reported as "conditioning on the action does not help the outcome". An exploratory interaction-capable arm, over the same solver, features and split, beats the additive one by +0.0292 CCC $[+0.0055, +0.0514]$, 9,952/10,000. That contrast is outside the corrected inferential family below and is not claimed as a result. It is reported for one purpose: to show that the additive null is a null about additive models. It is not evidence that a history-by-action interaction exists, and it is not corrected for the number of arms the artifact scores. What the conversation buys on this head is larger still: against the trained action-only

baseline, +0.0987 [+0.0080, +0.1958], 9,845/10,000. Head A and Head B thus point in opposite directions — the outcome result favours the conversation and not (additively) the action, the next-move result favours the action with no measurable contribution from the conversation — and neither interval establishes equivalence in the direction it fails to reject, which is what the artifact's null_scope_caveat also says. Both statements are limited to the predictors we fitted and to this split.

Multiplicity. The primary comparison is the full model against the trained action-only baseline, on two heads. Taking each head's two-sided bootstrap p as twice the smaller of s and $1 - s$, for s its share of resamples favouring the full model, gives $p = 0.031$ on the outcome and $p = 0.459$ on the next move; Holm over the two returns 0.062 and 0.459. The primary comparison does not reach significance after correction, and the outcome head's single-head interval excluding zero is reported in that light. No other family in this table is corrected, because no other family is claimed; in particular the history-by-action interaction contrast (+0.0292 CCC) is exploratory, uncorrected, and nothing in this paper rests on it.

What is not reported, and why. Mean MAE appears on Head A for context and ranks the forecasters in the opposite order to CCC. Neither metric overrides the other and we let neither do so silently, because they answer different questions about the same forecasts. The constant predictor scores CCC 0.0000 at MAE 0.8239; the full action-conditioned ridge scores CCC 0.1150 at MAE 0.8578. Practically: if what is wanted is the single number closest on average to the rating a participant will give, the constant predictor is the better choice, by 0.034 of a point on the source's 1–5 scale. If what is wanted is a forecast that moves with the item, only the learned arm does that at all, and weakly, at a Pearson correlation of 0.181 against a gold standard deviation of 1.063. The mechanism is in the artifact's outcome_metric_caveat: the learned arms put real spread on their forecasts (predicted SD 0.389 against the constant predictor's 0.000) at a weak correlation, which concordance rewards and absolute error punishes relative to sitting on the training mean. A reader choosing a forecaster for a decision should choose by which of those two properties the decision needs. is not usable: the arms emit point values, and under the deterministic two-point decoding RPS is exactly MAE minus $f(1 - f)$ for f the fractional part of the forecast, verified at run time to 4.4×10^{-16}, so its apparent ordering is an artifact of where each arm's forecasts sit between integers. deal_accuracy is excluded as degenerate: 97.5% of held-out dialogues reach a deal, so every configuration including the constant predictor scores the label prior. The estimand is item-pooled, so longer dialogues carry more weight; dialogue-equally-weighted variants and a dialogue-level CCC are recorded in the artifact and change no conclusion above.

The four outcomes separately, and the two that are excluded. The item counts, once, explicitly. The held-out split is the corpus's own and runs by dialogue: 41 annotated test dialogues yield 835 items, dividing into 440 final-outcome items (Head A) and 395 next-move items (Head B) on disjoint item sets (Appendix P.1). Each Head A item carries a six-outcome label vector, of which four Likert outcomes enter the head's score — each party's own satisfaction and each party's liking of the other — so the CCC is a mean of four per-outcome CCCs, each computed over all 440 items, and the head rests on $440 \times 4 = 1{,}760$ scored scalars. There is no smaller item count anywhere in this head. The full model scores 0.102 on the speaker's satisfaction, 0.123 on the partner's, 0.112 on the speaker's liking of the partner and 0.123 on the partner's liking of the speaker; the interaction arm improves all four (0.129, 0.142, 0.147, 0.159) and the trained action-only baseline reaches 0.009–0.027. No single outcome carries the head. The artifact scores two further outcomes, each party's point total, which are excluded from the head by construction: they are not Likert items, sit on a different scale (mean absolute error ≈ 2.8 against ≈ 0.8), and every arm is close to the constant predictor on them — 0.033 and 0.033 for the full model, -0.012 and -0.003 for the action-only baseline. They are reported here and nowhere else, so that the exclusion is visible rather than silent.

No prompted language-model forecaster was run on either head, and none is reported. There was no API key at the freeze, so the track is learned and trivial configurations only; all eight of them parsed every held-out item, 0 parse failures in 8 of 8 configurations. The comparison

a reader might want — a prompted forecaster against these baselines — is absent from v1 and named as absent.

	Static (casino/strategy)	Forecast (casino/satisfaction)
(a) The two halves, each read only down its own column		
Measured floor	0.5152	0.8087
floor policy	{non-strategic}	median (4)
Copy-from-prefix policy	0.0535	not applicable†
margin over floor	−0.4617	n/a
Claude Haiku 4.5	0.3130	0.7614
floor-normalized	−0.4170	−0.2475
Claude Sonnet 4.6	0.2271	0.7778
floor-normalized	−0.5941	−0.1617
Claude Sonnet 5	0.5178	0.8314
floor-normalized	+0.0054	+0.1188
Claude Opus 4.8	0.4247	0.8365
floor-normalized	−0.1866	+0.1452
Claude Opus 5	0.6662	0.8434
floor-normalized	+0.3116	+0.1815
Range over five models	0.4391	0.0821
floor-normalized	0.9057	0.4290
(b) The relationship: one (x, y) per model, $n = \text{five}$		
Pearson r		0.8431
95% bootstrap CI		[0.517, 1.000]
Spearman ρ		0.8000
95% bootstrap CI		[0.111, 1.000]
leave-one-model-out range		0.6000–0.8000
Kendall τ		0.6000
95% bootstrap CI		[0.000, 1.000]
(c) Rank reversals: point estimates over ten dependent pairs		
Model pairs compared		10
Pairs each model enters		4
Rank reversals (any)		2
Reversals surviving joint uncertainty + Holm	0 of 2 reversals survive Holm over all 10 model pairs	

†The copy-from-prefix family is undefined on a continuous target and the manifest records null there. A null is the audit declining to run, not a clean result, so no number is reported and none is pending.

Table 12: The static half against the forecasting half, on the 396 independent CaSiNo pairs both were measured on (792 episodes, one corpus; this is a paired-stimulus result, not a cross-source one). Panel (a) is read down each column and never across it: the two halves have different target types, different label cardinalities and floors −0.2936 apart, and no artifact in this release puts them on a common scale, so no static-minus-forecast quantity appears here or anywhere in the paper (§5.2). On the static half the reference point is the copy policy, not the floor: copying from the prompt scores below the floor, which is why this head carries the pairing. Panel (b) is model-level: what is paired is the stimulus, what is correlated is a model score, one (x, y) per model, so n is five – the 396 pairs set the precision of each coordinate, not the degrees of freedom of the coefficient. Five points make a coefficient one model can move, so each is reported only beside its interval and its leave-one-model-out range. Every interval in panel (b) is a percentile bootstrap over models — 10,000 resamples of the five models with replacement, seed 20260913, of which 9,986 were non-degenerate — and not over items: resampling the 396 pairs instead holds the model set fixed and yields a Spearman interval roughly three times narrower ($[0.600, 0.900]$), which is the precision of each coordinate and not of the coefficient. The width is the result. Five points cannot pin a correlation, and $[0.111, 1.000]$ on ρ is what they support; the leave-one-model-out range is reported beside it because at this n it is more legible than either bootstrap. Panel (c) is not a test. The ten pairs are functions of one five-vector of scores per axis, each model entering four of them. Joint uncertainty over both legs of each reversal within a single pair-preserving bootstrap replicate, Holm-corrected over all ten pairs — the family is every pair that could have reversed, not only the two that did, because correcting over the observed reversals alone is selection on the outcome — leaves zero of the two reversals standing, the smallest adjusted p being 0.309. No reversal in this paper is described as significant. The forecasting floor is the floor of the 396 pairs these runs scored. A floor that any input-independent policy reaches is a property of the item set it is measured on. The pair build computes each half's floor over all 792 pairs it holds, and every one of these five runs stamped that value. But a run scores only the 396 pairs that pass the independence gate, and on those 396 the best constant answer scores 0.8087, not 0.8065.

	Static casino/strategy		Forecast casino/satisfaction	
Model	score	95% CI	score	95% C
All pairs — n=396 pairs; measured floors 0.5152 (static) and 0.8087 (forecast)				
Claude Haiku 4.5	0.3130	[0.2767, 0.3500]	0.7614	[0.7412, 0.7809
Claude Sonnet 4.6	0.2271	[0.1984, 0.2559]	0.7778	[0.7588, 0.7961
Claude Sonnet 5	0.5178	[0.4739, 0.5617]	0.8314	[0.8106, 0.8523
Claude Opus 4.8	0.4247	[0.3834, 0.4665]	0.8365	[0.8169, 0.8554
Claude Opus 5	0.6662	[0.6248, 0.7069]	0.8434	[0.8239, 0.8624
Model-level fit over these n=5 points: $r = 0.8431$ [0.5166, 1.0000], $\rho = 0.8000$ [0.1111, 1.0000], $\tau_b = 0.6000$ [0.0000, 1.0000] 2 of 10 model pairs reverse between the halves; 0 survive Holm				
Same referent — n=191 pairs; measured floors 0.4660 (static) and 0.8102 (forecast)				
Claude Haiku 4.5	0.3235	[0.2711, 0.3776]	0.7565	[0.7277, 0.7866
Claude Sonnet 4.6	0.2438	[0.2011, 0.2869]	0.7709	[0.7435, 0.7984
Claude Sonnet 5	0.4948	[0.4311, 0.5576]	0.8154	[0.7840, 0.8469
Claude Opus 4.8	0.4420	[0.3826, 0.5022]	0.8338	[0.8050, 0.8613
Claude Opus 5	0.6444	[0.5846, 0.7029]	0.8416	[0.8115, 0.8704
Model-level fit over these n=5 points: $r = 0.8637$ [0.3612, 1.0000], $\rho = 0.8000$ [0.1111, 1.0000], $\tau_b = 0.6000$ [0.0000, 1.0000] 2 of 10 model pairs reverse between the halves; 0 survive Holm				
Different referent — n=205 pairs; measured floors 0.5610 (static) and 0.8073 (forecast)				
Claude Haiku 4.5	0.3032	[0.2541, 0.3529]	0.7659	[0.7390, 0.7927
Claude Sonnet 4.6	0.2115	[0.1746, 0.2500]	0.7841	[0.7585, 0.8098
Claude Sonnet 5	0.5392	[0.4764, 0.6000]	0.8463	[0.8171, 0.8732
Claude Opus 4.8	0.4085	[0.3492, 0.4659]	0.8390	[0.8134, 0.8646
Claude Opus 5	0.6866	[0.6276, 0.7435]	0.8451	[0.8195, 0.8695
Model-level fit over these n=5 points: $r = 0.8206$ [0.6146, 1.0000], $\rho = 0.8000$ [0.1111, 1.0000], $\tau_b = 0.6000$ [0.0000, 1.0000] 2 of 10 model pairs reverse between the halves; 0 survive Holm				
The split is 191 same-referent and 205 different-referent pairs (0.4823 of the census). The two subsets are shown SIDE BY SIDE and are not tested against each other; no difference test is computed and overlapping intervals are not evidence of equality. The fit on each subset has the same five models from one provider as the fit on the census.				

Table 13: Splitting the paired census by whether its two halves are about the same person. The same stored predictions as Table 12, re-aggregated on 191 pairs whose static anchor speaker is the surveyed participant and, separately, on the 205 pairs where it is not; the split is 48.2% same-referent. The two kinds of interval resample different things. Each model's two score intervals are percentile bootstraps over pairs, the cluster being the pair, since a pair's two halves share a prefix and a timestep. Each block's model-level fit interval is a percentile bootstrap over models, the five models drawn with replacement; that is the interval Table 12 prints, and the pair bootstrap is not an interval on a five-point fit because it holds the model set fixed. Both use the same machinery, resample count and seed as Table 12. This is a descriptive check, not a test: the two subsets are printed side by side, no statistic compares them, and overlapping intervals are not evidence of equality. Each subset's model-level fit rests on five models from one provider, so every limit the census coefficient carries carries here too. The two subsets have different measured floors on the static half, so no cell may be differenced across blocks (§5).

Model	Native score: earlier build 0b9e1d78	Native score: this build 139722f4	Δ	Rank: earlier build	Rank: this build
CaSiNo, strategy macro-F_1 — n=467 and n=467 of 7,500 frozen Core					
Claude Opus 5	0.6563	0.6563	+0.0000	1	1
Claude Opus 4.8	0.5351	0.5351	+0.0000	2	2
Claude Sonnet 5	0.4649	0.4649	+0.0000	3	3
Claude Sonnet 4.6	0.4489	0.4489	+0.0000	4	4
Claude Haiku 4.5	0.3650	0.3650	+0.0000	5	5
MELD, weighted F_1 — n=622 and n=697 of 7,500 frozen Core					
Claude Haiku 4.5	0.5939	0.5991	+0.0052	1	1
Claude Opus 5	0.5844	0.5907	+0.0063	2	2
Claude Sonnet 5	0.5770	0.5820	+0.0051	3	3
Claude Sonnet 4.6	0.5603	0.5701	+0.0099	4	4
Claude Opus 4.8	0.5522	0.5533	+0.0012	5	5
OpenToM, macro-F_1 — n=244 and n=226 of 7,500 frozen Core					
Claude Sonnet 5	0.6592	0.8481	+0.1890	1	1
Claude Opus 5†	0.5731	0.8119	+0.2388	3	2
Claude Sonnet 4.6†	0.6460	0.8028	+0.1568	2	3
Claude Opus 4.8	0.5201	0.7207	+0.2006	4	4
Claude Haiku 4.5	0.3455	0.4595	+0.1140	5	5

Table 14: The item set, not the models, decides one of the rankings. One set of stored model responses, bought once and never regenerated, replayed through the same prompt builder, parser and scorer against two builds of the benchmark; the build, and so the item list, is the only thing that differs. Every cell is a native per-dataset score — the adapter's own metric on its own items — so no human anchor enters this table and no aggregate is formed. The three blocks are three different metrics and nothing may be read across them. Neither column is a Core result: each is a partial, unbalanced subset of the frozen Core track — 1,390 of 7,500 items (affect 697, intent 467, tom 226), not the 1,401-item balanced evaluation sample, and the left one restricts the same responses to the earlier build's 1,333 overlapping items. On CaSiNo and MELD the ordering of the five models is identical on both builds and no pair reverses. On OpenToM one pair does (†): Claude Sonnet 4.6 leads Claude Opus 5 by 0.0729 on the earlier build and trails it by 0.0091 on this one, because this build quarantines the two open-ended location heads. The absolute scores move much further than the reversal does — up to +0.2388, for Claude Opus 5 — on answers that never changed. The left column additionally cannot be rebuilt: re-aggregating against the earlier build today fails prompt verification on 622 of its 1,333 items, every one a MELD item predating the context fix, where this build verifies clean on all 1,390 (§3.4, Appendix T.13).