Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Preliminary evaluation of a typed decision model for counselling dialogue coding using motivational interviewing

Elad Refoua
Bar-Ilan University
Mor Bar
Harvard Medical School
Shaked Seker
Bar-Ilan University
Hadar Dery
Trier University
Ziv Ben-Zion
University of Haifa
David Piterman
University of Haifa
Zohar Elyoseph
zohar.j.a@gmail.com

University of Haifa

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11596861/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Preliminary evaluation of a typed decision model for counselling dialogue coding using motivational interviewing

Elad Refoua[1,2], Mor Bar[3], Shaked Seker[1], Hadar Dery[4], Ziv Ben-Zion[5,6], David Piterman[2*], Zohar Elyoseph[2*]

1 Department of Psychology, Bar-Ilan University, Ramat-Gan, Israel

2 Department of Counseling and Human Development, University of Haifa, Haifa, Israel

3 Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

4 Clinical Psychology and Psychotherapy - Methods and Approaches, Department of Psychology, Trier University, Trier, Germany

5 School of Public Health, Faculty of Social Welfare and Health Sciences, University of Haifa, Haifa, Israel

6 Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA

* David Piterman and Zohar Elyoseph contributed equally to this work.

Corresponding author: Prof. Zohar Elyoseph, Faculty of Education, University of Haifa, 199 Aba-Hushi Avenue, Mount Carmel, Haifa 3498838, Israel. Email: zohar.j.a@gmail.com

Abstract

Typed decision models return probabilities over predefined options, potentially supporting automated coding of motivational interviewing transcripts. The study compared Jev 1.13 with ten experts and two large language models (LLMs), Claude Opus 5.5 and GPT-6 sol, on seven motivational interviewing demonstrations. LLMs reported confidence as probabilities. Jev showed high panel agreement, although Claude Opus 5.5 had higher point estimates on both tasks. Against the ten-expert plurality, its estimated calibration error was lower than both LLMs', clearly so only against Claude Opus 5.5. Calibration comparisons depended on the reference; a prespecified target was missed against single coders. Retained successful LLM calls cost 64-175 times as much as Jev's, with median response times 7-8 times as long. Costs and latencies were measured through one provider on one day. These preliminary findings motivate evaluation in independent clinical sessions and workflows combining automated coding with human review.

Introduction

Artificial intelligence has been studied in mental health since the 1960s[1], but large language models (LLMs) made it a mass phenomenon. In a U.S. survey, nearly half of LLM users with a mental health condition reported using one for psychological support[2], and commentators describe LLMs as de facto mental health providers[3]. Research has tested them on emotional awareness[4], suicide risk[5] and a widening range of clinical tasks such as detecting depression, supporting treatment and psychoeducation[6-10], and dedicated applications have already reached randomized trials[11].

One task has so far only partly yielded to automation: coding what is said in psychotherapy sessions - a core method for studying how therapy works and for checking that therapists deliver it as intended and to support clinical decisions in ongoing treatments[12-14]. Traditionally, trained raters manually assign each utterance to one category of a structured scheme. By hand it is slow: a proficient coder may need up to two hours to code a 20-minute motivational interviewing session[15], and thus most recordings of psychotherapy are never analysed[12]. Nevertheless, automating this process has proved difficult, for reasons specific to the nature of psychotherapist interactions. Codes follow the speaker's intention rather than the words alone: the same sentence can be a reflection or a question depending on its intonation[16], and whether a client's statement argues for or against change often depends on what was said earlier[17]. Therapy talk also follows its own conversational rules[18], so methods built for everyday or

commercial dialogue cannot simply be carried over. Experts themselves disagree, most on client codes[19], and supervised models trained on hundreds of coded sessions agreed well with human coders on common therapist codes but poorly on client codes[17]. The stakes are also high: psychotherapy is an unusually high-stakes domain for AI[6], in which a misread utterance can distort feedback to a therapist, a judgement of treatment quality or, in live use, the response to a person in distress. Whether LLMs can close this gap depends on how they reach their decisions.

LLMs were built to converse with people, which makes them attractive for psychotherapy research and practice, but two consequences limit them as decision-makers. First, they hallucinate. Pretraining and evaluation reward a confident guess over an admission of uncertainty, so even state-of-the-art LLMs produce plausible falsehoods[20,21], and any decision must still be parsed out of free text. Second, they are sycophantic. Tuned by reinforcement learning from human feedback (RLHF) to write the answers people prefer[22], they drift towards a user's stated view and endorse it, even when it is wrong[23]. In April 2025 OpenAI withdrew a ChatGPT update for "validating doubts, fueling anger, urging impulsive actions" and flagged the risk for mental health[24]; in clinical terms, sycophancy can mean affirming a delusion[25] or responding inappropriately to a person in crisis[26]. Their answers also shift with changes of prompt format that leave the meaning intact[27]. Nor can their confidence be trusted: stated in words, it is more generated text and tends towards overconfidence[28], and after preference tuning their token probabilities are poorly calibrated too[29]. Without calibrated confidence, in which answers given 90% confidence are right 90% of the time, a system cannot tell which decisions to act on and which to pass to a person[30].

Real-time use adds costs in time and money. Frontier models buy accuracy with reasoning at inference time[31], and some no longer let it be switched off: Claude Opus 5.5 always runs with adaptive thinking, deciding for itself how much to reason[32]. Independent API benchmarks report end-to-end response times extending to minutes for frontier models at high reasoning settings; these times include input processing, reasoning and answer generation[33,34]. Each call is also billed per token, at US$2-4 per million input tokens and five times as much per output token, reasoning included, for Claude Opus 5.5 and GPT-6 sol, the two frontier models we compared[32,35]. For a system that must decide on every utterance of every session as it happens, both costs multiply.

Typed decision models reverse each of these choices, and they have antecedents. In supervised classification, models such as BERT returned probabilities over a fixed set of labels after fine-tuning on labelled examples for the target task[36], and could be well calibrated within their training domain[37]; entailment-based classifiers, including BART-based implementations, scored any list of label descriptions without such examples[38,39,40]. Typed decision models join both properties to the language understanding of current large models, by their developer's account, and TypeSafe AI's Jev, released in September 2026 as the first 'System One' model, is the current expression of the technology[41]. Its release drew unusual attention: the developer reported clearing about 140,000 waitlist sign-ups within 36 hours[42], according to Firecrawl, opened access on 20 September and paused new sign-ups on 22 September, citing demand[43], and independent open models of the same shape appeared within days[44].

Such models are built for machines rather than people: software declares the possible answers in advance, as a closed list of options with their definitions, and the model returns a probability distribution over that list. Because it generates no text, it cannot hallucinate an answer outside the list, although it can still choose the wrong option, and there is nothing to parse. It is trained by reinforcement learning for calibrated decisions (RLCD), which rewards probabilities for matching outcomes where RLHF rewards answers for pleasing people; not trained to please, it should bend less to the user and to the prompt, and the developer states that it returns similar answers for similar inputs[41]. Its confidence is read off the returned distribution rather than written in words. And it samples all answers in parallel in a single query rather than token by token, with reported response times of 70-500 ms at US$0.042 per million input tokens and no charge for output. The developer claims accuracy comparable to LLMs, but only from its own evaluation, which scores the model against labels averaged from two LLMs[45].

Our aim was to test whether this class can serve as the decision layer of automated systems that code therapy sessions. Such systems must assign a code to every utterance, either after the session, for research and supervision, or, in clinical use, as it unfolds, and pass to a person only the codes they are unsure of. As a test case we used AnnoMI[19], one of the few high-quality open datasets in the field, in which experts in motivational interviewing (MI), a counselling approach that helps people strengthen their own motivation for change[16], coded every utterance of 133 demonstration sessions. MI coding follows a structured scheme, which makes it a good fit for this class of models: each utterance takes one code from a closed list, the codes relate to treatment outcomes[46], automated coding has supported real-time feedback during simulated

counselling practice[47], and a subset of seven sessions coded by all ten experts allows direct comparison with experts and LLMs.

The current study asked three questions: (1) Does the model agree with experts as well as frontier LLMs do? (2) Is its confidence calibrated well enough to decide which codes need human review? (3) Is it fast and cheap enough to code every utterance?

Results

Agreement with expert codes

The typed decision model showed high agreement with the expert panel. Its reported Cohen's κ is the mean of ten comparisons, each using the plurality code of nine experts as the reference. Because experts disagree on such codes[19,48], a coder that tracks the consensus is favoured when scored against it, so these comparisons describe agreement with the panel's modal codes and do not establish equivalence to expert coders. For therapist behaviour its κ was 0.849, 95% confidence interval (CI) [0.791, 0.900], inside the experts' range of 0.657–0.872 and 0.028 [−0.022, 0.076] above their mean of 0.821. For client talk type it was 0.749 [0.666, 0.824], 0.154 [0.081, 0.229] above the experts' mean of 0.595 (experts' range 0.474-0.738; Fig. 1a and Table 1). This exceeds every expert, a result the consensus reference would favour. Under the planned utterance bootstrap, both agreement-difference intervals between the typed model and Claude Opus 5.5 included zero and allowed Claude Opus 5.5 to agree better by up to 0.079 and 0.089. Against GPT-6 sol no difference was detected on therapist behaviour, and on client talk type the typed model was higher (0.136 [0.064, 0.213]). These analyses included all seven AnnoMI transcripts annotated by ten experts. The confidence intervals treat utterances as independent, although they are clustered within transcripts. Generalisation beyond these seven transcripts remains uncertain.

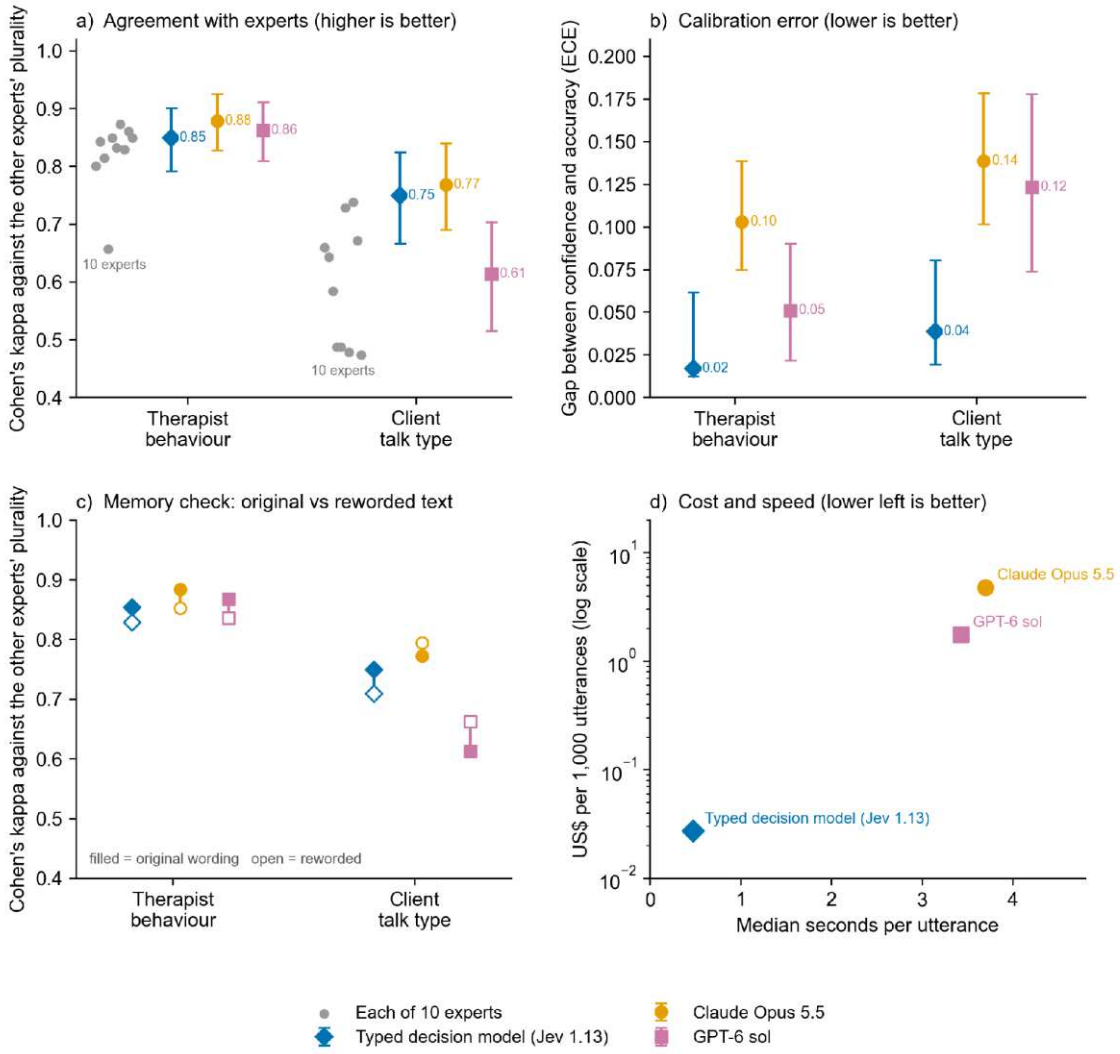


Fig. 1 | Agreement, calibration, rewording stability, cost and time for the typed decision model and two LLMs. a, Each expert's Cohen's κ against the plurality of the other nine experts, and each model's mean κ against the same ten references. The subset contains 428 utterances (216 therapist, 212 client); tied items are excluded separately for each reference. Small grey dots are the ten experts; diamonds are the typed decision model (Jev 1.13, TypeSafe AI); circles are Claude Opus 5.5; squares are GPT-6 sol. A coder that tracks the consensus is favoured when the consensus is the reference. **b**, Expected calibration error against the ten-expert plurality, on the 212 therapist and 202 client utterances without a tie; lower is better. For the LLMs, confidence is the probability the model stated for its label. **c**, Leave-one-out κ on the original utterances (filled markers) and on the corresponding versions from the rewording experiment (open markers), for the 214 therapist and 209 client utterances retained after checking. Claude models prepared and checked these versions, and no human checked them. **d**, US dollars per 1,000 utterances (log scale) against median seconds per utterance on the 428 ten-expert utterances, as billed and timed through OpenRouter on 23 September 2026. Error bars in **a** and **b** are 95% percentile bootstrap intervals over utterances (5,000 resamples), which treat utterances as independent.

Memorisation and rewording checks

Could this agreement come from memorisation? AnnoMI's codes have sat beside its transcripts in public since 2022[19], a known contamination risk[49]. Both memorisation checks had predictions committed before their runs. Their point estimates fell within the prespecified

comparison ranges and did not meet the operational memorisation predictions. On the 9,271 utterances coded by single experts, agreement with the published codes was $\kappa = 0.742$ [0.718, 0.765] for therapist behaviour and 0.509 [0.471, 0.547] for client talk type, inside the prespecified comparison ranges (0.67-0.87 and 0.46-0.66) and below the 0.90 predicted for memorisation. On all 4,817 client utterances, the set on which GPT-4.1 reached 0.51[50], it was 0.519 [0.481, 0.559]. On the 423 utterances retained after the rewording check[51], the typed model's leave-one-out κ decreased by 0.025 for therapist behaviour and 0.040 for client talk type (originals minus rewrites; Fig. 1c). The corresponding 95% utterance-bootstrap intervals were [−0.0001, 0.057] and [−0.004, 0.086]; transcript-bootstrap sensitivity intervals were [−0.010, 0.057] and [0.002, 0.099]. Only the client transcript-bootstrap interval excluded zero, so evidence of a decrease in client-code agreement depended on the resampling method. Both point estimates were below the prespecified 0.05 threshold, but all four intervals extended above it. These results do not establish that the decrease was 0.05 or less, and its size remains uncertain with only seven transcripts. The rewrites retained a mean Jaccard word overlap of 0.59 (Methods), and the checks do not rule out memorisation.

Calibration against expert references

Scored against the ten-expert plurality, the typed decision model's confidence in its chosen code tracked its accuracy. Its expected calibration error (ECE) was 0.017 [0.012, 0.061] for therapist behaviour and 0.038 [0.019, 0.080] for client talk type, against 0.103 and 0.139 for the probabilities Claude Opus 5.5 stated and 0.051 and 0.123 for those GPT-6 sol stated (Fig. 1b). Against Claude Opus 5.5 the difference excluded zero on both tasks. Against GPT-6 sol it excluded zero on client talk type and included zero on therapist behaviour (−0.034 [−0.063, 0.023]). The client ECE comparison with GPT-6 sol was sensitive to binning: the 95% confidence interval for the difference included zero with ten equal-width bins. Claude Opus 5.5 was under-confident and GPT-6 sol over-confident (Table 1). Against single coders on the full corpus its ECE was 0.054 and 0.103 (Supplementary Fig. 1), so our prediction of 0.05 or less on both tasks failed. An exploratory analysis added after that failure found similar values (0.048 and 0.105) for the same ten-expert answers scored against single experts, which is consistent with the single-coder reference explaining much of the gap. Against single experts the ranking did not hold for client talk type, on which the stated probabilities of Claude Opus 5.5 were better calibrated than the typed model's. Calibration against any single reference, a plurality included, is contested when experts disagree[52].

Cost and response time

Across the retained successful calls on the 428 ten-expert utterances, the typed model's mean cost was US$0.000027 per code; GPT-6 sol and Claude Opus 5.5 cost 64 and 175 times as much. Its median request latency was 475 ms, compared with medians 7.2 and 7.8 times as long for the two LLMs (Fig. 1d). Coding all 9,699 AnnoMI utterances cost US$0.25, against a projected US$16.93 and US$46.22 for the two LLMs. These are one day's prices and times through one provider, and coder time was not measured (Methods).

Discussion

This preliminary evaluation tested a typed decision model for automated coding of motivational interviewing demonstrations. Under the planned utterance bootstrap, the agreement-difference intervals between the typed model and Claude Opus 5.5 included zero. The typed model's retained calls were faster and less costly than both LLMs',and its estimated calibration error against the ten-expert plurality was lower, clearly so against Claude Opus 5.5.Whether these differences hold in independent clinical sessions remains untested.

Well-calibrated confidence estimates could support automated coding in research. At the 0.9 threshold of our analysis plan, 62% of the typed model's therapist codes and 50% of its client codes met that threshold, with accuracy 0.977 and 0.990 against the ten-expert plurality; against single coders on the full corpus, 60% and 49% met the same threshold, with accuracy 0.917 and 0.868. This selective-review workflow[30] was simulated on these answers. When read through their option-token probabilities, LLMs answer in the same form and can be well calibrated, if need be after rescaling[53]. A probability over a closed list makes calibration measurable and usable task by task, and this study cannot tell whether the calibration came from that format, from this model's training or from this dataset.

Speed and cost answer the third question only for coding after the fact. Despite the large relative differences, projected API costs for the full corpus were below US$50 for both LLMs, with Jev remaining the least expensive. Human coding costs were not measured, so the study does not quantify savings relative to human coders. The typed model's median of 475 ms per code, against about 3.4 s for GPT-6 sol and 3.7 s for Claude Opus 5.5, and its per-code cost fit what real-time monitoring needs, but real-time use itself was not tested. The calls were made after the sessions, on professional transcripts, through one provider, and each input held the

utterance after the target, which a live system would not yet have; a live system would also add speech recognition before each code.

If these results hold on real sessions, the main clinical gain would be to make coding routine rather than exceptional. Therapists could receive feedback on specific skills after every session; in a feasibility study, therapists found automated feedback of this kind easy to use and representative of their work[54]. Supervisors could use automated fidelity feedback to target supervision, a use that community therapists and clinical leaders found acceptable[55]. Services and trials could monitor fidelity in every session rather than a sample, and researchers could relate therapist and client behaviour to outcomes across thousands of sessions[12]. In each case, calibrated confidence would decide which codes a person must check. Such tools will also meet concerns about stigma and inappropriate crisis responses[26]. Legal limits on AI therapy include an Illinois law that bars AI from therapeutic decision-making but allows licensed professionals to use it for administrative and supplementary support[56]. A coder that informs a clinician's review, rather than acting on its own, suits a supporting role of this kind, although whether a given use is permitted must be judged under each jurisdiction's rules.

The same properties matter well beyond coding. Wherever software must make many judgements from a closed set, a model of this class could serve as the decision layer. Building on the developer's proposed uses in software workflows and automated verification[41], we suggest potential applications: in agentic systems that route requests, choose tools or decide when to hand over to a person; in safety monitors that flag risk in conversations with chatbots; in screening and triage of intake forms or messages; and in reviews, such as screening studies for a systematic review or checking a manuscript against a reporting guideline. In each, a calibrated probability decides what the machine may do alone and what goes to a person, and low cost and latency make it affordable to judge every item, in real time or in bulk, rather than a sample.

These results are preliminary, and their validity on real sessions is untested. They come from one public corpus of English demonstration sessions and one coarse coding scheme (four therapist and three client codes), in which about a quarter of therapist utterances were brief back-channels that are easy to code, and we did not test sycophancy or sensitivity to prompt format[23,27]. The model was weakest on client sustain talk (F1 score 0.554 against single coders), a code associated with worse outcomes[46]. Claude models wrote and checked the rewrites without a human check; LLM judges can favour their own generations[57], and the LLMs' client

point estimates were higher on the rewrites, by 0.022 and 0.050, with no intervals computed. We tested one version of each closed model, whose training data are undisclosed and whose behaviour can change between versions[58]. The LLM comparisons used two frontier models at their providers' default settings. Smaller models and alternative reasoning settings were not evaluated; the observed cost and latency advantages therefore apply to the configurations tested and may not extend to LLMs optimised for this coding task. Independent evaluations of Jev on other tasks are mixed. An open model of the same shape was better calibrated than Jev on three of four tasks, two of them in its training data[44], and a classifier trained on labelled examples beat Jev on the one dataset where such examples existed[59].

Systems built on this class will carry limits of their own. A closed list returns only the options it was given, so it must include a no-match option wherever none may fit. In the configuration tested, the model read text only and returned judgements without reasons, which makes its errors harder to audit. Jev is closed and hosted by one developer, with undisclosed architecture, training data and weights; open models of the same shape run locally[44] but were less accurate than Jev on an intent task outside their training data, and one instance cannot show what the class as a whole can do. Patient transcripts should not be sent to the service as offered today: under US health privacy rules, for example, a cloud service that processes protected health information for a clinic needs a business associate agreement[60], and the developer's privacy policy does not mention one[61]. The next tests are real sessions transcribed by speech recognition and coded as they unfold from preceding context only, other coding schemes and languages, rewrites checked by people, and selective review tested with human coders. We propose that researchers test the accuracy and calibration of this class on their own decision tasks, populations and languages, and adopt it only where it holds up. Where it does, such models could take over many of the small judgements that clinical systems now leave to people or to slow, costly LLMs, and pass to people the ones they are unsure of.

Methods

Design

We tested the class through its most prominent current instance, Jev 1.13, running it on all 9,699 utterances of AnnoMI and comparing it, on the 428 utterances coded by all ten experts[19], with those experts and with two frontier LLMs, Claude Opus 5.5 and GPT-6 sol, for agreement, calibration, cost and time. Two checks tested whether agreement could come from memorised codes; the second, which submitted the 428 utterances for rewording, also tests how far answers

depend on wording. Earlier automated MI coders were supervised classifiers that need sessions already coded for each scheme[17,62]. Prompted LLMs need no such data and reached Cohen's κ (chance-corrected agreement; 1 is perfect) against AnnoMI's published codes of 0.51 on all client utterances[50] and 0.53 on 40 sessions[63].

Data

AnnoMI[19] contains 133 professionally transcribed demonstrations of motivational interviewing (110 showing high-quality and 23 low-quality practice), taken from videos posted online by MI practitioners and organisations, with 9,699 utterances: 4,882 therapist and 4,817 client. We used its two coding tasks: the main behaviour of each therapist utterance (question, therapist input, reflection or other) and the talk type of each client utterance (change, sustain or neutral). Seven transcripts (428 utterances: 216 therapist and 212 client) were coded independently by all ten experts; on them we recomputed the experts' Fleiss' κ as 0.737 for therapist behaviour and 0.467 for client talk type, and these match the published values[19]. Each of the other 126 transcripts was coded by one expert (11-13 transcripts per expert). The repository[64] is public and has no licence file. The sessions are demonstrations, and no patient data were used.

Coding prompt

Each utterance was coded in a separate call. The input held one line naming motivational interviewing and the session topic, the five utterances before the target and the one after it, with the target marked. The question listed the task's options with definitions closely paraphrased from the AnnoMI paper[19]. Every model received the same input and question; the exact text is in the analysis code, available on request. The same coding questions and option definitions were used in the pilot and subsequent study runs.

The typed decision model was Jev 1.13 (TypeSafe AI; pinned on OpenRouter as typesafe/jev-1.13-20260917), called through OpenRouter's decision endpoint. The ten-expert and rewording runs (23 September 2026) requested the alias jev-latest; every response identified this build, and the full-corpus run requested it by its pinned identifier. For a choice question it returns the chosen option and a probability for each option; we took the chosen option as its code and that option's probability as its confidence. At the time of the study it was the developer's only released model, and both of its aliases pointed to it[65]. The developer publishes no results on public benchmarks[41]. It states that English is the model's primary training language[65]. Open models built to the same input and output format have since been released

by others, among them open-jev, Kev-0.8B and Laya, which an independent evaluation ran on a local machine[44].

The LLM controls were Claude Opus 5.5 (anthropic/claude-opus-5.5) and GPT-6 sol (openai/gpt-6-sol), called through OpenRouter's chat-completions endpoint with these identifiers fixed. Both LLMs ran at their default sampling settings. No reasoning setting was sent, so each model ran at its provider's default, and reasoning tokens were recorded. Each LLM received the typed model's input and question byte for byte, with one added instruction to answer in JSON with a label and a probability for every option, with the probabilities summing to 1. Neither model returns token log-probabilities through OpenRouter, so an LLM's confidence here is the probability it stated for its label. Stated probabilities were renormalised to sum to 1, a failed or unparseable call was retried up to twice, and no answer was missing. We used the stated probabilities as given; post-hoc recalibration, which repairs the calibration of token probabilities[53], might also narrow the gap for stated ones.

Agreement

On the ten-expert subset, each expert was scored against the plurality code of the other nine, and utterances on which those nine tied were left out. Each model was scored against the same ten references, and Cohen's κ was averaged over them (leave-one-out κ). We also computed κ between each coder and each single expert, beside the mean κ between pairs of experts, which was 0.737 for therapist behaviour and 0.470 for client talk type. On the 126 single-coder transcripts (9,271 utterances) the reference was the published code.

Calibration and Selective Coding

Expected calibration error (ECE) is the sample-weighted mean absolute difference between mean confidence and accuracy in four bins: [0, 0.5), [0.5, 0.7), [0.7, 0.9) and [0.9, 1][66]. On the ten-expert subset the reference was the ten-expert plurality, and tied utterances were left out, which left 212 therapist and 202 client utterances. The direction of miscalibration is mean confidence against accuracy. Supplementary Fig. 1 draws a reliability diagram in ten equal bins and shows the seven that hold at least 20 utterances; the ECE uses the four bins above. Selective coding reports the share of utterances with confidence of 0.9 or more, one of two thresholds named in the committed analysis plan, and the accuracy among them. The threshold was not validated on new data. ECE measures calibration of the selected label against the specified

reference; its value depends on binning and sample size, and estimates from 212 and 202 utterances are imprecise.

Memorisation Checks

The first check used the full corpus. We set κ of 0.90 or more against the published codes on the single-coder transcripts, on both tasks, as the operational prediction for memorisation; lower values cannot exclude contamination. For comparison, agreement within 0.10 of the pilot single-expert κ values (0.772 and 0.561) was prespecified, giving ranges of 0.67-0.87 for therapist behaviour and 0.46-0.66 for client talk type. These thresholds were descriptive probes of training-data exposure. The second check submitted all 428 ten-expert utterances for rewording. Of these, 67 remained textually unchanged; 58 had three words or fewer and nine were longer, using the run script's word-count rule. Claude Sonnet agents received inputs without expert codes and prepared the candidate versions. A Claude Opus agent checked each version against a supplied expert code. For 14 utterances with tied leading expert codes, the checker received one of those codes without an explicit tie indicator. The 5 rewrites it flagged as changing the code or making it more obvious were excluded from scoring, which left 214 therapist and 209 client utterances. Their rewrites still appeared in the context windows of other retained utterances. For utterances of more than five words, the original and the candidate version shared on average 0.59 of their combined distinct words (Jaccard overlap). We predicted a clear drop in leave-one-out κ on the rewrites under memorisation and a drop of 0.05 or less under judgement. The two LLMs coded the same candidate versions.

Cost and Time

Cost and latency were summarised from retained successful responses: the OpenRouter-reported charge and client-side request-to-response time, with latency reported as the median. One Claude call was retried after a failed request (an HTTP or network error, not an unparseable answer); the earlier attempt's charge and latency were not retained. The summaries exclude any unrecorded cost or retry delay for that attempt. All calls were made on 23 September 2026, one utterance per call, without batching and five requests at a time for every coder. Per-code cost, latency and the ratios between coders come from the 428 ten-expert utterances, which all three models coded. The typed model's cost for the full corpus is its actual bill, and the LLMs' full-corpus figures are projections from their per-code cost. The time human coders would spend was not measured.

Statistics and Reproducibility

Intervals are 95% percentile bootstrap intervals from 5,000 resamples (seed 20260923). On the ten-expert subset, utterances were resampled and paired across coders, and a difference between coders is the bootstrap interval of the paired difference. These intervals treat utterances as independent, although the utterances are clustered within seven transcripts. Inference with so few clusters is fragile, and a bootstrap over seven transcripts is itself unreliable[67], so we report transcript resampling only as a sensitivity check. For the typed model's margin over the experts' mean it gave 0.028 [−0.002, 0.057] for therapist behaviour and 0.154 [0.067, 0.215] for client talk type, against [−0.022, 0.076] and [0.081, 0.229] with utterances resampled; with seven clusters, neither interval should be read as precise. On the full corpus, whole transcripts were resampled. We report nominal bootstrap confidence intervals without adjustment for multiple comparisons and do not report *P* values. The comparisons are descriptive and conditional on the observed panel and transcript sample. The typed model gave the same label on 99.3% of the 428 ten-expert utterances across two runs; each LLM coded a fixed sample of 100 of them (50 per task) three times.

Predictions and the analysis plan were committed to version control before the corresponding runs: C1 (memorisation against judgement) and C2 (an ECE of 0.05 or less against single coders) before the first full-corpus call (commit 6327bd2), and F1 to F4 before any LLM call (commit 72a3d34). The paraphrase prediction was written into the run script before that run (commit 74263f7). These commits sit in a local repository, and no third-party registry holds them.

The first ten-expert analysis of the typed model had no predictions written in advance. The C1 agreement point estimates fell within the prespecified ranges; C2's 0.05 calibration threshold was exceeded on both tasks. For F1, Claude Opus 5.5's point-estimate differences were within 0.10 κ of the typed model on both tasks, while GPT-6 sol's client difference exceeded 0.10. For F2, each LLM's plurality ECE point estimate exceeded the typed model's by at least 0.03 on both tasks, although the interval for GPT-6 sol on therapist behaviour included zero. The predicted over-confidence direction was observed only for GPT-6 sol; Claude Opus 5.5 was under-confident against that reference. For F3, each LLM's paraphrase-drop point estimate was at most 0.05; uncertainty in those drops was not estimated. F4 (GPT-6 sol less repeatable than the typed model) needed no ruling, because all three coders gave the same label on 98% or more of repeated utterances. One exploratory reference-calibration analysis was added after C2

failed and is labelled wherever it appears. It was later recomputed on the untied utterances after the audit. During the full-corpus run, and before any of its results were computed, the typed model's code was switched to its own choice field, because two options had tied at the top in 8 of the first 2,873 answers; the ten-expert analyses were unaffected.

Literature Screen

We screened the 77 records from 2025 onwards that Semantic Scholar and OpenAlex list as citing AnnoMI, duplicates included, reading the full text or the available passages of 59 and the abstract or record of the other 18, and re-checked the 28 that evaluated, or might have evaluated, a model on AnnoMI. On the text available, none scored a model against the ten-expert subset or analysed the calibration of a model's confidence.

Ethics

The study analysed publicly released transcripts of demonstration sessions; it recruited no participants and used no patient data. Ethics approval was not required for this analysis.

Use of AI tools

Claude models (Anthropic) generated and checked the rewritten statements used in the rewording experiment. Claude also assisted with code development and quality checks. During the preparation and revision of this manuscript, Claude (Anthropic) was used for proofreading and language editing. The authors reviewed and edited the outputs as needed and take full responsibility for the accuracy, integrity and final content of the manuscript.

Data availability

AnnoMI is publicly available at https://github.com/uccollab/AnnoMI. The data supporting the findings of this study are available from the corresponding author upon request.

Code availability

The analysis code, the prompts and the scripts that generate the reported study results are available from the corresponding author upon request.

Competing interests

Ziv Ben-Zion reports consulting for Talkspace and research funding from Anthropic (AI for Science program), both outside the submitted work. All other authors declare no financial or non-financial competing interests.

Acknowledgements

This study received no specific funding. David Piterman thanks the Azrieli Foundation for its support.

Author contributions

E.R. contributed to conceptualization, supervision, data review, and manuscript review and editing. M.B. and S.S. contributed to data review and manuscript review and editing. H.D. and Z.B.-Z. contributed to manuscript review and editing. D.P. contributed to conceptualization and manuscript review and editing. Z.E. contributed to conceptualization and supervision. D.P. and Z.E. contributed equally to this work.

References

1. Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9,** 36-45 (1966). https://doi.org/10.1145/365153.365168
2. Rousmaniere, T., Zhang, Y., Li, X. & Shah, S. Large language models as mental health resources: patterns of use in the United States. *Pract. Innov.* **11,** 139-155 (2026). https://doi.org/10.1037/pri0000292
3. Rousmaniere, T., Goldberg, S. B. & Torous, J. Large language models as mental health providers. *Lancet Psychiatry* **13,** 7-9 (2026). https://doi.org/10.1016/S2215-0366(25)00269-X
4. Elyoseph, Z., Hadar-Shoval, D., Asraf, K. & Lvovsky, M. ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* **14,** 1199058 (2023). https://doi.org/10.3389/fpsyg.2023.1199058
5. Levkovich, I. & Elyoseph, Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR Ment. Health* **10,** e51232 (2023). https://doi.org/10.2196/51232

6. Stade, E. C. *et al.* Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Ment. Health Res.* **3,** 12 (2024). https://doi.org/10.1038/s44184-024-00056-z
7. Obradovich, N. *et al.* Opportunities and risks of large language models in psychiatry. *NPP Digit. Psychiatry Neurosci.* **2,** 8 (2024). https://doi.org/10.1038/s44277-024-00010-z
8. Omar, M. *et al.* Applications of large language models in psychiatry: a systematic review. *Front. Psychiatry* **15,** 1422807 (2024). https://doi.org/10.3389/fpsyt.2024.1422807
9. Hua, Y. *et al.* A scoping review of large language models for generative tasks in mental health care. *npj Digit. Med.* **8,** 230 (2025). https://doi.org/10.1038/s41746-025-01611-4
10. Jin, Y. *et al.* The applications of large language models in mental health: scoping review. *J. Med. Internet Res.* **27,** e69284 (2025). https://doi.org/10.2196/69284
11. Heinz, M. V. *et al.* Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI* **2,** AIoa2400802 (2025). https://doi.org/10.1056/AIoa2400802
12. Imel, Z. E., Steyvers, M. & Atkins, D. C. Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy* **52,** 19-30 (2015). https://doi.org/10.1037/a0036841
13. Hurlocker, M. C., Madson, M. B. & Schumacher, J. A. Motivational interviewing quality assurance: a systematic review of assessment tools across research contexts. *Clin. Psychol. Rev.* **82,** 101909 (2020). https://doi.org/10.1016/j.cpr.2020.101909
14. Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D. & Houck, J. M. The Motivational Interviewing Treatment Integrity code (MITI 4): rationale, preliminary reliability and validity. *J. Subst. Abuse Treat.* **65,** 36-42 (2016). https://doi.org/10.1016/j.jsat.2016.01.001
15. Flemotomos, N. *et al.* Automated evaluation of psychotherapy skills using speech and language technologies. *Behav. Res. Methods* **54,** 690-711 (2022). https://doi.org/10.3758/s13428-021-01623-4
16. Miller, W. R. & Rollnick, S. *Motivational Interviewing: Helping People Change and Grow* 4th edn (Guilford Press, 2023).
17. Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C. & Srikumar, V. A comparison of natural language processing methods for automated coding of motivational

interviewing. *J. Subst. Abuse Treat.* **65,** 43-50 (2016). https://doi.org/10.1016/j.jsat.2016.01.006

18. Peräkylä, A., Antaki, C., Vehviläinen, S. & Leudar, I. (eds) *Conversation Analysis and Psychotherapy* (Cambridge Univ. Press, 2008). https://doi.org/10.1017/CBO9780511490002
19. Wu, Z. *et al.* Creation, analysis and evaluation of AnnoMI, a dataset of expert-annotated counselling dialogues. *Future Internet* **15,** 110 (2023). https://doi.org/10.3390/fi15030110
20. Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. Why language models hallucinate. Preprint at https://arxiv.org/abs/2509.04664 (2025). https://doi.org/10.48550/arXiv.2509.04664
21. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630,** 625-630 (2024). https://doi.org/10.1038/s41586-024-07421-0
22. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* 35 27730-27744 (NeurIPS, 2022). https://doi.org/10.52202/068431-2011
23. Sharma, M. *et al.* Towards understanding sycophancy in language models. In *Proc. 12th International Conference on Learning Representations* (ICLR, 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html
24. OpenAI. Expanding on what we missed with sycophancy. 2 May 2025. https://openai.com/index/expanding-on-sycophancy/ (accessed 22 October 2026).
25. Østergaard, S. D. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophr. Bull.* **49,** 1418-1419 (2023). https://doi.org/10.1093/schbul/sbad128
26. Moore, J. *et al.* Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proc. 2025 ACM Conference on Fairness, Accountability, and Transparency* 599-627 (ACM, 2025). https://doi.org/10.1145/3715275.3732039
27. Sclar, M., Choi, Y., Tsvetkov, Y. & Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. In *Proc. 12th International Conference on Learning*

Representations (ICLR, 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html

28. Xiong, M. *et al.* Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *Proc. 12th International Conference on Learning Representations* (ICLR, 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/6733cf15e10e2cd1d59af033c3bb8507-Abstract-Conference.html
29. Tian, K. *et al.* Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5433-5442 (ACL, 2023). https://doi.org/10.18653/v1/2023.emnlp-main.330
30. Kompa, B., Snoek, J. & Beam, A. L. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digit. Med.* **4,** 4 (2021). https://doi.org/10.1038/s41746-020-00367-3
31. Snell, C., Lee, J., Xu, K. & Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *Proc. 13th International Conference on Learning Representations* (ICLR, 2025). https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html
32. Anthropic. Claude Opus 5.5 overview. n.d. https://platform.claude.com/docs/en/models/opus-5-5/overview (accessed 22 October 2026).
33. Artificial Analysis. Claude Opus 5.5 (Max, Default Fallback) vs GPT-6 Sol (Max). n.d. https://artificialanalysis.ai/models/comparisons/claude-opus-5-5-vs-gpt-6-sol (accessed 22 October 2026).
34. Artificial Analysis. Language Model API Performance Benchmarking Methodology. n.d. https://artificialanalysis.ai/methodology/performance-benchmarking (accessed 22 October 2026).
35. OpenAI. GPT-6 Sol model. n.d. https://developers.openai.com/api/docs/models/gpt-6-sol (accessed 22 October 2026).
36. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics 4171-4186 (ACL, 2019). https://doi.org/10.18653/v1/N19-1423

37. Desai, S. & Durrett, G. Calibration of pre-trained transformers. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing* 295-302 (ACL, 2020). https://doi.org/10.18653/v1/2020.emnlp-main.21
38. Lewis, M. *et al.* BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 7871-7880 (ACL, 2020). https://doi.org/10.18653/v1/2020.acl-main.703
39. Yin, W., Hay, J. & Roth, D. Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* 3914-3923 (ACL, 2019). https://doi.org/10.18653/v1/D19-1404
40. Facebook AI. facebook/bart-large-mnli [model card]. n.d. https://huggingface.co/facebook/bart-large-mnli (accessed 22 October 2026).
41. Almeida, D. Introducing System One Models & Jev. 15 September 2026. https://typesafe.ai/blog/introducing-system-one-models-and-jev (accessed 22 October 2026).
42. Almeida, D. Post reporting 140,000 waitlist admissions in less than 36 hours. 17 September 2026. https://x.com/CompleteSkeptic/status/2100454726462804333 (accessed 22 October 2026).
43. Fathima, H. What Is Jev? Inside TypeSafe's Decision-Only AI Model and Its Developer Use Cases. 23 September 2026. https://www.firecrawl.dev/blog/what-is-jev (accessed 22 October 2026).
44. 4esv. jev-eval [code repository]. 2026. https://github.com/4esv/jev-eval (accessed 22 October 2026).
45. TypeSafe AI. Workflow evals. n.d. https://evals.typesafe.ai/ (accessed 22 October 2026).
46. Magill, M. *et al.* A meta-analysis of motivational interviewing process: technical, relational, and conditional process models of change. *J. Consult. Clin. Psychol.* **86,** 140-157 (2018). https://doi.org/10.1037/ccp0000250

47. Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C. & Imel, Z. E. Development and evaluation of ClientBot: patient-like conversational agent to train basic counseling skills. *J. Med. Internet Res.* **21,** e12529 (2019). https://doi.org/10.2196/12529

48. Plank, B. The 'problem' of human label variation: on ground truth in data, modeling and evaluation. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* 10671-10682 (ACL, 2022). https://doi.org/10.18653/v1/2022.emnlp-main.731

49. Jacovi, A., Caciularu, A., Goldman, O. & Goldberg, Y. Stop uploading test data in plain text: practical strategies for mitigating data contamination by evaluation benchmarks. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5075-5084 (ACL, 2023). https://doi.org/10.18653/v1/2023.emnlp-main.308

50. Ali, S. *et al.* Automated coding of counsellor and client behaviours in motivational interviewing transcripts: validation and application. In *Proc. 2nd Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication (NLPAI4Health)* 25-54 (ACL, 2025). https://doi.org/10.18653/v1/2025.nlpai4health-main.4

51. Zhu, W. *et al.* CLEAN-EVAL: clean evaluation on contaminated large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* 835-847 (ACL, 2024). https://doi.org/10.18653/v1/2024.findings-naacl.53

52. Baan, J., Aziz, W., Plank, B. & Fernández, R. Stop measuring calibration when humans disagree. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* 1892-1915 (ACL, 2022). https://doi.org/10.18653/v1/2022.emnlp-main.124

53. Kadavath, S. *et al.* Language models (mostly) know what they know. Preprint at https://arxiv.org/abs/2207.05221 (2022). https://doi.org/10.48550/arXiv.2207.05221

54. Imel, Z. E. *et al.* Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* **56,** 318-328 (2019). https://doi.org/10.1037/pst0000221

55. Creed, T. A. *et al.* Knowledge and attitudes toward an artificial intelligence-based fidelity measurement in community cognitive behavioral therapy supervision. *Adm. Policy Ment. Health* **49,** 343-356 (2022). https://doi.org/10.1007/s10488-021-01167-x

56. Illinois Department of Financial and Professional Regulation. Gov. Pritzker Signs Legislation Prohibiting AI Therapy in Illinois. 4 August 2025. https://idfpr.illinois.gov/news/2025/gov-pritzker-signs-state-leg-prohibiting-ai-therapy-in-il.html (accessed 22 October 2026).
57. Panickssery, A., Bowman, S. R. & Feng, S. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems* 37 (NeurIPS, 2024). https://doi.org/10.52202/079017-2197
58. Chen, L., Zaharia, M. & Zou, J. How is ChatGPT's behavior changing over time? *Harv. Data Sci. Rev.* **6**(2)**,** https://doi.org/10.1162/99608f92.5317da47 (2024).
59. ickma2311. Jev vs. the right baselines: a pre-registered independent evaluation [code repository]. 2026. https://github.com/ickma2311/jev-baselines-eval (accessed 22 October 2026).
60. U.S. Department of Health and Human Services. Guidance on HIPAA & Cloud Computing. 23 December 2022. https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html (accessed 22 October 2026).
61. TypeSafe AI. Privacy policy. 19 November 2025. https://typesafe.ai/legal/privacy-policy (accessed 22 October 2026).
62. Atkins, D. C., Steyvers, M., Imel, Z. E. & Smyth, P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implement. Sci.* **9,** 49 (2014). https://doi.org/10.1186/1748-5908-9-49
63. Lim, K., Jung, Y. C. & Kim, B. H. Evaluating motivational interview quality using large language models and hidden Markov models. *BMC Psychiatry* **25,** 908 (2025). https://doi.org/10.1186/s12888-025-07391-1
64. Wu, Z. *et al.* / uccollab. AnnoMI [dataset repository]. n.d. https://github.com/uccollab/AnnoMI (accessed 22 October 2026).
65. TypeSafe AI. Models. n.d. https://docs.typesafe.ai/models (accessed 22 October 2026).
66. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proc. 34th International Conference on Machine Learning* 1321-1330 (PMLR 70, 2017). https://proceedings.mlr.press/v70/guo17a.html

67. Cameron, A. C., Gelbach, J. B. & Miller, D. L. Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* **90,** 414-427 (2008). https://doi.org/10.1162/rest.90.3.414

Tables

Table 1 | Coders compared on the 428 utterances coded by all ten experts.

Task and coder	Leave-one-out κ [95% CI]	Typed model minus this coder, κ [95% CI]	ECE [95% CI]	Typed model minus this coder, ECE [95% CI]	Mean confidence / accuracy	Confidence of 0.9 or more: share, accuracy	Drop on rewording
Therapist behaviour (216 utterances)							
Experts, mean (range)	0.821 (0.657-0.872)	-	-	-	-	-	-
Typed decision model (Jev 1.13)	0.849 [0.791, 0.900]	-	0.017 [0.012, 0.061]	-	0.877 / 0.887	62%, 0.977	0.025
Claude Opus 5.5	0.878 [0.827, 0.924]	−0.029 [−0.079, 0.019]	0.103 [0.075, 0.138]	−0.086 [−0.109, −0.032]	0.809 / 0.910	46%, 1.000	0.032
GPT-6 sol	0.862 [0.808, 0.910]	−0.013 [−0.063, 0.036]	0.051 [0.022, 0.090]	−0.034 [−0.063, 0.023]	0.930 / 0.896	77%, 0.982	0.031
Client talk type (212 utterances)							
Experts, mean (range)	0.595 (0.474-0.738)	-	-	-	-	-	-
Typed decision model (Jev 1.13)	0.749 [0.666, 0.824]	-	0.038 [0.019, 0.080]	-	0.849 / 0.876	50%, 0.990	0.040
Claude Opus 5.5	0.768 [0.690, 0.839]	−0.018 [−0.089, 0.055]	0.139 [0.102, 0.178]	−0.100 [−0.139, −0.046]	0.748 / 0.886	24%, 1.000	−0.022
GPT-6 sol	0.613 [0.515, 0.703]	0.136 [0.064, 0.213]	0.123 [0.074, 0.178]	−0.085 [−0.141, −0.010]	0.910 / 0.787	68%, 0.912	−0.050

Each expert is scored against the plurality of the other nine experts. Each model is scored against all ten such references, and its reported Cohen's κ is their mean (leave-one-out κ). Utterances with a tied plurality are excluded separately for each reference. A coder that tracks the consensus is favoured by this reference. ECE, expected calibration error against the ten-expert plurality in four confidence bins, with tied utterances left out (212 therapist and 202 client utterances). Confidence is the returned probability of the chosen option for the typed decision model and the stated probability of the chosen label for the LLMs. Differences are the typed model minus the coder in that row; a negative ECE difference indicates lower estimated four-bin ECE against the ten-expert plurality; its interval describes uncertainty in that difference. Brackets

give 95% percentile bootstrap intervals over utterances, paired across coders (5,000 resamples). Mean confidence above accuracy indicates over-confidence. Drop on rewording is κ on the original utterances minus κ on the corresponding versions from the rewording experiment (214 therapist and 209 client utterances). Coverage and accuracy at confidence of 0.9 or more are point estimates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf