

Small Model, Careful Pipeline: An Autoresearch Method for Closing the Gap to Frontier LLMs on Long, Multi-Document Summarisation

Bartłomiej Głowacki* Joanna Głowacka Bohdan Głowacki Ewa Cylwik

AITHENTICA P.S.A. / Therapy Support, Poland

bartlomiej.glowacki@aitherapy.support

Abstract

Frontier language models summarise long, multi-document material well but are expensive to run routinely; small models are cheap but lose information over long inputs. We present an *autoresearch* method that narrows this gap without human raters: a fully model-evaluated development loop, combined with a division of labour in which code does all counting and ordering, a sentence-level classifier returns calibrated judgments, and the small model only writes short sections. Our testbed is pre-session briefs for cognitive behavioural therapy, built from 12–20 transcripts of two synthetic patients. Over 30 versions and 16 blind review rounds, a low-cost model rose from a gold-checklist score of 0.647 to 0.846 (0.879 with best-of-3 verifier selection), against 0.782 for a frontier reference (GPT-5.5). In blind model-based review the two were indistinguishable (mean paired difference +0.05, SD 0.49), while the same unchanged reference brief scored 3.50–4.03 across 13 rounds, which makes judge noise a result in its own right. A brief costs about USD 0.03 to generate.

1 Introduction

Many practical summarisation problems share three properties: the input is long and spread over many documents, the output must follow a fixed structure with counts, dates and verbatim evidence, and the output must be produced routinely, so cost matters. Frontier LLMs handle such tasks reasonably well, but running one over all source documents before every output multiplies quickly. Small models are weak exactly where these tasks are hard: they lose information from the middle of long inputs (Liu et al., 2024), they merge partial drafts poorly, and they conflate categories that look alike on the surface.

We use one such task as a testbed: a pre-session brief for a CBT therapist. Before each session the therapist needs a three-minute picture of the whole therapy so far: core beliefs with session counts, intermediate rules, recurring distortions, typical ABC episodes, tested predictions, risk, the last session and what to bring to today’s session (Beck, 2020). The material grows to 10–20 transcripts of about 6,500 words each. The task has a known structure (Beck’s cognitive model and session structure), a natural notion of verifiable evidence (the patient’s verbatim words) and several traps (a belief first worded by the therapist; a metaphor that is not a death wish). It is therefore a good place to ask a methodological question:

How far can a low-cost model, embedded in a carefully engineered pipeline and improved by an automated, model-evaluated loop, approach the output quality of a frontier model on long, multi-document summarisation?

Our answer has two parts. First, a **division of labour**: whatever can be counted, ordered or looked up goes to deterministic code (program-aided generation; Gao et al., 2023; Chen et al., 2023); fine-grained judgments that need no text generation go to a cheap sentence-level classifier that returns probabilities; the small model writes only short sections from code-prepared inputs; and run-to-run variance is exploited by best-of- N selection with a reference-free verifier (Cobbe et al., 2021; Brown et al., 2024). Second, an **autoresearch loop** in the spirit of recent automated experimentation (Karpathy, 2026; Lu et al., 2024): every version is generated three times, graded on a weighted gold-anchored checklist by a classifier and reviewed blind by an LLM supervisor from a different model family, which always also sees one unchanged frontier reference so that the judge’s own drift is measured; failures that persist in all

*Corresponding author.

three runs drive one targeted change per round; a fixed stop rule ends the loop.

Contributions.

1. A formal description of the autoresearch loop (Algorithm 1) and of the division-of-labour pipeline (§4), instantiated on multi-session CBT briefs.
2. Evidence that the loop brought a low-cost model level with a frontier reference on content coverage (0.846 vs 0.782; 0.879 with best-of-3) and indistinguishable from it in blind model-based review at this sample size (§6), at a few US cents per brief (§6.7).
3. A quantified account of LLM-judge noise on an *identical* input across 13 rounds (SD 0.18, range 0.53 on a 1–5 scale), with the design consequences we drew from it (§6.3, §7).
4. An ablation of what helped and what did not, including several plausible additions that hurt (§6.4), and an item-level error analysis (§6.6).

This is a methods paper. All evaluation is model-based by design so that the loop can run unattended; we make no claims beyond the method and its testbed.

2 Related Work

LLMs on psychotherapy transcripts.

Component-guided summarisation of counselling sessions has been benchmarked across eleven LLMs with clinician review on 191 sessions (Adhikary et al., 2024). Simulated patients built from CBT cognitive models are used to train and evaluate clinicians (Wang et al., 2024); our synthetic patients follow the same idea at the level of whole therapies. A multi-session, multi-therapy benchmark finds that direct reconstruction from raw material flattens therapeutic arcs and recommends a structured intermediate representation (Pan et al., 2026); our per-session extraction layer is such a representation. A qualitative study of ChatGPT case conceptualisation reports difficulty separating vulnerabilities from coping strategies and handling abstract relational constructs (Buatini et al., 2026); the same confusions (automatic thought vs rule vs core belief) are the hardest items in our checklist (§6.6).

Long inputs and salience. Models use the middle of long contexts poorly (Liu et al., 2024). Two-stage approaches that first select or plan salient content and then generate from the plan im-

prove long-document summaries (Du et al., 2025); keyphrase-level salience hints help across models, with mixed effects on faithfulness (Xu et al., 2024). We use a classifier-produced salience score per patient sentence both as a planning signal (tested, not kept) and as a reference-free verifier for best-of- N (kept).

Self-correction, verification and test-time compute.

Iterative self-refinement (Madaan et al., 2023) and verification chains (Dhuliawala et al., 2024) can improve outputs, but intrinsic self-correction without external feedback often does not (Huang et al., 2024); our rubric-guided self-revision step (§6.4) is a further negative instance on a small model. Rubric-based evaluation and refinement has been proposed for medical dialogue (Chen et al., 2026). Sampling several candidates and selecting with a verifier is a simple and robust use of test-time compute (Cobbe et al., 2021; Lightman et al., 2024; Wang et al., 2023; Brown et al., 2024; Snell et al., 2024); we use $N = 3$ with a classifier verifier.

Program-aided generation. Offloading computation to code improves reliability on numerical tasks (Gao et al., 2023; Chen et al., 2023). We extend the idea from arithmetic to bookkeeping over documents: session lists, counts, ranges, chains of predictions and their status, quote verification and the ordering of an agenda.

LLM judges and automated research loops.

LLM-as-a-judge evaluation correlates with human preference but shows position, verbosity and self-enhancement biases (Zheng et al., 2023) and is inconsistent across samples and prompt variants (Stureborg et al., 2024). We add a longitudinal measurement: the same brief judged by the same prompt in 13 rounds. Automated research loops that propose, run and keep or discard one change at a time have recently been popularised for model training (Karpathy, 2026) and for end-to-end paper generation (Lu et al., 2024); our loop applies the same discipline to pipeline engineering with model-based evaluation. Where applicable we follow the TRIPOD-LLM reporting guideline (Gallifant et al., 2025).

3 Task and Testbed

Input and output. The input is the set of transcripts $T_{1:N} = (T_1, \dots, T_N)$ of one patient up to session N . The output is a JSON brief b with a

screen (what the therapist reads in about three minutes) and *details*. The screen holds: context and working hypothesis; up to four core beliefs, each with session count n , session range, state (active, recurring, weakened, challenged, new) and a verbatim patient quote; two dysfunctional rules and one new adaptive rule; up to three recurring distortions; one or two typical ABC episodes and a vicious circle; a predictions bar; numbers (mood, measures); self-risk and violence lines; the last session; and a *today* agenda. The details hold candidates, typical ABCs, open threads, homework and contradictions. Every quote on the screen must be verbatim and spoken by the patient.

Synthetic patients. Two synthetic Polish-speaking patients were generated for this purpose with the authors’ layered synthetic-patient method: a clinical profile sheet, a therapeutic arc, a session-by-session plan and a voice style guide are written first, and the sessions are then generated sequentially, each conditioned on the plan and on the preceding sessions, as time-stamped ASR-style transcripts with hesitations and self-corrections (about 50 minutes per session): **Anita**, 20 sessions, 131,498 words, depression and anxiety treated in J. Beck-style CBT; and **Markus**, 12 sessions, 83,978 words, whose therapist works partly in schema therapy, so that the brief must translate the material back into CBT terms. Each patient has a **gold reference** written independently from a full reading of the transcripts: the expected conceptualisation at given checkpoints plus explicit traps. Checkpoints are Anita after S12 and S19 and Markus after S11 (three checkpoints). No real patient data were used.

4 Method

4.1 Division of labour

Figure 1 shows the final pipeline (v28). Three kinds of component appear, each used for what it is good at:

- **Small LLM** (LUNA, a low-cost deployment, §4.5): short generative steps with a fixed schema; per-session extraction, a critic, a hypothesis step and six section writers.
- **Sentence classifier** (JEV): judgments that need no generation, returned as typed answers with probabilities; applied to every patient sentence once, and reused as a reference-free verifier and as the checklist grader.

- **Deterministic code:** everything that can be counted, ordered, looked up or verified by string matching.

4.2 Pipeline

Per-session layer (once per new session). For each session i the small LLM produces an extraction $x_i = X(T_i)$: ABC episodes, core-belief signals with slot (helplessness, unlovability, worthlessness, others, world) and evidence type (in-vivo, reported from the week, retrospective), rules, distortions with quotes, predictions with percentages and outcomes, homework, risk signals, violence, therapeutic-relationship events, mood and measures. A critic call $k_i = K(T_i)$ answers 12 yes/no questions (one per distortion) with a verbatim patient quote each; a distortion is kept only if it is in $x_i \cap k_i$ and its quote passes the quote gate. A third call $h_i = H(x_i)$ derives core-belief hypotheses by the downward arrow and stores them only as a hypothesis layer, never as evidence. Independently, the classifier labels every patient sentence $s_{i,j}$ (with the preceding therapist turn and the patient’s previous words as context): $\ell_{i,j} = C(s_{i,j}, \text{ctx}_{i,j})$ with CBT level (automatic thought, rule, belief about self, belief about others/world, emotion, behaviour, fact, new perspective, echo), time frame (now/recent, retrospective, hypothetical), core-belief slot, literal self-risk (death, self-harm vs metaphor such as “as if I did not exist”), violence, and clinical salience 0–4 with a reason (risk, belief/rule, change/result, life event, therapeutic relationship). In total 11,464 patient sentences were labelled (7,980 Anita, 3,484 Markus).

Bookkeeping (code). From $x_{1:N}, k_{1:N}, h_{1:N}, \ell$ code computes the section inputs $I = B(x_{1:N}, k_{1:N}, h_{1:N}, \ell)$: session lists and n and ranges for each belief (a session counts only if the extraction or the classifier has in-vivo or reported evidence there); prediction chains linked across sessions, with status *resolved* only when an outcome sentence exists; the verbatim-quote gate (normalised substring match against the patient’s lines; a quote found only in the therapist’s lines is dropped); removal of therapist echoes; the risk-screening reminder when the last direct assessment is ≥ 4 sessions old; a scrub of technical words; and, from v28, the *today* agenda as a contract in Beck’s session order: (1) a planned measurement or end-of-phase closure before a break; (2) the test or prediction set in the last

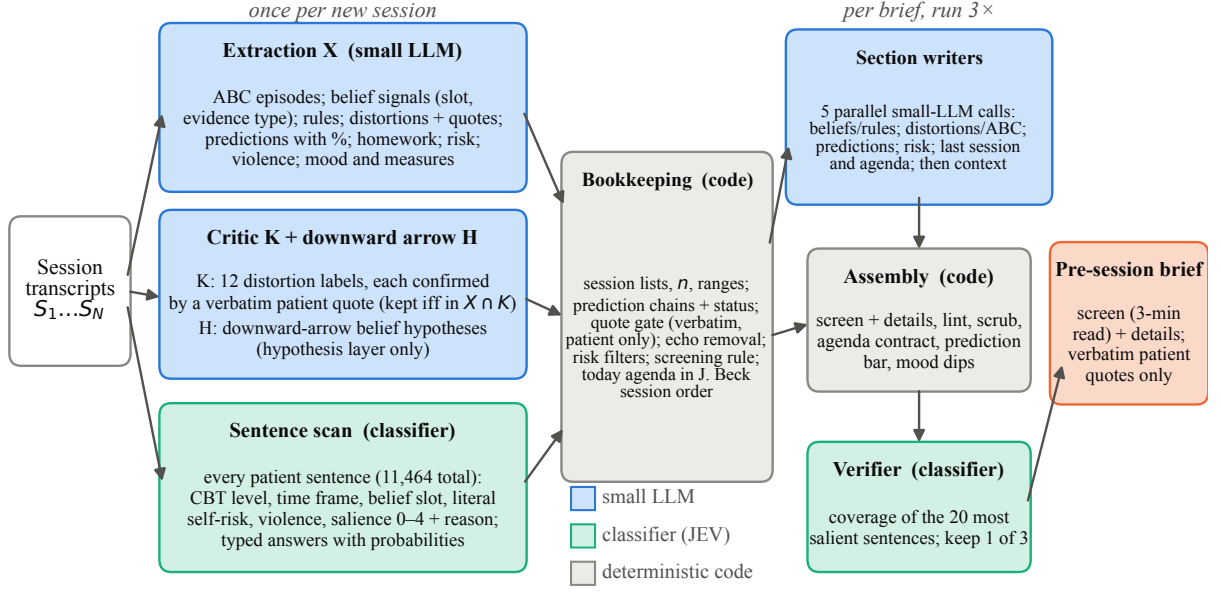


Figure 1: Final pipeline (v28). Left: once per new session the small LLM extracts structure (X), confirms distortions with verbatim quotes (K) and derives downward-arrow hypotheses (H), while the classifier labels every patient sentence. Centre: deterministic code does all bookkeeping. Right: per brief, five section writers run in parallel (then a context writer), code assembles and lints the screen, the pipeline is run three times and a reference-free classifier verifier keeps one.

session; (3) homework from the last session; (4) risk screening if due; (5) the model’s remaining items, deduplicated.

Section writers and assembly. Five small-LLM calls run in parallel, $y_m = W_m(I_m)$ for $m \in \{\text{beliefs/rules, distortions/ABC, predictions, risk, last session/agenda}\}$, followed by a context/numbers call that sees the top beliefs. Each writer sees only the data for its section, with session numbers attached to every item, and is instructed never to invent session numbers or quotes. A lint step re-asks the beliefs writer if a belief is conditional (then it is a rule). Code assembles the screen and details, re-runs the quote gate, removes echoes, refills empty quote fields with verbatim sentences from the extraction, builds the prediction bar, appends the most recent literal risk signal and recent mood dips, and flags beliefs whose wording the therapist proposed first.

4.3 Best-of- N with a reference-free verifier

The brief-level part of the pipeline is run $N = 3$ times, $b^{(1)}, b^{(2)}, b^{(3)}$. The verifier takes the $K = 20$ patient sentences with the highest classifier salience (bonus for the last third of sessions and for risk) and asks the classifier, for each, whether its meaning is reflected in the brief; the verifier

score is the mean probability,

$$V(b) = \frac{1}{K} \sum_{k=1}^K P(s_k \text{ covered} \mid b), \quad (1)$$

$$b^* = \arg \max_r V(b^{(r)}).$$

The verifier never sees the gold reference or the checklist. Its cost is a single classifier call per candidate.

4.4 The autoresearch loop

Let π denote a pipeline version (prompts, code, settings), $\mathcal{C}$ the set of checkpoints, $R = 3$ the number of runs and b_{REF} a fixed frontier reference brief per checkpoint. The loop (Figure 2, Algorithm 1) evaluates each version with two independent model-based metrics (§5): the gold-checklist score G and the blind supervisor score Q . Each round starts from the best version so far; a change is accepted only if it improves G and does not lower Q within the same round; failures that persist in all R runs and the supervisor’s objections are diagnosed and the literature is searched for a remedy before the next single, targeted change. The loop stops after two consecutive rounds without improvement. The cost constraint is fixed in advance: no model more expensive than the

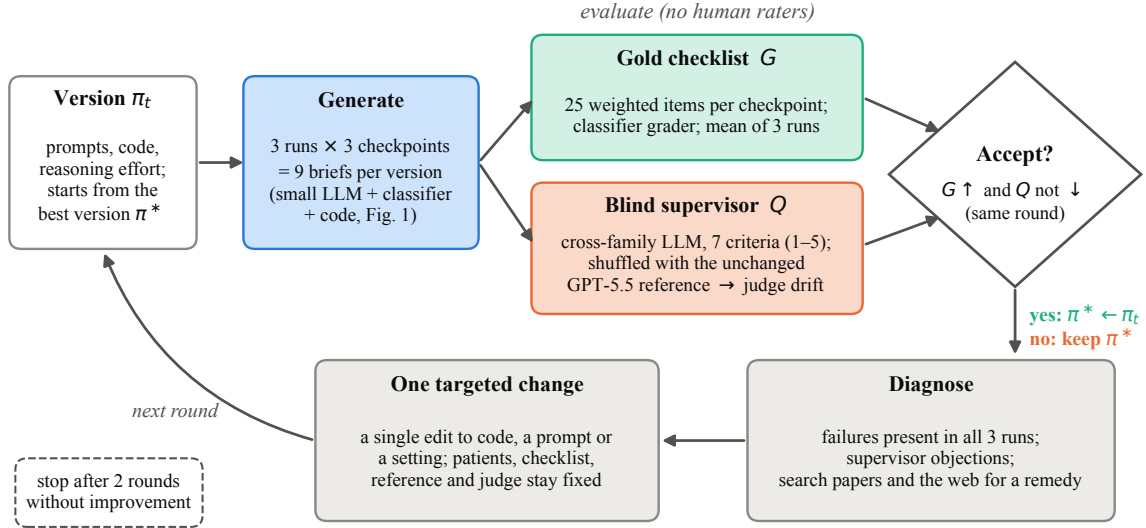


Figure 2: The autoresearch loop. Each round generates nine briefs (three runs at three checkpoints), scores them with two independent model-based metrics, and accepts a version only if the gold-checklist score rises and the blind supervisor score does not fall against the unchanged frontier reference in the same round. Failures present in all three runs, the supervisor’s objections and a literature search drive one targeted change per round. Over the project the loop ran through 30 versions and 16 blind review rounds and stopped at v28.

Algorithm 1 Autoresearch loop (as run; the proposal step was partly manual).

Require: π_0 ; checkpoints $\mathcal{C}$; runs R ; fixed reference briefs b_{REF} ; patience p

- 1: $\pi^* \leftarrow \pi_0$; $t \leftarrow 0$
- 2: **while** $t < p$ **do**
- 3: $\delta \leftarrow \text{PROPOSE}(\text{diagnosis, literature})$
- 4: $\pi' \leftarrow \pi^* \oplus \delta$ ▷ one targeted change
- 5: $b_c^{(r)} \leftarrow \pi'(T^c)$ for $r \leq R, c \in \mathcal{C}$
- 6: $G(\pi') \leftarrow \text{mean}_{r,c} G(b_c^{(r)})$ ▷ checklist
- 7: $Q \leftarrow \text{BLINDREVIEW}(\{b_c^{(1)}\} \cup \{b_{\text{REF},c}\})$
shuffled, unlabelled, same round
- 8: $\text{diagnosis} \leftarrow$ items failed in all R runs,
supervisor objections
- 9: **if** $G(\pi') > G(\pi^*)$ **and** $Q(\pi') \geq Q(\pi^*)$ **then**
- 10: $\pi^* \leftarrow \pi'$; $t \leftarrow 0$
- 11: **else**
- 12: $t \leftarrow t + 1$ ▷ keep π^* ; try a branch
- 13: **return** π^* with best-of- R selection by V

Role	Model and list price (USD per 1M tokens, in/out)
Extraction, critic, hypotheses, section writers	LUNA: low-cost GPT-6-family deployment on Azure OpenAI (gpt-6-luna, version 2026-09-22, Data Zone Standard, EU); 0.12 / 0.60
Sentence scan, verifier, checklist grader	TypeSafe JEV classifier (jev-1.14.0); typed answers with probabilities; 0.042 / output free
Frontier reference	GPT-5.5 (gpt-5.5, version 2026-04-24) on the same Azure tenant and deployment type; pipeline v9; 5.50 / 33.00
Blind supervisor	Claude Opus 5.5 (claude-opus-5-5); separate subscription

Table 1: Models. The supervisor comes from a different model family than the systems it judges (GPT), which removes self-enhancement bias (Zheng et al., 2023) from the comparison; the classifier is a third family.

small LLM or the classifier may be used inside the pipeline.

4.5 Models

Table 1 lists the models. From v25 the section writers run with the provider’s *high* reasoning-effort setting (v25h); v25x tested the highest setting. Open-weight models at a similar price were tried in an exploratory step and dropped (content-filter rejections on 15–30% of clinical texts; invalid structured output).

5 Evaluation Protocol

Gold checklist. For each checkpoint, 25 weighted statements (weights 1–3; total weight 54, 55 and 55) were derived from the gold reference, covering core beliefs (5 items), rules, distortions, ABC, predictions, self-risk, violence (Markus), last session, today agenda, CBT purity and technical leaks. Examples: “the agenda includes repeated BDI-II/GAD-7/PHQ-9 with item 9 and closure before the summer break”; “the brief does not present ‘I could just not be

there’ (S10) as a passive death wish”; “no distortion on the screen has $n > 9$ (material: 12 sessions)”. The classifier grades each statement against the brief JSON (technical fields removed) and returns a probability p_k ; the score is

$$G(b) = \frac{\sum_k w_k \mathbf{1}[p_k(b) \geq 0.5]}{\sum_k w_k}. \quad (2)$$

Every version from v14 on was run three times; we report the mean over runs and checkpoints (198 graded briefs in total).

Blind supervisor. An LLM prompted as a demanding certified CBT supervisor received, per checkpoint, three or four briefs under random letters (re-drawn per checkpoint), the gold reference and the transcripts, and scored seven criteria on 1–5: accuracy, CBT levels, session trace (lists, counts, chronology, checked against transcripts), quotes, usefulness (three-minute read), CBT purity and calibration (hypotheses, risk and violence, belief states, prediction status). The overall score Q is the unweighted mean of the seven. The supervisor also verified every quote by script and wrote ranked objections per brief, which fed the diagnosis step. The frontier reference brief was included *unchanged* in every round from R10 on (13 rounds), so the judge’s drift on an identical input is observed directly. Letters, keys and scores were never shown to the supervisor.

Frontier reference. The reference briefs were generated once with GPT-5.5 in pipeline v9 (one extraction call per session and one brief call per checkpoint, with up to two schema retries) and then reused as files. The reference is therefore an earlier, unoptimised pipeline with a frontier model, not a frontier pipeline tuned with the same loop; it was graded on the checklist once (single run).

Statistics. With three checkpoints and three runs the design supports descriptive statistics only. We report means, standard deviations between runs, same-round paired differences and ranges; we do not report significance tests, and we treat any supervisor difference smaller than the judge’s own round-to-round SD (§6.3) as uninterpretable.

6 Results

6.1 Content coverage (gold checklist)

The small-model line started well below the reference (v9: 0.647 vs 0.782), crossed it at v25 (0.785)

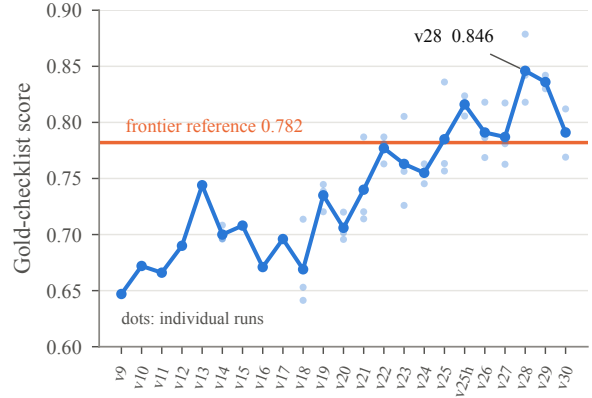


Figure 3: Gold-checklist score by pipeline version (small-model line from v9; mean over three checkpoints, dots show individual runs where three runs exist). The horizontal line is the frontier reference (GPT-5.5, pipeline v9, one run).

	Anita S12	Anita S19	Markus S11
<i>Gold checklist (0–1)</i>			
v28, mean of 3 runs	0.963	0.800	0.776
SD between runs	0.000	0.066	0.058
v28, best-of-3 (verifier)	0.963	0.836	0.836
GPT-5.5 reference (1 run)	0.963	0.655	0.727
<i>Blind supervisor (1–5), same round</i>			
R28: v28 / reference	3.6 / 4.0	4.1 / 3.9	3.6 / 3.7
R29: v28 / reference	3.4 / 3.9	4.0 / 3.1	3.9 / 3.7

Table 2: Final version v28 against the frontier reference per checkpoint. Checklist means over checkpoints: v28 0.846, best-of-3 0.879, reference 0.782. Supervisor means over checkpoints: R28 3.77 vs 3.87; R29 3.77 vs 3.57.

and reached **0.846** at v28 (Figure 3; Table 2). All three v28 runs (0.818, 0.842, 0.879) lie above the reference. Per checkpoint, v28 equals the reference on Anita S12 (0.963 vs 0.963), and exceeds it on Anita S19 (0.800 vs 0.655) and Markus S11 (0.776 vs 0.727). Two later versions (v29, v30) did not improve on v28 and the loop stopped.

6.2 Blind supervision

In the last three review rounds the best small-model version matched the reference within judge noise: R25 3.67 vs 3.60 (v25h), R28 3.77 vs 3.87 and R29 3.77 vs 3.57 (v28). Individual v28 checkpoints reached 4.1 (Anita S19, R28) and 4.0 (R29); its weakest checkpoint was Anita S12 (3.6, 3.4), where the supervisor missed a belief reactivated in S12 and a duplicated prediction. Over the six same-round pairs (v28 vs reference, R28–R29) the paired difference is $-0.4, +0.2, -0.1, -0.5, +0.9, +0.2$:

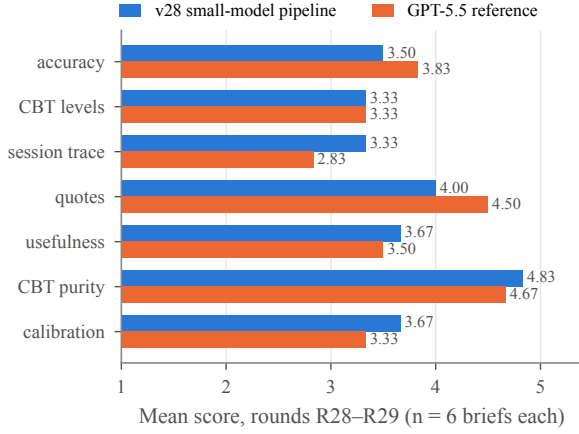


Figure 4: Blind supervisor scores by criterion for v28 and the frontier reference, mean over rounds R28–R29 (three checkpoints each, $n = 6$ brief scores per system).

mean $+0.05$, SD 0.49 . The average target of 4.0 was reached by neither system in any round. Figure 4 breaks the two systems down by criterion: v28 is ahead on session trace (3.33 vs 2.83), calibration (3.67 vs 3.33), usefulness and CBT purity, behind on accuracy (3.50 vs 3.83) and quotes (4.00 vs 4.50), and equal on CBT levels (3.33).

6.3 Judge noise as a result

Because the reference brief never changed, its scores measure the supervisor alone (Figure 5a). Across 13 rounds the round mean was 3.71 with SD 0.18 and range 3.50 – 4.03 (0.53 on a five-point scale); per checkpoint the SD was 0.21 (Anita S12), 0.29 (Anita S19) and 0.18 (Markus S11), with single-brief scores spanning 3.1 – 4.1 . Per criterion, the round-to-round SD ranged from 0.16 (accuracy, usefulness) to 0.46 for *quotes* and *CBT purity*, although the quotes were script-verified and identical in every round. The noise is comparable to the whole improvement the loop produced on the same scale (the best small-model version moved from 3.37 to 3.77 between R10 and R28).

Two consequences follow. First, any cross-round comparison of absolute scores is unsafe; only same-round, same-prompt comparisons carry information (Figure 5b), and even those should be read against an SD of about 0.2 for a three-checkpoint mean. Second, a single judge cannot be the sole metric of an unattended loop; the classifier-graded checklist, which is deterministic given the brief, carried the accept/reject decisions, and the supervisor served as a veto and as a source of diagnoses. The paired series in Figure 5b never-

theless shows a clear trend through the noise: ten negative differences in a row, then three near zero.

6.4 Ablations

Figure 6 and Table 3 isolate the changes of the last six rounds. Three changes helped: (i) fixing defects in deterministic code found by the supervisor (an empty vicious-circle field, session numbers formatted as “SS10”, empty quote fields, therapist echoes removed even when the patient had said the words), (ii) the *high* reasoning-effort setting for the section writers ($+0.031$), whereas the *highest* setting cost -0.010 relative to the default while tripling output tokens, and (iii) moving the agenda, prediction bar and the most recent risk signal into code ($+0.030$). Three plausible additions did not help: a classifier-built salience list per section with a coverage-repair pass (v26), rubric-guided self-revision of the assembled brief (v27), and both of them combined with high effort (v26h, v27h), which were the worst versions of the round. A code-built list of *all* risk signals since the last assessment (v29) pulled in thoughts voiced during ABC exercises and an emotion rating (“hopelessness, six”), which the supervisor flagged as noise (R29: 3.40); removing the list again (v30) did not recover v28, which suggests an interaction with the second v29 change (forcing two “others” beliefs onto the screen).

6.5 Best-of-3 selection

Run-to-run differences within one version were often larger than differences between versions: at a single checkpoint the spread between runs reached 0.17 (v26, Anita S12) and 0.13 (v28, Anita S19). Selecting one of three runs with the reference-free verifier improved the checklist score for seven of eight versions (Figure 7); for v28 from 0.846 to **0.879** , close to the oracle 0.885 . The exception is v30 ($0.791 \rightarrow 0.775$), where the verifier preferred a run that the checklist ranked last. Because the verifier and the checklist grader are the same classifier, part of the gain may reflect shared preferences; the oracle gap (0.006 for v28) bounds what any selector could add. The brief selected by the verifier in R29 (v29r3) scored 3.15 with the supervisor, so verifier selection did not translate into supervisor preference in that round.

6.6 Error analysis

Figure 8 pools the three checkpoints and counts the checklist weight failed in all three runs. From

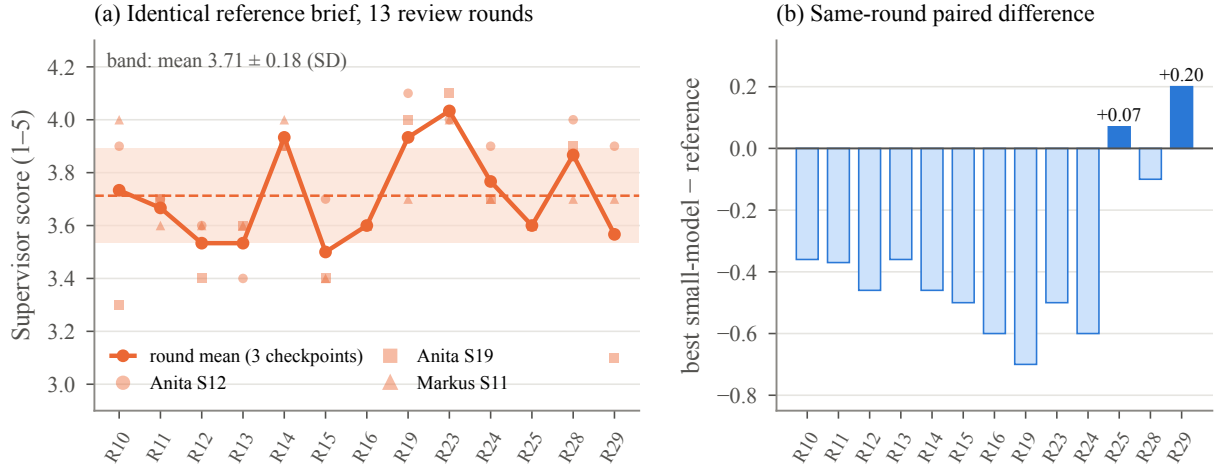


Figure 5: (a) Blind supervisor score of the *identical* frontier reference brief in 13 review rounds (same prompt, same files): round means (line), per-checkpoint scores (markers), overall mean $\pm$ SD (band). (b) Same-round paired difference between the best small-model version of that round and the reference: negative in all ten rounds R10–R24, then +0.07, −0.10, +0.20.

Version	Change (cumulative unless stated)	Checklist	SD (runs)	Supervisor, same round
v22	baseline after 22 iterations	0.777	0.012	3.53 (R23), 3.00 (R24)
v25	fix four deterministic-code defects	0.785	0.044	—
v25h	+ reasoning effort <i>high</i>	0.816	0.009	3.67 (R25), 3.57 (R28)
v25x	+ reasoning effort <i>highest</i> instead	0.775	0.022	—
v26	+ per-section salience list and coverage repair (classifier)	0.791	0.025	3.40 (R25)
v27	+ rubric-guided self-revision	0.787	0.028	3.47 (R25)
v26h / v27h	v26 / v27 with high effort (branches)	0.738 / 0.763	0.037 / 0.018	—
v28	v25h + code agenda (Beck order), prediction bar, latest risk signal, mood dips, therapist-first flag	0.846	0.031	3.77 (R28), 3.77 (R29)
v29	+ full code-built risk list, two “others” beliefs	0.836	0.006	3.40 (R29)
v30	v29 without the full risk list	0.791	0.022	—

Table 3: Ablation summary. Checklist = mean of three runs over three checkpoints; SD = standard deviation of the three run means. Earlier rounds (not shown) found the same for a code-imposed timeline of belief states (v15) and for fusing three drafts into one (v17): neither improved on v14 (0.700) on the checklist (0.708, 0.696) and both lost in blind review (R14 3.47 $\rightarrow$ R15 3.00, R16 3.00).

v22 to v28 the persistent failures fell from 25 to 15 of 164 weight units: technical leaks and distortion items were eliminated, prediction failures halved (duplicate predictions remain at Anita S19), and core-belief failures fell from 13 to 10. The remaining v28 failures are concentrated in the subtlest judgments of the cognitive model: the state of a belief about others that is reactivated in S12 towards the therapist (marked “weakened”), belief ranges that stop before the latest material, the qualification that Markus’s helplessness concerns relationships rather than work, and listing recent aggressive impulses as verbalised but not acted on. v29 regressed on distortions, rules and violence. The reference fails 36 weight units in its single run, mostly on the today agenda (8 of 19) and core beliefs (17 of 35), and it is the only system that

misses risk and last-session items; its single-run score should be read with the run-to-run spread of §6.5 in mind.

6.7 Cost

Table 4 gives tokens and list-price cost from the call logs. A single v28 brief costs about USD 0.034 in section writing and takes about 2.5–3 minutes of wall time with the five writers in parallel; a best-of-3 brief about USD 0.11. The per-session layer costs about USD 0.07 per new session, most of it the classifier scan of every sentence (about 0.9M input tokens per session, output free). Figure 9 relates cost to quality: the high reasoning-effort setting roughly doubles the cost of a brief and pays for itself; the highest setting costs a further 50% and does not. At Data Zone list prices

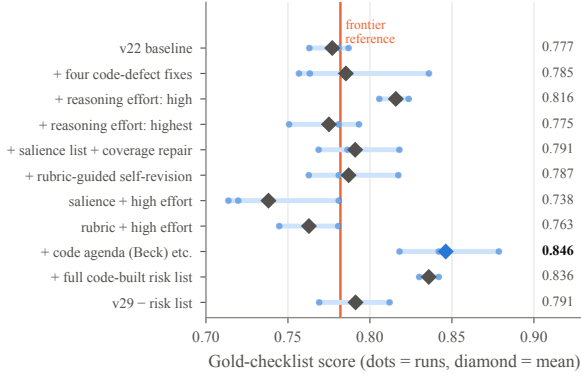


Figure 6: Ablation on the gold checklist (rounds 25–30). Dots are the three runs (mean over checkpoints), the diamond the mean; the vertical line is the frontier reference. Rows are cumulative from v22 unless stated; “salience + high effort” and “rubric + high effort” are branches.

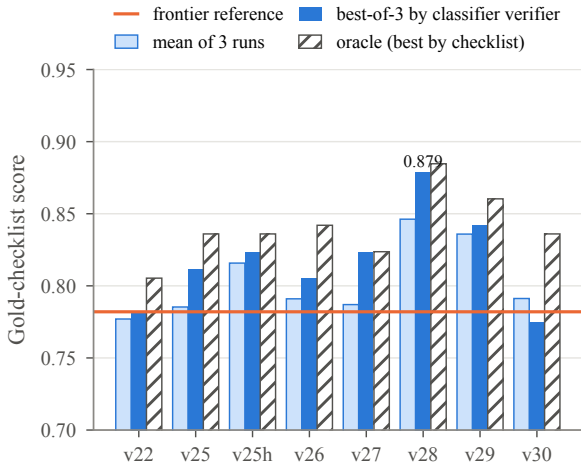


Figure 7: Mean of three runs, best-of-3 selected by the reference-free classifier verifier, and the oracle (best run by the gold checklist), per version.

(USD 5.50 / 33.00 per million tokens for GPT-5.5 against 0.12 / 0.60 for the small LLM) the reference brief costs about 19 times a single v28 brief and about 6 times a best-of-3 brief; per new session plus brief the ratio is about 6 (USD 1.02 vs 0.18). The verifier adds about 6.5k classifier input tokens per call, under USD 0.001 per brief. The whole development effort of rounds 25–30 (eleven versions $\times$ three runs $\times$ three checkpoints, plus scans) cost USD 2.87 on the small LLM; the entire project consumed 14.4M input and 6.4M output small-LLM tokens (USD 5.57). The blind supervisor ran on a separate subscription and is not included.

core beliefs (35)	13/35	9/35	10/35	10/35	17/35
rules (11)	1/11	3/11	1/11	3/11	2/11
distortions (14)	2/14	–	–	2/14	–
ABC / circle (6)	–	–	–	–	–
predictions (13)	4/13	4/13	2/13	2/13	2/13
self-risk (29)	–	–	–	–	2/29
violence (9)	3/9	1/9	2/9	3/9	2/9
last session (11)	–	–	–	–	3/11
today agenda (19)	–	2/19	–	–	8/19
other (10)	2/10	–	–	–	–
CBT purity (7)	–	–	–	–	–
	v22	v25h	v28	v29	GPT-5.5 (1 run)

Figure 8: Checklist weight failed in all three runs, by content area (area weight in parentheses; three checkpoints pooled, total weight 164). Reference: items failed in its single run.

Step (v28)	In	Out	USD
<i>Once per new session</i>			
$X + K + H$ (small LLM)	36–43k	11–13k	0.012
Sentence scan, labels + salience (classifier)	1.35M	—	0.057
<i>Per brief</i>			
Section writers, high effort (small LLM)	56k	45k	0.034
Best-of-3: two extra runs (as above)	112k	90k	0.07
Verifier, 3 calls (classifier)	20k	—	<0.001
Total, one new session + best-of-3 brief			≈0.18
<i>Frontier reference (GPT-5.5, v9), Data Zone list price</i>			
Extraction, per session	19k	8k	0.37
Brief, per checkpoint (no retries)	58k	10k	0.65

Table 4: Tokens and list-price cost from call logs (means over runs and checkpoints). Of the 45k output tokens per v28 brief, about 38k are reasoning tokens.

6.8 Example

Table 5 shows part of a v28 screen. The two agenda items that the supervisor had repeatedly flagged as missing in earlier versions (the planned measurement with PHQ-9 item 9; closure before the summer break) are now placed first by the code contract, and the prediction bar lists each test once with the session of its outcome.

7 Discussion

Why the division of labour works. The decisive changes were not better prompts but a re-assignment of responsibilities. A small model asked to count sessions, keep chains of predictions straight, order an agenda and quote verbatim

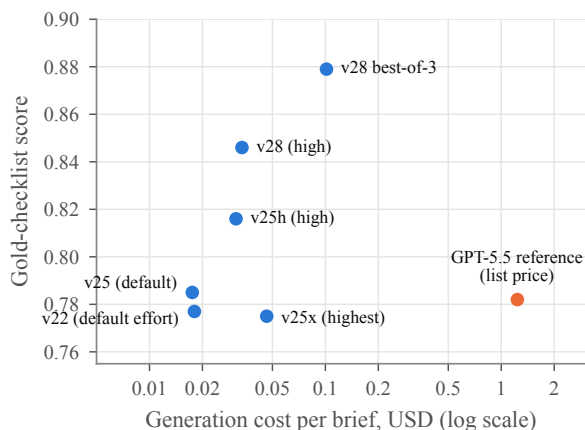


Figure 9: Generation cost per brief (section writers only; log scale) against gold-checklist score. GPT-5.5 includes schema retries, at list price.

Today: (1) Administer BDI-II, GAD-7 and PHQ-9 (compare with baseline); discuss PHQ-9 item 9 — last direct risk assessment: S2. (2) Closure before the break: summary of change, plan for the break (warning signs, contacts), return date. (3) Review homework from S19: a short note in the evidence log about the conversation with Bartek (done? what was noticed?).

Predictions: Bartek 70 % $\times$ (S9) $\cdot$ therapist 80 % $\times$ (S13) $\cdot$ brother 65 % $\times$ (S14) $\cdot$ Bartek 60 % $\times$ (S16) $\cdot$ father 90 % $\checkmark$ (S18).

Core beliefs: “I can’t cope when difficulties appear” — $n = 12$, S1–S19, recurring; “Those close to me withdraw in the face of my weakness” — $n = 9$, S5–S19, weakened.

Numbers: mood first 4.5/10 (S2); peak 6.5 (S9, S17); dips 2–2.5 (S13) and 3.5 (S10, S18); last 5.5 (S19). Baseline BDI-II 24, GAD-7 13, PHQ-9 16 (S2); no repeat measure since S2.

Table 5: Excerpt from the v28 screen for Anita before session 20 (run 1; translated from Polish). The agenda items (1)–(3) and the prediction bar are built by code; the belief lines are written by the small LLM from code-prepared inputs. The full screen has 768 words.

across 130,000 words fails in ways that are individually small and collectively disqualifying; code does all of this exactly, and once a defect in that code is found it stays fixed. Judgments that need no generation—is this sentence a literal statement about death or a metaphor? is it said now or recalled from childhood? how salient is it for the next session?—are well suited to a classifier that returns probabilities, costs USD 0.042 per million input tokens and can be applied to every sentence once. What remains for the small model is short writing from well-prepared, session-numbered material, where its weaknesses matter least. The pattern is not specific to CBT: any long-input summarisation task with countable structure, verifiable evidence and a known output order admits the

same split.

Why extra generation steps hurt. Fusing three drafts (v17), a code-imposed belief timeline (v15), salience hints with coverage repair (v26), rubric-guided self-revision (v27) and the two combined with high effort (v26h, v27h) all failed to improve on their base, several of them markedly. The common thread is that each adds a generation step in which a small model can lose or distort information, and that a small model revising its own output without new external evidence does not reliably improve it (Huang et al., 2024). By contrast, every change that *removed* work from the model (code agenda, code prediction bar) or that *selected* among outputs without rewriting them (best-of-3) was safe. The reasoning-effort result fits the same reading: more internal deliberation helped up to a point, after which longer reasoning traces (124k vs 64k reasoning tokens per two briefs) coincided with lower scores.

The judge is part of the experiment. The 13-round record of an identical brief (SD 0.18, range 0.53) is, to our knowledge, a rare direct measurement of LLM-judge drift on a long, structured input, and its size is sobering: it equals the improvement the whole loop produced on the same scale. Earlier rounds of this project drew conclusions from cross-round comparisons that the later record showed to be noise. The loop as finally run therefore (a) compared only within a round, (b) used the deterministic checklist for accept/reject and the judge for veto and diagnosis, and (c) generated three runs per version. We also note a subtler threat: in R29 the supervisor inferred “two families” of briefs from structural fingerprints (a template phrase, a field name), so blinding by shuffling alone is partial; neutralising surface differences between systems before review would strengthen the protocol.

What stayed hard. Risk lines were the hardest component. Both under-reporting (a recent signal missing from the screen) and over-reporting (a metaphor, an exercise, an emotion rating turned into a signal) were penalised, and rule-based lists over-reported. The best design kept only the most recent literal signal and the date of the last assessment and left ambiguity to the reader. The remaining checklist failures are the subtlest judgments of the cognitive model (belief state after a reactivation; scope qualifications), where the classifier and

the small model disagree with the gold reference in the same direction.

8 Limitations

- **Testbed size.** Two synthetic patients, three checkpoints and three runs per version support descriptive statistics only; transfer to other tasks and inputs is argued, not shown.
- **Development on the evaluation set.** All versions were tuned against the same checkpoints and judges, so the final scores are optimistic and partly reflect adaptation to the checklist wording and to the supervisor’s preferences. Mitigations in the loop: a reference-free verifier for selection, a fixed reference in every review round and a second, independent metric; a held-out patient and checkpoints generated after freezing the pipeline would be the proper test.
- **Judge noise.** The supervisor’s score for an unchanged output moved by up to 0.53 across rounds and by up to 0.46 (SD) on single criteria; the reported supervisor differences between v28 and the reference are within this noise. Multiple judge samples, judge ensembles and pairwise rather than absolute scoring would shrink it.
- **Shared classifier.** The checklist grader, the sentence scanner and the best-of- N verifier are the same classifier; gains from selection may partly reflect shared preferences, bounded by the oracle gap.
- **Reference system.** The frontier reference is one earlier pipeline variant with GPT-5.5, graded once; it is not an optimised frontier pipeline, and its own run-to-run variance is unknown.
- **Model drift.** Provider deployments change; the small-model name and the frontier price are marked for confirmation. Prompts, code, checklists, gold references, round logs and the synthetic transcripts are available from the authors on request for research use.

9 Conclusion

An automated, model-evaluated development loop brought a low-cost language model—combined with a sentence classifier for fine judgments and deterministic code for everything that can be counted—level with a frontier reference on content coverage (0.846 vs 0.782; 0.879 with best-of-3) and within judge noise of it in blind review, at a few US cents per output. The gains came

from the division of labour, from fixing deterministic code and from selecting among runs, not from adding generation steps; and the judge’s own variance proved large enough to be a result in itself. The loop and the pattern are general; the CBT brief was the testbed.

Ethics Statement

Only synthetic transcripts were used; no human participants, patients or clinicians were involved, and no human raters took part in the evaluation by design. Transcripts were processed only on the organisation’s own cloud deployments. Third-party datasets consulted during internal exploration are not reported here. The authors develop the product in which this pipeline may be used (conflict of interest). An AI assistant (Claude, Anthropic) was used for coding, analysis and drafting; the authors are responsible for the content.

References

- Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K. Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. [Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study](#). *JMIR Mental Health*, 11:e57306.
- Judith S. Beck. 2020. *Cognitive Behavior Therapy: Basics and Beyond*, 3rd edition. Guilford Press, New York.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. arXiv:2407.21787.
- Matilde Buattini, Donald Barjami, Lorenza Paponetti, Dalila Torres, Rosita Borlimi, and Gabriele Caselli. 2026. [Evaluating strengths, limitations, and future directions of ChatGPT in psychological analysis within case conceptualization: A qualitative analysis](#). *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 20(1):article 4.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.
- Yinzhu Chen, Abdine Maiga, Hossein A. Rahmani, and Emine Yilmaz. 2026. Automated rubrics for reliable evaluation of medical dialogue systems. arXiv:2601.15161.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv:2110.14168*.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. *Chain-of-verification reduces hallucination in large language models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578.
- Xiaotang Du, Rohit Saxena, Laura Perez-Beltrachini, Pasquale Minervini, and Ivan Titov. 2025. Enhancing long document long form summarisation with self-planning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the ACL (IJCNLP-AACL 2025)*. *ArXiv:2512.17179*.
- Jack Gallifant, Majid Afshar, Saleem Ameen, et al. 2025. *The TRIPOD-LLM reporting guideline for studies using large language models*. *Nature Medicine*, 31:60–69.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pages 10764–10799.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.
- Andrej Karpathy. 2026. autoresearch: an agent loop that proposes, runs and keeps or discards one experiment at a time. Software, <https://github.com/karpathy/autoresearch>.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. *Lost in the middle: How language models use long contexts*. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv:2408.06292*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.
- Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, Jingyuan Zhao, Qin Chen, Ningning Zhou, Kai Chen, and Liang He. 2026. PsychEval: A multi-session and multi-therapy benchmark for high-realism AI psychological counselor. *arXiv:2601.01802*. To appear in *Findings of ACL 2026*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv:2405.01724*.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. PATIENT-Ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, FL. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.
- Lei Xu, Mohammed Asad Karim, Saket Dingliwal, and Aparna Elangovan. 2024. Salient information prompting to steer content in prompt-based abstractive summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 35–49. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

A Supervisor criteria

Each brief was scored 1–5 on: (1) *accuracy* of the conceptualisation against the gold reference

and transcripts; (2) *CBT levels* (automatic thought vs rule vs core belief; belief slots); (3) *session trace* (session lists, counts, ranges and chronology, checked against transcripts); (4) *quotes* (verbatim, spoken by the patient, supporting the item they are attached to); (5) *usefulness* for a three-minute pre-session read; (6) *CBT purity* (no schema-therapy or other-modality vocabulary on the screen); (7) *calibration* (hypotheses marked as such; risk and violence neither under- nor over-stated; belief states and prediction statuses consistent with the material).

B Version history

v1–v4: three initial variants (brief from transcripts; brief from per-session extractions; brief from an incrementally updated card) with frontier models; variant B (extractions) retained. v5–v9: code-assembled screen from candidates, prediction bar from data, quote gate, screening rule, echo filter; v9 with GPT-5.5 is the frozen reference. v9–v13 (small model): critic K and downward arrow H ; per-session writers. v14: sentence classifier added. v15–v16: belief timeline (dropped), ambiguity list for risk, technical scrub. v17: fusion of three drafts (dropped). v18–v21: deterministic screening, “others” belief call, recency items, agenda from the plan. v22: baseline of the ablation. v25–v30: Table 3. Blind review rounds: R8, R8.1, R9 (frontier models only) and R10–R16, R19, R23–R25, R28, R29 (13 rounds with the unchanged reference).

C Judge noise, per criterion

Round-to-round SD of the identical reference brief’s three-checkpoint mean, 13 rounds: accuracy 0.16, CBT levels 0.32, session trace 0.19, quotes 0.46, usefulness 0.16, CBT purity 0.46, calibration 0.26 (overall 0.18). Means: 3.90, 3.13, 3.00, 4.23, 3.69, 4.49, 3.51.