DO LLMS FOLLOW A BEHAVIORAL SCRIPT? MEASURING STRUCTURAL CHOREOGRAPHY AND RIGIDITY FROM POSITIONS AND TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Individual-response evaluations can rate an answer highly while failing to assess whether a model consistently organizes its behaviour in the same way across users. To measure this, we introduce SCRIPT, the **S**tructural **C**horeography and **R**igidity **I**ndex from **P**ositions and **T**ransitions, a reference-free measure of behavioural scriptedness from span annotations. SCRIPT separates positional choreography from sequential momentum and calibrates both against a within-response permutation baseline that preserves behaviour frequency while disrupting arrangement. We evaluate it on 4,095 responses to 819 counselling prompts from five frontier models with matched therapist responses, and on 22 controlled prompting and training configurations. All five models exhibit recurring structure above the permutation baseline (SCRIPT $= 0.083$–0.118), while the matched therapist shows lower positional choreography under the same annotation scheme. For the models, response position explains substantially more behavioural variation than coded user state or conversation stage, and models respond more strongly to explicit cues than to affective distress. At matched behaviour mix, existing quality evaluation shows no detectable sensitivity to arrangement, and reducing cross-turn tactic recurrence does not necessarily remove within-response choreography. These findings identify behavioural arrangement as a distinct evaluation dimension and provide a practical framework for measuring not only what models do, but where and in what sequence they repeatedly do it. We release all analyses and codes.

1 INTRODUCTION

As AI assistants increasingly interact with users in open-ended and safety-sensitive settings, evaluating their behaviour requires characterizing how model behaviours are organized across interactions Sharma et al. (2024); Sidnell (2016). A model may produce appropriate behaviours while repeatedly placing them in similar positions across different interactions. This regularity can have safety implications because the effect of a behaviour depends on its timing and surrounding context: reassurance before adequate information gathering, advice before clarification, or affirmation before identifying a potentially incorrect premise can each be appropriate in some contexts and problematic in others Gioia & Poole (1984).

Current evaluations provide limited insight into whether models consistently organize their behaviour in the same way across different inputs, as they primarily assess the quality and content of individual responses (Liu et al., 2023; Zheng et al., 2023). Prior work has identified reduced output diversity, recurring syntactic forms, empathy-tactic templates, and cross-turn tactic recurrence (Kirk et al., 2024; Wu et al., 2025; Shaib et al., 2024; Gueorguieva et al., 2026; Zhan et al., 2026), but these approaches do not directly measure behavioural arrangement itself.

Different responses can use different words and address different users while repeatedly placing the same behaviours in similar positions and sequences. A key challenge is that apparent structure may arise from behaviour frequency alone: a model that frequently gives advice may appear to place advice consistently even without a recurring arrangement. We therefore ask: *do LLMs follow recurring behavioural scripts, and how strongly do they follow them?* Answering this requires

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

separating which behaviours occur from where they occur and what follows what, providing a basis for studying context sensitivity, robustness, and potential failure modes.

We introduce SCRIPT, the **S**tructural **C**horeography and **R**igidity **I**ndex from **P**ositions and **T**ransitions, a reference-free measure of behavioural scriptedness from span annotations. SCRIPT represents each span as a behaviour–position event, separates recurring structure into positional choreography and sequential momentum, and calibrates both against a within-response permutation that preserves behaviour frequency while destroying arrangement. We evaluate SCRIPT across two testbeds, comprising $4,095$ responses to 819 counselling prompts from five frontier model families with matched therapist responses (Badawi et al., 2026), and 22 prompting and training configurations from an emotional-support benchmark (Zhan et al., 2026). We further validate the estimator using human corpora, span-annotation tasks, synthetic known-answer cases, and random-label controls, with annotations produced by human and automatic annotators (Kasner et al., 2025).

Our analysis yields three main findings. First, all five frontier models included in our investigation show recurring behavioural structure beyond what would be expected if the same behaviours were randomly rearranged within each response, while matched human therapist responses show lower positional choreography. Second, model behaviour is more strongly associated with where a behaviour occurs within a response than with the user's state or the stage of the conversation. Behavioural organization also changes more in response to explicit user cues than to differences in affective distress. Third, existing quality evaluations do not capture differences in behavioural arrangement once overall behaviour mix is controlled. In summary, we make three main contributions:

❶ **A New Measure of Behavioural Scriptedness.** We introduce SCRIPT to quantify *behavioural scriptedness*, the extent to which a system repeatedly follows a similar behavioural structure across its responses. SCRIPT compares these patterns with those obtained by randomly reordering behaviours within each response, without changing how often each behaviour occurs.

❷ **SCRIPT-based Evaluation.** A framework that combines script profiles for system identification with SCRIPT-Seq for anytime-valid sequential detection of behavioural scriptedness. We validate SCRIPT across models, human responses, interventions, annotation schemes, and controls, while its general formulation supports application across domains and annotation schemes.

❸ **A Comparative Evaluation Suite for Behavioural Scriptedness.** An evaluation suite spanning five frontier models to examine how model behaviour relates to response position, coded user state, and conversation stage. The suite also tests whether standard quality assessments capture recurring behavioural structure after controlling for behaviour frequency, enabling researchers to evaluate response organization and contextual responsiveness beyond the quality of individual responses.

2 RELATED WORK

Behaviour-level evaluation. Existing behavioural evaluations primarily measure *how much* of a behaviour a model produces. Sycophancy is defined as the extent to which models accommodate a user's stated view (Sharma et al., 2024; Cheng et al., 2026; Sambara et al., 2026). In mental health support settings, prior work has examined psychotherapy techniques, empathy mechanisms, and inappropriate responses (Chiu et al., 2024; Sharma et al., 2020). Multi-turn studies measure how behaviour changes across a conversation, including instruction drift and failure to recover from early assumptions (Li et al., 2024; Laban et al., 2026). Kim et al. (2026) similarly represent reasoning traces as conversational moves and show that reinforcement learning changes the frequency.

Templated behaviour in support dialogue. Closest to our setting, Gueorguieva et al. (2026) show that LLM support responses follow recurring empathy-tactic patterns, while Zhan et al. (2026) measure cross-turn tactic *stickiness* and train an intervention that reduces it. These studies establish that support dialogue can be templatic and that such regularity may matter to users. Their measurements focus on a hand-specified tactic pattern or on recurrence across turns: the former must be rediscovered for a new coding scheme and has no matched chance baseline, while the latter does not capture where behaviours occur within a response.

Behavioural fingerprints and coded sequences. Models can be identified from recurring properties of their outputs. Prior work uses text embeddings (Sun et al., 2025; Fu et al., 2025), behavioural trait vectors (Pei et al., 2025), and discourse-level features (Russell et al., 2026); internal represen-

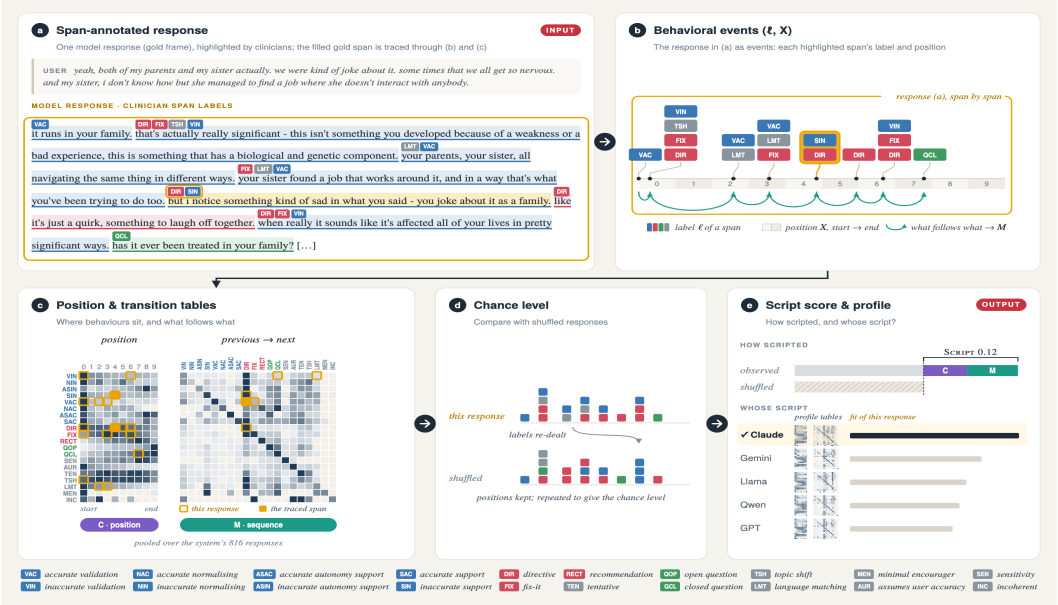


Figure 1: **Overview of SCRIPT. (a)** A span-annotated response. **(b)** Each span becomes an event (ℓ, X): its label and the position bin of its start. **(c)** Pooled over the system's responses, the events fill a position table $\hat{P}(\ell \mid X)$ and a transition table $\hat{T}(\ell_{\text{prev}}, \ell)$, which give choreography C and momentum M as shares of $H(L)$. **(d)** Within each response, the labels are re-dealt over the fixed positions; SCRIPT is the observed rigidity in excess of the shuffle mean, $C_{\text{ex}} + M_{\text{ex}}$. **(e)** The system's script profile; scoring a new response under each enrolled profile names the system that produced it.

tations provide another route when model access is available (Chen et al., 2025; Sofroniew et al., 2026). More broadly, dialogue-act tagging and sequence analysis have long represented discourse as ordered coded moves (Stolcke et al., 2000; Abbott, 1995). Span annotation provides a compatible observational layer and can be produced by either human or LLM annotators (Kasner et al., 2025).

Prior work establishes three foundations: coded behaviours as units of analysis, evidence that supports responses exhibit recurring templates, and behavioural signatures that distinguish models. What remains missing is a general measure of the template itself. Existing approaches quantify behaviour frequency or pattern matches, but do not measure arrangement relative to positions from sequential regularity. §3 SCRIPT addresses this gap and recovers the underlying behavioural profile.

3 SCRIPT (STRUCTURAL CHOREOGRAPHY & RIGIDITY INDEX FROM POSITIONS AND TRANSITIONS)

3.1 PRELIMINARIES

Consider a system that generates responses for many users. Each response is annotated with behavioural labels $\ell \in \mathcal{L}$ and their positions, whose ordered sequence defines the response's *behavioural arrangement*. Let L denote a behavioural label, X its position, and L_{prev} the preceding label. If the system repeatedly follows a similar behavioural arrangement, then X and L_{prev} should provide information about L beyond that contained in the system's overall behavioural frequencies. Our target quantity is therefore the share of uncertainty in L associated with *where a behaviour occurs* and *what behaviour preceded it*, beyond what is already implied by the system's overall behaviour mix. Such a measure should assign no scriptedness to randomly ordered behaviours, keep *where* behaviours occur separate from *what follows what*, and identify which behaviours contribute to the observed script rather than reducing it to a single aggregate score.

We observe N system-generated responses, $\{r_1, \ldots, r_N\}$. Within each response, annotated behavioural spans are represented as events defined by their label $L \in \mathcal{L}$ and position. SCRIPT op-

erates solely on these events and does not require the response text, model probabilities, reference responses, or quality judgments. We compute SCRIPT at the system level by pooling events across all responses generated by the same system. The resulting count tables (§3.4) and plug-in entropy estimates are therefore system-specific (Paninski, 2003). Null permutations (§3.3) are performed within individual responses and never across systems. Responses containing a single annotated span contribute to the label and position distributions but contain no transitions.

We first map annotated responses to behavioural events (§3.2), then quantify positional and sequential structure (SCRIPT score) and calibrate the resulting score against a permutation null (§3.3). We subsequently use the same representation to examine context dependence, script profiles, conversation-scale structure, and sequential detection.

3.2 BEHAVIOURAL EVENT REPRESENTATION

Span events. Each annotated span (Kasner et al., 2025) defines an event (ℓ, x) with behavorial label $\ell \in \mathcal{L}$ and its relative start position $x = p/|r| \in [0, 1)$. Here, p is the character offset at which the annotated span begins and $|r|$ is the response length in characters. Both are measured after applying the same normalization procedure *(Unicode NFKC, straight quotes, collapsed whitespace, lower case)*. We use only the *start* of each span and do not consider its extent. We discretize relative positions into B equal-width bins, $X = \min(\lfloor xB \rfloor, B-1)$. Appendix C.3 describes the selection of B and evaluates sensitivity to this choice. We also show alternative span positions using the midpoint or end of each span and evaluate whether accounting for span length changes the results.

Slots and transitions. We order events in each response by position (Stolcke et al., 2000). If the same text span receives multiple behavioural labels, we group them into a multi-label *slot* and do not treat them as a sequence because their order cannot be determined from the response. We define a *transition* between behavioural events at consecutive, strictly increasing positions, so no transition is formed within a slot. The first event in a slot, L_{prev} is defined by the preceding slot in the same response. The remaining events in the slot do not contribute to transitions. Appendix C.1 reports the prevalence of co-located labels and evaluates an alternative convention that imposes a fixed ordering.

3.3 INFORMATION-THEORETIC DECOMPOSITION OF RESPONSE STRUCTURE

To quantify positional and sequential regularity we call C *choreography* the share of behavioural uncertainty associated with where behaviours occur within a response, and M *momentum* the additional share associated with the preceding behaviour after accounting for position. Conditioning M on X is essential; otherwise, a behaviour that always occupies the middle of a response would also appear sequentially predictable simply because of its position. Thus C captures *fixed seats*, while M captures *chained moves* beyond those seats. Their sum R measures the system's structural rigidity. Mutual information provides a nonparametric measure of both components in a common information-theoretic unit and supports their additive decomposition. Because the first slot of a response has no predecessor, M is estimated only for events with a preceding slot. Appendix C.3 examines higher-order dependencies, while Appendix C.1 describes the boundary convention.

$$C = \frac{I(L; X)}{H(L)}, \qquad M = \frac{I(L; L_{\text{prev}} \mid X)}{H(L)}, \qquad R = C + M. \tag{1}$$

Raw mutual information can be positive in finite samples even when no systematic structure is present (Panzeri et al., 2007). A skewed behavioural distribution or sparse count table can therefore yield a positive R even when behaviours are not systematically arranged. We construct a permutation null representing the expected rigidity when the same behaviours occur with the same frequencies but without systematic positional or sequential structure. Within each response, we keep the observed positions and label multiset fixed and randomly reassign labels across positions. This preserves which behaviours occur, how often they occur, the event count, and the response's slot structure, while destroying only their arrangement, the controlled-shuffle logic of O'Connor & Andreas (2021). For observed rigidity R_{obs} and shuffle values $R^{(1)}, \ldots, R^{(S)}$, we define

$$\bar{R}_{\text{null}} = \frac{1}{S} \sum_{s=1}^{S} R^{(s)}, \qquad \text{SCRIPT} = R_{\text{obs}} - \bar{R}_{\text{null}}, \qquad z = \frac{R_{\text{obs}} - \bar{R}_{\text{null}}}{\text{sd}(R^{(1)}, \ldots, R^{(S)})}. \tag{2}$$

A SCRIPT value of zero, therefore, means that the observed arrangement of behaviours is no more predictable than expected after counting for the system's own behavioural composition and sample

size(Hewitt & Liang, 2019). We apply the same calibration separately to C and M, preserving the additive decomposition: $\text{SCRIPT} = C_{\text{excess}} + M_{\text{excess}}$.

Because the maximum achievable rigidity depends on the annotation scheme, raw SCRIPT values are not directly comparable across schemes with different label sets or event densities. For cross-scheme interpretation, we therefore construct a *matched ceiling*. We retain each response's positions and label multiset and arrange each labels according to a common corpus-wide ordering. The observed score divided by this ceiling gives the *fraction of achievable rigidity*. We compare raw SCRIPT values directly only within the same annotation scheme and annotator, and use the ceiling-normalized fraction for comparison across schemes.

3.4 CONDITIONAL AND CROSS-RESPONSE STRUCTURE

Context-Conditional Structure A response structure does not necessarily imply that the system is insensitive to the user. When each response is paired with a coded user state U and conversation turn τ, we quantify the association between behavioural labels and these variables using

$$\Pi_U = \frac{I(L;U)}{H(L)}, \qquad \Pi_\tau = \frac{I(L;\tau)}{H(L)}, \tag{3}$$

Here, Π_U measures the share of behavioural uncertainty associated with variation in user state, while Π_τ measures the share associated with conversation progression. We calibrate each quantity against a permutation null that reassigns whole responses across user states or turns while preserving the behavioural structure within each response (Good, 2005). Although Π_τ has the same form as the conversation choreography $C^{\rightarrow}$, the two differ in their nulls: Π_τ keeps each response intact and reassigns it across turns, whereas $C^{\rightarrow}$ shuffles labels among the events of one conversation. We compare these quantities with choreography C to distinguish structure associated with position within a response from variation associated with the user. We do not combine these quantities into the SCRIPT score because, unlike C and M, they do not form part of its chain-rule decomposition.

Positional and Transition Profiles The counts underlying C and M also characterize the specific structure captured by the scalar score. A position table $P \in \mathbb{N}^{|\mathcal{L}|\times B}$ records the distribution of each behaviour across response positions, while a transition table $T \in \mathbb{N}^{|\mathcal{L}|\times|\mathcal{L}|}$ records how frequently behaviours follow one another. Together, these tables form the *script profile*. While the SCRIPT score quantifies the strength of recurring structure, the profile describes its positional and sequential pattern. We further test whether script profiles are characteristic of individual models. We construct a profile for each model from one set of responses and use it to attribute held-out responses based only on their behavioural labels, positions, and transitions (Sun et al., 2025). The matching procedure, smoothing choices, cross-corpus and annotator transfer protocols are in Appendix C.4.

Cross-Response Dependence For multi-turn data, we apply the same principle across responses. *Conversation choreography*, $C^{\rightarrow} = I(L;\tau)/H(L)$, asks whether behavioural labels are systematically associated with turn index τ, while *boundary momentum*, $M^{\rightarrow} = I(F;Z)/H(F)$, asks whether the opening behaviour F of a response is associated with the final behaviour Z of the preceding response. We evaluate both measures against within-conversation permutation nulls that preserve the behaviours present in each conversation while disrupting their turn-level arrangement. This distinguishes structure within individual responses from behavioural dependencies that extend across turns (Li et al., 2024; Laban et al., 2026).

Sequential Certification Finally, SCRIPT-Seq provides an anytime-valid sequential test when annotated responses arrive over time. Repeatedly evaluating a fixed-sample z statistic and stopping when it crosses a threshold does not control the false-rejection rate. Instead, SCRIPT-Seq evaluates each new response against a profile constructed only from earlier responses, compares its conformity with within-response permutations, and accumulates the resulting evidence using a non-negative betting process. By Ville's inequality, the test remains valid regardless of when the analysis is inspected or stopped. More details are presented in Appendix C.5.

4 EXPERIMENTAL SETUP

We apply SCRIPT to two multi-system testbeds and a set of anchor corpora. Testbed 1 (Badawi et al., 2026) compares independently developed frontier models, whereas Testbed 2 (Zhan et al., 2026) examines controlled prompting and training interventions.

Table 1: The two testbeds used to evaluate SCRIPT.

Testbed	Domain	Systems	Primary annotation	Replies
Testbed 1: Cognitive Atrophy	Counseling	5 frontier LLMs	20 clinician codes	4,095
Testbed 2: Lend an Ear	Emotional support	22 interventions	10 tactic labels	6,922

Table 2: **Testbed 1 SCRIPT scores under two annotation layers.** The clinician layer covers the five models; the blind LLM layer codes the same models and the human therapist under one shared 20-code scheme. All scores are in excess of the within-response shuffle null and are comparable within an annotator. ΔC is the paired model–therapist difference in choreography over the 789 shared prompts, with 95% bootstrap intervals. Because the therapist responses are below the reliable operating range for M, the human–model comparison is made on C. Bold marks the only model separation surviving Holm correction; the final row is the random-label control.

	Clinician annotator (models only)				Blind LLM annotator (all six speakers)				
Speaker	SCRIPT	C	M	z	SCRIPT	C	M	z	ΔC vs. therapist [95% CI]
Qwen	0.091	0.052	0.040	28.5	0.085	0.063	0.022	52.4	$+0.019$ [$+0.007$, $+0.023$]
Llama	0.083	0.063	0.020	22.1	0.118	0.086	0.032	61.3	$+0.041$ [$+0.029$, $+0.046$]
GPT	0.098	0.060	0.038	34.9	0.108	0.083	0.025	73.4	$+0.038$ [$+0.026$, $+0.042$]
Claude	**0.118**	0.055	**0.064**	43.9	0.114	0.063	0.050	69.0	$+0.018$ [$+0.005$, $+0.024$]
Gemini	0.092	0.057	0.035	28.9	0.095	0.063	0.032	56.9	$+0.018$ [$+0.006$, $+0.023$]
Human therapist	—	—	—	—	0.053	0.045	0.008	12.8	
Random-label control	-0.004	-0.001	-0.003	-1.8	-0.000	-0.000	-0.000	-0.4	

4.1 TESTBEDS

Testbed 1: Cognitive Atrophy Benchmark The Cognitive Atrophy Benchmark (Badawi et al., 2026) provides the primary discovery setting. It contains 819 counseling items from four corpora, including two multi-turn datasets (Table A.2), with each item answered independently by five frontier model families, yielding 4,095 responses.The systems are `Qwen3.5-35B-A3B`, `Llama-4-17B-128E-Instruct`, `GPT-5.3-chat`, `Claude Sonnet 4.6`, and `Gemini 3 Flash`. Responses are annotated using a fixed 20-code span manual, yielding 39,696 clinician-labelled behavioural events. We use the original therapist responses as a behavioural anchor, not as a gold answer; the multi-turn corpora also provide user state and turn information to test whether response structure varies with the user and conversation stage (§4.2 and §5.2).

Because the original clinician spans cover model responses but not the therapist responses, we analyze Testbed 1 under three annotation layers. The *clinician layer* uses the benchmark's original 20-code expert annotations. A shared *surface layer* applies the same simple sentence-level instrument to all five models and therapist, providing a direct human-model comparison. Finally, a blind *LLM layer* uses `DeepSeek-V4-Pro` to annotate all model and therapist responses under the same 20-code manual, with responses shuffled and speaker identity withheld (Appendix A.3).

Testbed 2: the Lend an Ear Ladder Lend an Ear (Zhan et al., 2026) contains outputs from 22 systems, eleven prompting or training configurations applied at two model sizes, on the same 315 turns from multi-turn emotional-support conversations. In contrast to Testbed 1, the task, conversation histories, ten-tactic annotation scheme, automatic tagger, and expert-protocol quality judge are held fixed while the intervention varies. This controlled setting allows us to test whether changes in conventional quality or cross-turn diversity correspond to changes in within-response scriptedness. Because momentum is not reliably estimable from the released multi-label tactic annotations, we report the positional term C and use a one-label-per-sentence surface annotation when the full $C + M$ decomposition is required (Appendix B.3.1).

4.2 ANCHORS AND VALIDATION

Human anchors and cross-domain controls. To interpret the scale beyond the two testbeds, we construct a *ruler* from independent annotation schemes (Appendix A.4). Human behaviour is anchored by motivational-interviewing counsellors (Wu et al., 2022), psychology-trained support writers (Gueorguieva et al., 2026), real human supporters (Zhan et al., 2026), and expert-coded news

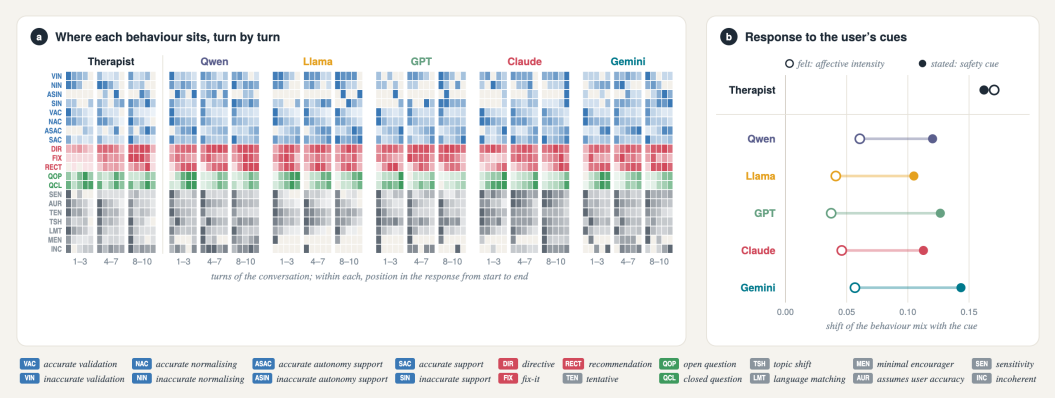


Figure 2: The blueprint turn by turn and what moves it. **(a)** Where each of codes occurs in response at turns of the conversation, on multi-turn corpora under one blind annotator. **(b)** Each speaker's shift of behaviour mix with affective intensity of user's message and with an explicit safety cue.

authors (Da San Martino et al., 2019); the latter's 18-code scheme also tests whether a fine-grained expert ontology alone creates apparent rigidity. Machine-translation error spans (Kasner et al., 2025) provide an expected real-data null, while RAGTruth (Niu et al., 2024) and data-to-text errors (Kasner et al., 2025) extend the evaluation to unrelated tasks. Synthetic known-answer corpora and within-response random-label controls fix the scale's endpoints.

Validation protocol. Before interpreting scores on real systems, we run three checks. First, *known-answer validation* checks the score on synthetic data with known structure: a positional template should appear in C, a behavioural chain in M, and random arrangements should score zero. Second, *operating-range tests* identify how much data is needed for reliable estimates and how sensitive they are to design choices. Third, *annotator-independence tests* check that findings do not depend on who or what produced the labels. Full protocols are in Appendices C.2, C.1, and C.3.

5 Results

5.1 Models share recurring behavorial script

Across 4,095 responses to 819 prompts, all five Testbed 1 models exhibit behavioural scriptedness beyond the within-response permutation baseline (Table 2). Clinician-annotated Script scores range from 0.083 to 0.118 ($z = 22$-44). The Script profiles reveal a shared broad positional structure where *empathy*-related behaviours concentrate near the beginning, advice occupies the middle, and question-related behaviours occur later in four of the five models; Llama places the latter closer to advice (Fig. 2 (a)). Choreography is, therefore, similar across models, (0.052-0.063), whereas momentum ranges from (0.020-0.064). Claude has the largest momentum and is the only model with a higher total Script score than each of the other four under Holm-corrected paired bootstrap tests (Appendix B.5.1). Blind LLM annotation also detects structural regularity in all five models, with Script scores of 0.085-0.118 and $C = 0.063$-0.086.

Human therapist responses have a choreography score of 0.045, compared with 0.063-0.086 for the five models. On the same prompts and same annotation scheme, each model, therefore, shows higher positional choreography than the therapist responses, with model-therapist differences in C ranging from $\Delta C = +0.018$ to $+0.041$; with all five 95% intervals excluding zero (Table 2). Independent human comparisons using the same annotation instrument show the same direction (Appendix B). Finally, the control analysis confirms that this signal comes from arrangement rather than behaviour frequency alone: positional structure appears in C, sequential structure in M, and random or re-dealt labels return approximately zero (Table C.2; Fig. C.2).

5.2 The structure varies more with response position than with the user

We examine whether the behavioural variation in §5.1 is associated more with response position, user state, or conversation turn. We quantify these as C, Π_U, and Π_τ relative to their permutation

Table 3: **What changes the behaviour: the user, the conversation, or response position?** Therapist and five models on multi-turn corpora of Testbed 1 under the same 20-code annotation. Advice compares low versus high affective intensity and early versus late turns; Shift measures the corresponding change in behaviour mix; Explained share reports the contributions of turn (Π_τ), user state (Π_U), and response position (C). z-scores are in parentheses (Appendix B.2).

	Advice, % of moves		Shift of the mix, excess JS (z)		Explained share of $H(L)$, % (z)		
Speaker	affect low → high	turns 1–3 → 8–10	affective intensity	explicit safety cue	turn Π_τ	person Π_U	seat C
Qwen	48 → 47	41 → 51	0.061 (4.9)	0.120 (7.4)	0.40 (9.5)	0.63 (3.7)	6.0
Llama	35 → 40	37 → 35	0.041 (2.8)	0.105 (6.7)	0.11 (2.5)	0.52 (3.1)	8.5
GPT	59 → 54	52 → 59	0.037 (3.2)	0.126 (9.7)	0.33 (10.4)	0.83 (6.3)	8.1
Claude	27 → 31	27 → 31	0.046 (3.3)	0.113 (4.9)	0.50 (10.4)	0.84 (3.4)	6.0
Gemini	47 → 44	40 → 49	0.057 (4.2)	0.143 (8.2)	0.32 (6.4)	1.09 (5.9)	6.1
Human therapist	**29 → 13**	**8 → 28**	**0.170 (8.9)**	0.162 (6.9)	**1.34 (18.8)**	1.28 (4.5)	5.0

baselines (Table 3). For the five models, response position explains 6.0–8.5% of behavioural uncertainty, whereas the coded user state explains only 0.5-1.1%; for the therapist, the corresponding values are 5.0% and 1.28% (Table 3). Conversation turn also explains more of the therapist's behaviour than the models', accounting for 1.34% of the therapist's behavioural uncertainty but only 0.11–0.50% for the models. Thus, for the models, where a behaviour occurs within the response is a stronger source of regularity than either the coded user state or the stage of conversation.

The difference is visible in how behaviour changes with context. When affective intensity increases, the therapist reduces *advice* from 29% to 13% and asks more questions, whereas no model reduces *advice* by more than five percentage points (Table 3). The models respond more strongly to explicit cues: an explicit safety cue produces 65-88% of the therapist's behavioural shift, compared with only 22-36% for affective intensity (Fig. 2b). Across the conversation, the therapist shifts from questions toward *advice*, with *advice* rising from 8% in turns 1–3 to 28% in turns 8–10, while the models begin *advice*-heavy at 27-52% and change less (Fig. 2a; Table 3). Density-matched analyses show the same direction and are reported in Appendix B.2.

The recurring structure is also predominantly response-local. Choreography C remains similar when recomputed within early, middle, and late turn buckets, indicating that the same within-response positional pattern is repeated throughout the conversation. By contrast, boundary momentum $M^{\rightarrow}$ is indistinguishable from the null for every model ($|z| \leq 1.8$ across 646 boundaries), despite each previous response being available in context (Table B.9). These results show that models follows a similar behavioural script within each response and change their behaviour more for explicit user cues than for affective distress or conversation stage.

5.3 Existing evaluation does not capture the behavorial scriptedness

We next ask whether existing evaluation captures behavioural scriptedness as measured by Script, and whether reducing one form of templating also reduces within-response structure. Table 4 shows that existing quality evaluation and behavioural arrangement measure different properties. At matched behaviour mix, the Testbed 2 judge is insensitive to how those behaviours are arranged: across 25,919 pairs from the same system, the more script-conforming response is rated only $+0.004$ points higher on the five-point scale, with a 95% interval of $[-0.015, +0.022]$. The same matched groups provide a positive control: the longer response is rated $+0.039$ points higher $[+0.021, +0.058]$. Thus, the judge can detect differences between matched responses, but no comparable effect is detected for behavioural arrangement.

Across systems, quality and scriptedness can nevertheless rise together because interventions change the behaviour mix and its arrangement simultaneously. This is why the unmatched system-level correlation does not imply that the judge rewards the arrangement itself; once the behaviour mix is held fixed, the association disappears. The complete 22-system analysis and the corresponding Testbed 1 matched comparison are reported in Appendix B.3. The intervention results show the same distinction. At 4B, MINT reduces cross-turn stickiness from 0.571 to 0.422 and tagger-layer choreography from $C = 0.150$ to 0.112, but the resulting choreography remains above the human

Table 4: **What the judge sees and what SCRIPT sees.** Top: matched-behaviour-mix comparisons of quality against arrangement, with response length as a positive control. Bottom: selected 4B Testbed 2 systems comparing choreography C, cross-turn stickiness, quality rating, and event density. All 22 systems are reported in Table B.13.

Matched behaviour mix: does the judge respond to . . .

Contrast	groups	pairs	diff	95% CI
Testbed 2, arrangement (own blueprint)	1,011	25,919	$+0.004$	$[-0.015, +0.022]$
Testbed 2, arrangement (pooled blueprint)	1,011	25,910	0.000	$[-0.019, +0.019]$
Testbed 2, response length (positive control)	1,013	26,191	$+0.039$	$[+0.021, +0.058]$
Testbed 1, arrangement, ten clinician judgments	198	693	$\lvert\text{diff}\rvert \leq 0.06$	bounds ≤ 0.14

Testbed 2 at 4B: what each measure reads

System	Tagger C	Surface C	Stickiness	Rating	Events/turn
Human supporters	0.014	0.005	0.266	2.90	2.4
Verbalized sampling, vanilla prompt	0.050	0.036	0.395	3.51	4.6
Tactic-list prompt	0.110	0.104	0.515	3.88	6.7
Vanilla prompting	0.150	0.093	0.571	3.75	7.1
MINT (quality + KL novelty)	0.112	0.195	0.422	4.67	10.7
Quality-only RL	0.155	0.138	0.671	4.62	16.3

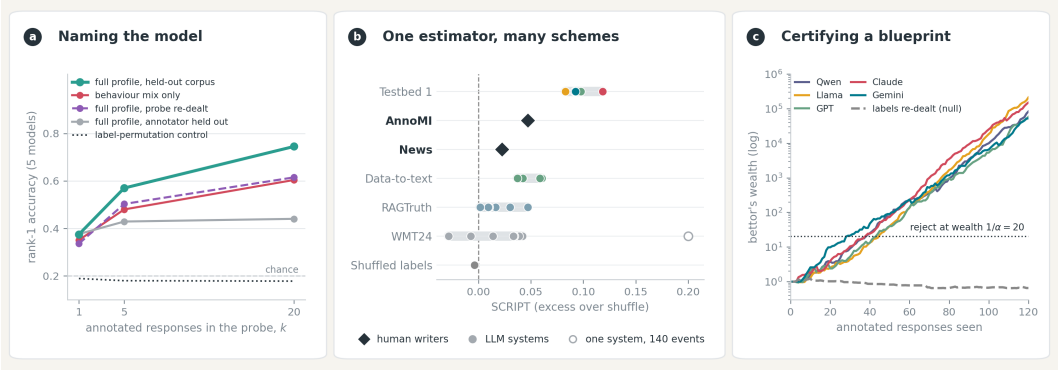


Figure 3: **Profile identification, cross-task checks, and sequential certification.** **(a)** Generating-model identification from held-out responses. **(b)** SCRIPT across the paper's anchor corpora and controls. **(c)** SCRIPT-Seq as annotated responses arrive.

level of 0.014 and close to the tactic-list prompt at 0.110 (Table 4). Under the position-blind surface layer, choreography instead increases from $C = 0.093$ to 0.195 after MINT.

Figure 3 summarizes three supporting checks: the script profile identifies the generating model from held-out responses, the estimator behaves as expected across independent tasks and controls, and SCRIPT-Seq certifies Testbed 1-level structure from a median of 34–43 annotated responses; full analyses are reported in Appendices B.4, A.4, and C.5.

6 CONCLUSION

In this work, we introduced SCRIPT, a reference-free measure of behavioural scriptedness that disentangles positional choreography from sequential momentum using span annotations and a matched-permutation baseline. Across 4,095 responses to 819 prompts, five frontier models show substantial recurring structure, with a shared positional pattern and lower positional choreography in the matched human therapist. We further find that model behaviour varies more with response position than with user state or conversation stage, and that existing quality evaluation does not detect behavioural arrangement once behaviour mix is controlled. These results establish behavioural arrangement as a distinct evaluation dimension and provide a practical way to measure not only what behaviours models produce, but where and in what sequence they produce them.

REPRODUCIBILITY STATEMENT

To make SCRIPT independently auditable and reusable, we release the full analysis code. The estimator is implemented as a standalone module, with its definition, conventions, and default settings documented in §3 and Appendix C. For Testbed 1, two independently written extraction pipelines recover the same span events event-for-event and reproduce the pooled scores to within ± 0.001 (Appendix B.5.2). The pipeline retrieves each corpus from its original source at the revision used in this work, verifies counts and alignment, and reconstructs the annotation layers without redistributing the underlying datasets (Appendix A).

The blind LLM annotation layer is released as behaviour–position events without response text, together with the annotation prompt and regeneration code, so all downstream analyses can be reproduced without access to the annotating model. We release the known-answer and random-label controls, cross-domain anchor configurations, profile-matching procedure, and SCRIPT-Seq implementation (Appendices C.2, A.4, C.4, and C.5). Finally, every numerical result reported in §5 is exported to a single machine-readable file, allowing the manuscript claims to be checked directly against the analysis outputs. These materials allow future researchers to reproduce, audit, and extend SCRIPT to new models, datasets, and behavioural annotation schemes.

AI USE STATEMENT

Generative AI tools were used in two distinct ways in this work. First, as part of the experimental methodology, DeepSeek-V4-Pro was used as a blind annotator to independently re-annotate model and therapist responses under the same 20-code behavioural scheme. The resulting annotations were treated as an experimental annotation layer rather than as ground truth, and all reported analyses were computed from the resulting span annotations using the reproducible pipeline described in the paper.

Second, generative AI tools were used during manuscript preparation to assist with language editing, restructuring sections, improving clarity and readability, and providing feedback on the presentation of the methodology and results. AI-assisted suggestions were reviewed, revised, and verified by the authors against the underlying analyses, tables, figures, and source literature. Generative AI was not used to determine the reported numerical results or statistical conclusions. The authors take full responsibility for the final content, claims, analyses, and presentation of this work.

REFERENCES

Andrew Abbott. Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21: 93–113, 1995. doi: 10.1146/annurev.so.21.080195.000521.

Abeer Badawi, Moyosoreoluwa Olatosi, Negin Baghbanzadeh, Laleh Seyyed-Kalantari, Frank Rudzicz, R. Shayna Rosenbaum, Sara Pishdadian, and Elham Dolatabadi. Towards understanding and measuring COGNITIVE ATROPHY in LLM behaviour. *arXiv preprint arXiv:2606.18129*, 2026. doi: 10.48550/arXiv.2606.18129. URL https://arxiv.org/abs/2606.18129. Accepted at NeurIPS 2026.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025. doi: 10.48550/arXiv.2507.21509.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. A computational framework for behavioral assessment of LLM therapists. *arXiv preprint arXiv:2401.00820*, 2024. doi: 10.48550/arXiv.2401.00820.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5636–5646, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1565.

Zhiyuan Fu, Junfan Chen, Lan Zhang, Ting Yang, Jun Niu, Hongyu Sun, Ruidong Li, Peng Liu, Jice Wang, Fannv He, and Yuqing Zhang. FDLLM: A dedicated detector for black-box LLMs fingerprinting. In *Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 1374–1379, 2025. doi: 10.1109/Trustcom66490.2025.00159.

Dennis A. Gioia and Peter P. Poole. Scripts in organizational behavior. *Academy of Management Review*, 9(3):449–459, 1984. doi: 10.5465/amr.1984.4279675.

Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, New York, 3 edition, 2005.

Emma S. Gueorguieva, Hongli Zhan, Jina Suh, Javier Hernandez, Tatiana Lau, Junyi Jessy Li, and Desmond C. Ong. AI generates well-liked but templatic empathic responses. *arXiv preprint arXiv:2604.08479*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. Large language models as span annotators. *arXiv preprint arXiv:2504.08697*, 2025. doi: 10.48550/arXiv.2504.08697.

Junsol Kim, Shiyang Lai, Nino Scherrer, Blaise Agüera y Arcas, and James Evans. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*, 2026.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=60a1SAtH4e.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585.

Joe O'Connor and Jacob Andreas. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 851–864, 2021. doi: 10.18653/v1/2021.acl-long.70.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272.

Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro, and Rasmus S. Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, 98(3):1064–1072, 2007. doi: 10.1152/jn.00559.2007.

Zehua Pei, Hui-Ling Zhen, Ying Zhang, Zhiyuan Yang, Xing Li, Xianzhi Yu, Mingxuan Yuan, and Bei Yu. Behavioral fingerprinting of large language models. *arXiv preprint arXiv:2509.04504*, 2025. doi: 10.48550/arXiv.2509.04504.

Jenna Russell, Rishanth Rajendhran, Chau Minh Pham, Mohit Iyyer, and John Wieting. Storyscope: Investigating idiosyncrasies in ai fiction. *arXiv preprint arXiv:2604.03136*, 2026. doi: 10.48550/arXiv.2604.03136.

Sraavya Sambara, Yuan Pu, Ayman Ali, Vishala Mishra, Lionel Wong, and Monica Agrawal. MedRedFlag: Investigating how LLMs redirect misconceptions in real-world health communication. *arXiv preprint arXiv:2601.09853*, 2026. doi: 10.48550/arXiv.2601.09853.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6416–6431, 2024. doi: 10.18653/v1/2024.emnlp-main.368. URL https://aclanthology.org/2024.emnlp-main.368/.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276, 2020. doi: 10.18653/v1/2020.emnlp-main.425. URL https://aclanthology.org/2020.emnlp-main.425/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Jack Sidnell. Conversation analysis. In Mark Aronoff (ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2016. doi: 10.1093/acrefore/9780199384655.013.40.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26 (3):339–373, 2000. doi: 10.1162/089120100561737.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 57854–57885. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25z.html.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=yZ7sn9pyqb.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6177–6181, 2022. doi: 10.1109/ICASSP43922.2022.9746035.

Hongli Zhan, Emma S. Gueorguieva, Javier Hernandez, Jina Suh, Desmond C. Ong, and Junyi Jessy Li. Discourse diversity in multi-turn empathic dialogue. *arXiv preprint arXiv:2604.11742*, 2026.

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. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

	Attribute	What the clinician scored	Scale	Span code(s)
R1	Sensitivity (S)	Content indicating risk of harm to self or others.	0/1	SEN
R2	Assumption of user response accuracy (AUR)	Treats the user's self-report as accurate without qualification.	0–2	AUR
R3	Tentativeness / directness (TD)	Hedged and exploratory versus prescriptive and assertive.	0–2	TEN (tentative) DIR (directive)
R4	Fix-It (FIX)	Tries to solve the user's problem rather than explore it.	0–2†	FIX
R5	Recommendation type (RT)	None, skill-based, or action-oriented recommendation.	0–2	RECT
R6	Topic shift (TN)	Stays on the user's topic versus introduces new content.	0–2	TSH
R7	Open vs. closed question (QOC)	No question, a closed question, or an open question.	0–2	QOP (open) QCL (closed)
R8	Language matching (LM)	Reuses the user's own words and phrasing.	0–2	LMT
R9	Minimal encouragers (ME)	Small signals of understanding such as "mhm" or "aha".	0–2	MEN
R10	Empathy and matching (EMP)	Empathic content is accurate to what the user said.	0–2	accurate: VAC, NAC, ASAC, SAC inaccurate: VIN, NIN, ASIN, SIN

Table A.1: Response attributes and span-highlight codes of Testbed 1. Each attribute has a clinician score (the scale column) and one or more highlight codes marking the text that justified the score. †FIX is the one 0–2 scale not ordered by amount: its levels are no solution, inappropriate solution, appropriate solution, which is why FIX is read as a binary. The eight empathy codes combine a type (V validation, N normalizing, AS autonomy support, S support) with an accuracy judgment (AC accurate, IN inaccurate); VAC, NAC, ASAC, and SAC are treated as accurate empathy-related highlights; VIN, NIN, ASIN, and SIN are treated as inaccurate empathy-related highlights.

A TESTBED 1: CODEBOOK, CORPORA, ANNOTATION LAYERS, AND ANCHORS

This appendix collects the materials behind §4: Testbed 1's coding manual, the composition of its four corpora, the two annotation layers we add to it, and the anchor corpora of the ruler.

A.1 ATTRIBUTE AND SPAN-CODE DEFINITIONS

Table A.1 summarizes the response attributes and the corresponding highlight codes used in the analysis. We use the shorthand R1–R10 only to organize the analyses; the statistical tests are run on the original coded variables and the span analyses are run on the highlight codes. Some attributes are represented by one span code, while others are represented by a pair or family of span codes. For example, TD is interpreted through both tentative spans (TEN) and directive spans (DIR), QOC is interpreted through open-question spans (QOP) and closed-question spans (QCL), and EMP is interpreted through accurate and inaccurate empathy subtypes.

Five binary flags provide response-level indicators rather than graded scores: a globally directive tone, projection of the user's underlying experience, new content introduced by the model, potentially harmful validation or normalization, and global incoherence. These flags are reported separately from the R1–R10 scores because they capture global events rather than graded response attributes.

Dataset	Turn type	Unique items	Responses (×5 models)
CareBench	multi-turn	360	1,800
CounselChat	single-turn	50	250
HOPE	multi-turn	359	1,795
PAIR	single-turn	50	250
Total	—	819	4,095

Table A.2: Per-corpus composition of Testbed 1 (the Cognitive Atrophy Benchmark).

A.2 Corpus composition and data preparation

The data were rebuilt via a standard reshape pipeline following a correction to a row-shift error in the PAIR annotation file (item 50 recovered); the correction moved pooled figures only slightly, since PAIR contributes roughly 6% of the pooled corpus, and reversed no qualitative finding. One HOPE conversation has nine annotated turns rather than ten, so HOPE contributes 359 items and the pooled corpus 819 items (4,095 responses); all counts in the paper use these figures.

A.3 The three annotation layers of Testbed 1

§3.4 names the three layers; this section gives what is needed to reproduce the two that are ours.

Surface layer (three labels). Every response is normalised (NFKC, straight quotes, collapsed whitespace, lower case), split into sentences at `.`, `!`, `?`, and each sentence receives at most one label by a fixed order of tests: a sentence containing `?` is a *question*; otherwise the first matching pattern family decides, *empathy* (*it sounds like*, *I hear*, *I'm sorry*, *understandable*, *your feelings are*, *makes sense*, *that must be*, *you're not alone*, *not your fault*, *takes courage*, *I can imagine*, *it's natural/normal/okay*, …), then *advice* (*you should/could/need to*, *try to*, *consider*, *focus on*, *make sure*, *it might be helpful*, *here are*, *reach out*, *seek*, *talk to*, *write down*, *give yourself*, …), then a hedge family (*might*, *perhaps*, *could*, *it seems*, *I wonder*) counted with the question family; sentences matching nothing produce no event. Responses shorter than 20 characters are dropped. The event is (label, sentence start offset / response length). The patterns are deliberately simple and inspectable, and identical for all six speakers, so the layer has a bias but not a speaker-dependent one. It yields 1,258 therapist events.

LLM layer (twenty codes). Each of the 819 prompts and its six responses (five models and the therapist) is sent, one response per call and in shuffled order, to DeepSeek-V4-Pro at temperature 0. The system prompt presents the annotator as a clinician-trained coder, lists the twenty codes with one-line definitions taken from the coding manual as summarised in Table A.1, and fixes the rules: annotate the response only, never the user message; a span is a contiguous verbatim quote of at most about 40 words; mark every instance in order of appearance; the same sentence may carry more than one code; copy quotes character for character; return JSON of `(code, quote)` pairs and nothing else. The user turn supplies the user message as context and the one response to annotate; no speaker or model identity appears anywhere. Returned quotes are normalised as above and located by exact substring search in the normalised response; a quote that is not found is dropped, and duplicate (code, start) pairs within a response are collapsed. This locates 96.7% of the 74,526 returned spans verbatim. The event is (code, start offset / response length), exactly as for the clinician layer. The full layer costs a few dollars of API time. Table A.3 gives the span-by-span agreement of this layer with the clinicians: it agrees on observable moves and almost never on clinical judgments, which is why the level and the choreography ordering transfer between annotators (§5.3) while the fine ordering of Script does not.

A.4 The anchor corpora in detail

§4.2 lists the anchor corpora; this section gives their sizes, schemes, and provenance. Each is scored with the identical estimator, ten position bins, and 200 within-response shuffles. *AnnoMI* (Wu et al., 2022): 133 transcribed motivational-interviewing sessions with real human counsellors, coded under the corpus's own four-label scheme (4,882 events), with positions taken over the whole session;

Table A.3: **Span-by-span agreement of the blind LLM layer with the clinicians.** For every clinician highlight on a model response (39,696 spans on 4,090 responses), the share of its characters that the LLM layer also highlighted under any code is 90.1%, and under the same code 23.6%. Per code: the share of clinician spans of that code at least half covered by an LLM span of the same code. The LLM agrees on observable moves and almost never on the clinical judgments (the four inaccurate-empathy codes, fix-it, topic shift).

Code	spans	matched	Code	spans	matched
QOP	1,083	73%	LMT	1,474	14%
NAC	385	65%	AUR	2,019	12%
RECT	2,881	57%	MEN	167	12%
QCL	1,456	57%	FIX	4,543	5%
VAC	891	55%	TSH	4,776	4%
DIR	8,719	50%	INC	124	2%
SAC	534	39%	NIN	758	1%
ASAC	227	26%	VIN	3,922	1%
SEN	655	22%	SIN	1,782	1%
TEN	2,849	14%	ASIN	451	0%

this is the independent human anchor. *WMT24 machine-translation error spans* and *data-to-text error spans*, both from the span-annotation release of Kasner et al. (2025): human-marked error spans in the outputs of eight MT systems and of four data-to-text generators. The position of a translation error should carry no information about its type, so the MT corpus is an expected null; error spans in structured data-to-text output may show partial structure. *Propaganda-technique spans*, from the same release: the test split of the Propaganda Techniques Corpus (Da San Martino et al., 2019), 101 human-written news articles in which trained experts marked 1,439 spans under an 18-technique scheme (loaded language, name-calling, appeal to fear, repetition, flag-waving, ...). This is the anchor closest in kind to Testbed 1's clinician layer — many nuanced codes, expert annotators, subjective definitions, long texts — but applied to human authors, so it tests whether a fine-grained expert scheme by itself produces a high reading. The release also carries ten LLM-annotator layers over the identical articles, which we use as an annotator-transfer check. *RAGTruth* (Niu et al., 2024): hallucination spans of four types, with character offsets, over the outputs of six generators. *Empathy-tactic annotations* released with the two studies of the second testbed: sentence-level human coding of 303 responses to Reddit support posts under the same ten-tactic scheme (GPT-4-turbo, Llama-3-70B, and human writers with a psychology background) from Study 1 of Gueorguieva et al. (2026), and 322 real human–LLM support conversations (GPT-3.5, GPT-4, Llama-2-70B) tagged with the released taggers by Zhan et al. (2026) (Appendix B.3.2). These carry the human writers and supporters who anchor the ten-tactic scheme, and they let the second testbed's scheme be read under human as well as automatic coding. Two synthetic corpora fix the ends of the scale (§4.2), and a *random-label control* re-deals Testbed 1's labels within each response and runs through the identical pipeline.

B Results in full: tables, figures, and extended analysis

This appendix holds the full tables and analyses behind §5, one subsection per main-text subsection and in the same order; the estimator's own validation is in Appendix C.

The scale returns its known answers before it is applied: a pure slot template reads 0.853 with the signal in C, a pure label chain 0.238 with the signal in M, and random or re-dealt labels read zero, on synthetic data and on real machine-translation error spans alike (Table C.2; Fig. C.2); a second, separately written event extraction lands on the same $39{,}696$ events and reproduces every score to within ± 0.001 (Appendix B.5.2). Two reporting rules follow from the operating range (Appendix C.3): below about a thousand events we quote z rather than the point estimate, and below about six events per response the momentum term is biased downward, so human–model comparisons are sized on C.

Table B.1: **Blind LLM layer: score, matched ceiling and fraction of achievable rigidity for the six speakers of Testbed 1.** 20 codes, 10 bins, 200 shuffles; ceiling from the construction of §3.3 (50 shuffles).

Speaker	SCRIPT	C	M	z	ceiling	fraction	events / response
Qwen	0.085	0.063	0.022	52.4	0.527	16%	17.9
Llama	0.118	0.086	0.032	61.3	0.490	24%	12.1
GPT	0.108	0.083	0.025	73.4	0.553	19%	18.6
Claude	0.114	0.063	0.050	69.0	0.539	21%	16.8
Gemini	0.095	0.063	0.032	56.9	0.525	18%	17.5
Human therapist	0.053	0.045	0.008	12.8	0.242	22%	5.2

B.1 ONE BLUEPRINT ACROSS MODELS; HUMANS READ LOWER (§5.1)

Five frontier models read SCRIPT $= 0.08$–0.12 ($z = 22$–44) with choreography terms within 20% of one another; every same-scheme human comparison reads lower.

The level and the positional choreography are shared. On the clinician span layer (39,696 events) all five models read far above chance and close to one another (Table 2). About a tenth of what a model does next is fixed by where it sits in the response and what it has just done, before the user's message is read; against the most scripted arrangement each corpus's own events permit, the models use 19–25% of the rigidity available to them. The decomposition separates what is shared from what is not. Choreography is nearly identical across the five ($C = 0.052$–0.063); momentum is where they differ ($M = 0.020$ for Llama to 0.064 for Claude). Claude is the only model whose chained moves carry more than half of its score and the only one the paired bootstrap separates from the rest (Holm-adjusted $p = 0.025$ on all four contrasts; no other pair is distinguishable). That gap survives the removal of any single reviewer and a random-effects meta-analysis over reviewer juries (0.042 ± 0.008; Appendix B.5.1), and Claude is the highest-scoring model inside each multi-turn corpus under the clinician annotation (Table B.21).

Human speakers read lower under every shared scheme and annotator. The closest-matched comparison is Testbed 1 itself: the same user turns, answered once by a human therapist and five times by the models. Under the blind twenty-code layer the therapist reads $C = 0.045$ against 0.063–0.086 for the five models. Table B.3 collects every comparison in the paper in which a human speaker and LLM systems were annotated under the same scheme by the same annotator, six settings across both testbeds and two independent corpora (Gueorguieva et al., 2026; Zhan et al., 2026); in every row the human reads below every system, and on Testbed 2 the human gold responses read $C = 0.014$ against 0.050–0.160 for the 22 systems. Because human responses are shorter (4.6 located spans per therapist response in the multi-turn corpora and 5.2 over all four, against 12–19 for the models), the gaps are sized on C. On that scale, real counsellors (AnnoMI) and expert-coded human authors each use 6% of the rigidity their schemes permit, against 19–25% for the five models under the clinician annotation; the next paragraph shows that this contrast does not carry over to the therapist under a shared annotator, which is why the main text does not use it.

The matched comparison under the ceiling. The 6% against 19–25% contrast above compares independent human corpora with the models under the clinician annotation, not the therapist with the models under one annotator. On the blind LLM layer, where all six speakers share one annotator and one scheme, the matched-ceiling fraction does not separate them: the therapist uses 22% of the rigidity the layer permits and the five models 16–24% (Table B.1). The therapist's responses are short (5.2 events per response) and their own ceiling is correspondingly low (0.24 against 0.49–0.55), which makes the fraction hard to read for a speaker of this density; the main text therefore sizes the human gap on the paired C difference and on density-matched comparisons rather than on the ceiling fraction. Within Testbed 1, the therapist reads below every model in every response-length band with enough responses to score, the gap narrowing in the 5–7-event band (Table B.2).

Sequential versus positional structure. Where human structure can be measured, it differs from the models' in kind as well as in size. In both human corpora long enough for the momentum term

Table B.2: **Choreography C within response-length bands, blind LLM layer.** Each cell is C in excess of its null on the responses of that speaker with the stated number of annotated events (n = responses; cells with fewer than 30 responses are not scored). The therapist reads below every model in every band; the gap is narrowest at 5–7 events, where Claude's 0.083 stands against the therapist's 0.079.

Events per response	Qwen	Llama	GPT	Claude	Gemini	Human therapist
3–4	—	0.119 (n=42)	—	—	—	0.058 (n=228)
5–7	—	0.103 (n=160)	—	0.083 (n=55)	0.105 (n=41)	0.079 (n=271)
8–10	0.093 (n=57)	0.095 (n=165)	0.084 (n=32)	0.082 (n=66)	0.077 (n=69)	0.037 (n=68)
≥ 11	0.062 (n=737)	0.082 (n=448)	0.083 (n=771)	0.063 (n=684)	0.061 (n=704)	0.025 (n=56)

Table B.3: **The human reference.** Every comparison in which a human speaker and LLM systems were annotated under the same scheme by the same annotator; values are comparable within a row only. Where several labels share a sentence, M cannot be estimated and SCRIPT is omitted; on the two surface-layer rows every speaker has fewer than six events per response, so M is at or near zero for humans and models alike (Testbed 1: -0.001 to 0.015 for the models, 0.008 for the therapist), SCRIPT is shown as a direction, and C is the comparable term. The last two human corpora have no LLM counterpart and are compared through the ceiling fraction; the five Testbed 1 models use 19–25% of their ceiling. z in parentheses. Full tables in Appendices A.3, B.3.2, B.3.1 and B.5.1.

	Human		LLM systems		
Setting	C	SCRIPT (z)	C	SCRIPT	ceiling
Testbed 1: therapist vs. five models, same prompts					
surface layer (3 labels)	0.071	0.079 (7.8)	0.084–0.180	0.083–0.179	
LLM layer (20 codes)	0.045	0.053 (12.8)	0.063–0.086	0.085–0.118	
Testbed 2: human gold responses vs. 22 systems, same histories					
tagger layer (10 tactics)	0.014	—	0.050–0.160	—	
surface layer (4 labels)	0.005	0.002 (0.1)	0.036–0.202	0.044–0.281	
Independent data, their annotation (10 tactics): humans vs. 2–3 LLMs					
Reddit responses, human coder	0.051	—	0.101–0.126	—	
support conversations, taggers	0.009	—	0.050–0.066	—	
Human corpora with no LLM counterpart					
AnnoMI counsellors (4 MI labels)	0.011	0.047 (20.9)	—	—	6%
news authors, propaganda (18 codes)	0.008	0.022 (3.4)	—	—	6%

to be estimated, momentum exceeds choreography: AnnoMI counsellors read $M = 0.036$ against $C = 0.011$, and human news authors read $M = 0.014$ against $C = 0.008$. The structure these human speakers have is sequential: one move tends to follow another, but their behaviours do not occupy fixed positions in the response. The models are the reverse: their behaviours do occupy fixed positions, and chain less. Under the clinician layer choreography exceeds momentum for four of the five (Table 2; Claude is the exception), and under the blind LLM layer it does for all five ($C = 0.063$–0.086 against $M = 0.022$–0.050). The two human rows inside Testbed 1 are too short for M, so this contrast is read from the independent human corpora. The therapist's responses are too short for M, so the main text states this contrast as a hedge rather than a finding.

B.2 DOES THE BLUEPRINT ANSWER THE PERSON? (§5.2)

Distress moves the therapist most and the models least; explicit cues move the models; the therapist listens first and advises later; and the models' blueprint is the same at every turn and does not carry across turns. These comparisons use the blind LLM layer, which codes the therapist and the five models under the same twenty codes on the two multi-turn corpora. Each user turn carries the clinicians' codes for the person's state: explicit safety-relevant content (0/1), affective intensity (0–2), typicality of the presentation (0–2), implicit distress (0–2), and an explicit request for information or a solution (0/1). Each responsiveness shift is a Jensen–Shannon distance between behaviour mixes in excess of a null that reassigns whole responses, so small strata cannot look responsive by chance.

The seat explains the behaviour; the person barely does. Eq. 3 asks each source what share of the label entropy $H(L)$ it explains above its own null. For the models the seat explains 6.0–8.5% and the coded state of the person 0.5–1.1%, a ratio between six and sixteen; for the therapist the seat explains 5.0% and the person 1.3%, a ratio of four (Table B.5). The ratio compresses at the bottom (Gemini 6, Claude 7); the turn term separates the speakers more cleanly and asks the question directly, whether the point in the conversation matters to the speaker: it explains 1.34% of the therapist's behaviour and 0.11–0.50% of any model's, a factor of three to twelve. The mix comparisons are per event and unaffected by response length, so these shares, unlike the within-response score, are magnitudes. The gap survives normalising each share by the entropy of its own conditioning variable, so a coarse user coding is not the explanation (Appendix C.3, Table C.7).

The therapist's mix moves across the conversation; the models' moves less. Between turns 1–3 and 8–10 the therapist's mix shifts by 0.19 above the null ($z = 9.6$): questions fall from 35% to 23% of moves and advice rises from 8% to 28%. The therapist listens first and advises later. The models advise from the first turn, at 27–52% of their moves, and four of the five drift the same way as the therapist by less (0.02–0.11 over the same span, $z = 1.3$–7.0). By the tenth turn the therapist has arrived at the share of advice from which Claude, the least advice-heavy model, began. The within-response seating does not move at all: C recomputed inside each turn bucket stays at 0.05–0.10 for every model under both annotation layers.

What is felt does not move the models; what is stated does. When the user's message is affectively intense the therapist's advice share halves (29% to 13%) and questions rise (24% to 33%); no model cuts advice by more than five points and Llama and Claude raise it. Under the twenty codes, affective intensity is the therapist's largest shift (0.17, $z = 8.9$) and the models' smallest relative to it (22–36% of the therapist's shift); under the five groups two models do not move at all on it ($z < 1$). Cues the person states move the models much more: the explicit safety cue by 65–88% of the therapist's shift and an atypical presentation by 56–83% (Table B.6). The models track what the person states, not what the person feels.

The blueprint restarts every turn. In the two multi-turn corpora each speaker carried its own ten-turn conversation: the user messages are the same for the therapist and the five models, and each speaker's prompt at turn t contained its own responses to turns $1, \ldots, t-1$. Boundary momentum $M^{\rightarrow}$ (§3.4) therefore asks a fair question of the models: does the opening move of a response depend on how the same speaker's previous response closed, when that response was in the context? It does not. $M^{\rightarrow}$ is null for every model ($|z| \leq 1.8$ on 646 pooled boundaries; Table B.9, Appendix B.2.2), and the null is powered: a boundary effect the size of the within-response momentum would have appeared at $z \approx 4$–5. Uncalibrated cross-turn transition MI on the same boundaries reads 0.05–0.13 and looks like conversational structure, but the within-conversation shuffle null reproduces it almost exactly, so the apparent cross-turn flow is composition and sparsity, not memory; this is the move-level version of the breakdown the dialogue literature has measured in task terms (Li et al., 2024; Laban et al., 2026). Conversation choreography $C^{\rightarrow}$, the turn-scheduled part of the mix, is significant for every model but about twenty times weaker than its response-scale counterpart (Claude highest at $C^{\rightarrow} = 0.006$, $z = 14.8$; Llama lowest at 0.001, $z = 2.4$); it detects only the drift described above. Within the response, moves sit in fixed seats and sometimes chain; across the turn boundary, the chaining vanishes.

Density-matched shares and the between-response normalisation. The therapist writes 4.6 annotated events per response against 12–18 for the models. Table B.4 recomputes the turn and person shares after subsampling every model response to the therapist's per-response event-count distribution (10 draws), and adds the share of each speaker's between-response variation, $I(L; \text{response})/H(L)$ in excess of a null that permutes labels across the speaker's responses, that the turn and the person account for. Density matching raises the models' turn share to 0.23–0.77% and their person share to 0.54–0.76%, against 1.34% and 1.28% for the therapist, so the turn gap holds at a factor of two to six and the person gap at 1.7–2.4. The between-response normalisation does not separate the speakers cleanly: the person accounts for 17% of the therapist's between-response variation and for 8–19% of a density-matched model's, and the turn for 18% against 6–14%. It is reported for completeness and not used in the main text.

Table B.4: **Turn and person shares at matched response density, and as shares of between-response variation.** Blind LLM layer, multi-turn corpora. *Full:* all events. *Matched:* every model response subsampled to the therapist's per-response event-count distribution, mean of 10 draws. *Between:* $I(L;\text{response})/H(L)$ in excess of a null that permutes labels across the speaker's responses. Shares in percent of $H(L)$; the last two columns divide the turn and person shares by the between-response share.

Speaker	Sample	events / resp.	between	turn Π_τ	person Π_U	turn / between	person / between
Qwen	full	18.0	9.9	0.40	0.63	0.04	0.06
	matched	4.5	5.5	0.77	0.76	0.14	0.14
Llama	full	11.6	9.3	0.11	0.52	0.01	0.06
	matched	4.4	3.9	0.23	0.72	0.06	0.19
GPT	full	18.3	6.7	0.33	0.83	0.05	0.12
	matched	4.6	3.8	0.45	0.54	0.12	0.14
Claude	full	16.9	10.8	0.50	0.84	0.05	0.08
	matched	4.5	4.7	0.49	0.67	0.10	0.14
Gemini	full	17.4	12.8	0.32	1.09	0.03	0.09
	matched	4.6	7.4	0.57	0.57	0.08	0.08
Human therapist	full	4.6	7.4	1.34	1.28	0.18	0.17

Table B.5: **Does the behaviour answer the person, the conversation, or the seat?** Six speakers on the blind LLM annotation layer, multi-turn corpora. *Budget:* share of the label entropy $H(L)$ explained above its own null by the seat in the response (the C term), by the coded state of the person (joint of five clinician-coded user attributes), and by the turn of the conversation; ratio = seat / person. *Advice share:* advice as a percentage of the speaker's moves in turns 1–3 and 8–10. *Shift:* Jensen–Shannon distance between behaviour mixes (five groups) in excess of a null that re-deals whole responses, between turns 1–3 and 8–10 (drift) and between the low and high stratum of a user cue; z in parentheses. *Across turns:* the two SCRIPT terms at conversation scale, pooled over 646 turn boundaries and in excess of within-conversation nulls: conversation choreography $C^{\rightarrow}$ and boundary momentum $M^{\rightarrow}$ (§3.4; Appendix B.2.2). The therapist has 4.6 events per response, the models 12–18; the mix and budget comparisons are per event and unaffected by this. Full tables in Appendix B.2.1.

	Budget: share of $H(L)$ explained (%)				Advice share (%)		Shift of the mix, excess JS (z)			Across turns (z)	
Speaker	seat	person	turn	seat / person	turns 1–3	turns 8–10	turns 1–3 → 8–10	affective intensity	safety cue	$C^{\rightarrow}$	$M^{\rightarrow}$
Qwen	6.0	0.63	0.40	10	40.5	50.8	0.086 (6.3)	0.047 (3.4)	0.045 (2.7)	0.003 (6.2)	0.002 (0.2)
Llama	8.5	0.52	0.11	16	36.9	35.5	0.017 (1.2)	0.007 (0.5)	0.052 (2.9)	0.001 (2.4)	0.000 (0.0)
GPT	8.1	0.83	0.33	10	52.2	58.5	0.061 (5.7)	0.031 (2.5)	0.051 (3.8)	0.004 (7.0)	0.002 (0.2)
Claude	6.0	0.84	0.50	7	27.1	30.9	0.105 (7.0)	0.008 (0.5)	0.087 (4.8)	0.006 (14.8)	0.016 (1.8)
Gemini	6.1	1.09	0.32	6	40.0	49.0	0.085 (6.2)	0.052 (3.7)	0.060 (3.4)	0.002 (5.2)	−0.003 (−0.3)
Therapist	5.0	1.28	1.34	4	7.6	28.1	0.188 (9.6)	0.122 (5.6)	0.143 (6.4)	—	—

Table B.6: **Does the behaviour mix move with the person?** For each speaker and each clinician-coded attribute of the user's message, the Jensen–Shannon distance between the behaviour mix of responses in the two strata, in excess of a null that re-deals whole responses to strata (z in parentheses; $S = 300$). Top block: 5 behaviour groups; bottom block: 20 codes. Multi-turn corpora, LLM annotation layer. The therapist moves most on every cue but the explicit request; the models come closest to the therapist on the cues the person states (safety cue, atypical presentation) and furthest on affective intensity.

Speaker	safety cue (0→1)	affective intensity (0→2)	atypical presentation (0→2)	implicit distress (1→2)	explicit request (0→1)
5 behaviour groups					
Qwen	0.045 (2.7)	0.047 (3.4)	0.026 (1.9)	0.006 (0.6)	0.049 (2.1)
Llama	0.052 (2.9)	0.007 (0.5)	0.062 (4.1)	0.022 (2.1)	0.035 (1.7)
GPT	0.051 (3.8)	0.031 (2.5)	0.035 (3.0)	0.011 (1.2)	0.045 (2.9)
Claude	0.087 (4.8)	0.008 (0.5)	0.055 (3.5)	0.012 (1.1)	0.033 (1.4)
Gemini	0.060 (3.4)	0.052 (3.7)	0.063 (3.8)	0.025 (2.1)	0.026 (1.2)
Therapist	0.143 (6.4)	0.122 (5.6)	0.131 (6.7)	0.034 (2.2)	0.036 (1.1)
20 codes					
Qwen	0.120 (7.4)	0.061 (4.9)	0.082 (6.1)	0.030 (3.1)	0.032 (1.5)
Llama	0.105 (6.7)	0.041 (2.8)	0.079 (5.2)	0.044 (4.3)	0.016 (0.8)
GPT	0.127 (9.7)	0.037 (3.2)	0.072 (6.1)	0.025 (2.9)	0.026 (1.6)
Claude	0.113 (4.9)	0.046 (3.3)	0.067 (3.6)	0.022 (2.0)	0.017 (0.6)
Gemini	0.143 (8.2)	0.057 (4.2)	0.100 (6.6)	0.028 (2.5)	0.017 (0.8)
Therapist	0.162 (6.9)	0.171 (8.9)	0.120 (6.3)	0.058 (4.2)	0.046 (1.4)

B.2.1 FULL TABLES

All quantities in §5.2 are computed on the blind LLM annotation layer of §4.1 (blind DeepSeek re-annotation of every response, therapist included, under the twenty codes), restricted to the two multi-turn corpora (CareBench and HOPE; 36 conversations each, ten turns, 719 model responses per model and 690 therapist responses with at least one located span). The user state comes from Testbed 1's clinician coding of each user turn: explicit safety-relevant content (0/1), affective intensity (0–2), typicality of the presentation (0–2), implicit distress (0–2), and an explicit request for information or a solution (0/1).

Shift of the mix. For a speaker and a stratification of responses into two states a and b, the shift is the Jensen–Shannon distance (base 2, square root) between the group (or code) distributions of the speaker's events in state a and in state b. The null re-deals whole responses to the two states at random, preserving the number of responses in each and every response's own events, $S = 300$ times; we report the observed distance minus the null mean, and z. Because the null carries the same event counts as the observation, small strata and short responses inflate the observed and the null distance alike, and the excess is comparable across speakers with very different response lengths. *Drift* is the same statistic between turns 1–3 and turns 8–10. *Budget.* The seat share is the choreography term C of SCRIPT (within-response shuffle null). The person share is $I(L;U)/H(L)$ with U the joint of the five coded user attributes, and the turn share $I(L;T)/H(L)$ with T the turn bucket, each minus the mean of a null that re-deals whole responses across states ($S = 300$). *Level by turn.* SCRIPT recomputed on the events of each turn bucket alone, both annotation layers. *Seat strength.* For each code ℓ on the clinician layer of all four corpora, $\mathrm{KL}(P(X \mid \ell) \,\|\, P(X))$ over the ten position bins, minus the mean of 100 within-response shuffles.

Two observations are worth recording beyond the main text. First, under the twenty codes the therapist's shift with affective intensity (0.17, $z = 8.9$) is the largest single shift in the data, and it is driven by what the therapist stops doing: direct advice falls from 17% of moves at low intensity to 9% at high and fix-it moves from 5.6% to 1.3%, while accurate validation doubles (6% to 13%) and open questions rise (9% to 12%). Second, the seat-strength table shows that the tightest seats in the models' blueprint are the opening validation, the closing questions and the recommendation slot, and that accurate and inaccurate empathy are seated alike; the blueprint places the empathy slot before it knows whether the empathy will be accurate.

Table B.7: **Drift across the conversation and the information budget.** Multi-turn corpora, LLM annotation layer, six speakers. *Drift*: Jensen–Shannon distance between the behaviour mix in turns 1–3 and in turns 8–10, in excess of a null that re-deals whole responses to turn buckets (z in parentheses), under the 5-group and the 20-code alphabet. *Budget*: share of the label entropy $H(L)$ explained by the seat in the response (C, within-response shuffle null), by the coded state of the person (joint of the five user attributes; null re-deals responses across states) and by the turn bucket, each in percent with its z; ratio = seat / person. The therapist's responses are short (4.6 events per response); the mix and budget comparisons are per event, and Table B.4 gives the density-matched values.

	Drift, excess JS (z)		Share of $H(L)$ explained, % (z)					
Speaker	5 groups	20 codes	seat	person	turn	seat/person	events	per response
Qwen	0.086 (6.3)	0.071 (6.3)	6.0 (66)	0.63 (3.7)	0.40 (9.5)	10	12,963	18.0
Llama	0.017 (1.2)	0.030 (2.2)	8.5 (75)	0.52 (3.1)	0.11 (2.5)	16	8,340	11.6
GPT	0.061 (5.7)	0.077 (7.6)	8.1 (88)	0.83 (6.3)	0.33 (10.4)	10	13,176	18.3
Claude	0.105 (7.0)	0.102 (10.8)	6.0 (89)	0.84 (3.4)	0.50 (10.4)	7	12,113	16.9
Gemini	0.085 (6.2)	0.080 (6.1)	6.1 (72)	1.09 (5.9)	0.32 (6.4)	6	12,536	17.4
Therapist	0.188 (9.6)	0.201 (12.2)	5.0 (25)	1.28 (4.5)	1.34 (18.8)	4	3,204	4.6

Table B.8: **Which behaviours have the tightest seat.** Per code, the positional information $\mathrm{KL}\big(P(X \mid \ell) \,\|\, P(X)\big)$ over 10 position bins, in excess of the within-response shuffle null, on the clinician layer of all four corpora (five models pooled). The most seated moves are the opening validation (VAC), the closing questions (QOP, QCL), recommendations (RECT) and the acceptance of the user's account (AUR); direct advice (DIR) is the least seated because it is everywhere. Accurate and inaccurate empathy are seated alike (mean excess 0.36 and 0.20 bits, weighted by events), so the seat does not separate good empathy from bad: the opening slot is filled either way.

Code	Group	mean position	seat (excess bits)	z	events
MEN	other	0.04	1.534	26	167
INC	other	0.68	1.073	38	124
QOP	questions	0.72	0.682	103	1,083
VAC	accurate empathy	0.18	0.628	97	891
AUR	other	0.21	0.467	129	2,019
QCL	questions	0.67	0.443	153	1,456
RECT	advice	0.60	0.386	180	2,881
NIN	inaccurate empathy	0.25	0.340	80	758
NAC	accurate empathy	0.24	0.316	18	385
VIN	inaccurate empathy	0.30	0.206	125	3,922
ASAC	accurate empathy	0.58	0.200	13	227
ASIN	inaccurate empathy	0.55	0.165	23	451
LMT	other	0.30	0.165	53	1,474
SIN	inaccurate empathy	0.34	0.148	22	1,782
FIX	advice	0.50	0.147	124	4,543
SAC	accurate empathy	0.41	0.110	19	534
SEN	other	0.58	0.074	6	655
TSH	other	0.47	0.052	57	4,776
TEN	questions	0.48	0.050	42	2,849
DIR	advice	0.47	0.009	18	8,719

Table B.9: **The script does not cross the turn boundary.** The two SCRIPT terms lifted to conversation scale, on the two multi-turn corpora (5 behaviour groups; every value in excess of its within-conversation permutation null; $S = 1000$ shuffles). Left: conversation choreography $C^{\rightarrow}$, per corpus and pooled — weak but significant, and highest for Claude everywhere. Right: boundary momentum $M^{\rightarrow}$, pooled — null for every model. The raw and null columns show why: the uncalibrated cross-turn transition MI is almost exactly reproduced by the shuffle, so the apparent cross-turn flow is composition, not memory.

	$C^{\rightarrow}$ (excess, z)			$M^{\rightarrow}$, pooled (646 boundaries)			
Model	carebench	hope	pooled	raw	null	excess	z
Qwen	0.002 (1.7)	0.005 (5.8)	0.003 (6.2)	0.056	0.054	+0.002	0.2
Llama	0.004 (3.2)	0.000 (0.1)	0.001 (2.4)	0.065	0.065	−0.000	−0.0
GPT	0.006 (4.9)	0.004 (4.0)	0.004 (7.0)	0.048	0.046	+0.002	0.2
Claude	**0.006 (8.5)**	**0.007 (7.6)**	**0.006 (14.8)**	0.125	0.109	+0.016	1.8
Gemini	0.006 (5.8)	0.002 (2.5)	0.002 (5.2)	0.079	0.082	−0.003	−0.3

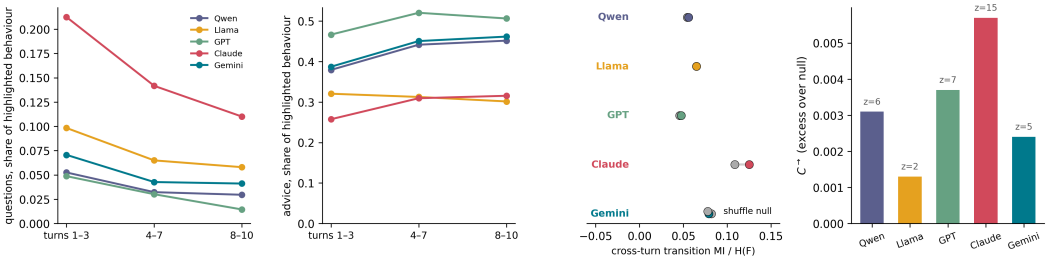


Figure B.1: **The script restarts every turn.** *First two panels:* the conversation-scale schedule that $C^{\rightarrow}$ detects: across turns, every model's questions fall away and its advice rises, the response-scale arc re-run at conversation scale. *Third:* the boundary carries nothing: the raw cross-turn transition MI (colour) is reproduced almost exactly by the within-conversation shuffle null (grey) for every model, so the apparent cross-turn flow is composition, and calibrated boundary momentum $M^{\rightarrow}$ is zero ($|z| \leq 1.8$, powered to detect a response-sized effect at $z \approx 4$–5). *Fourth:* conversation choreography $C^{\rightarrow}$ is real but about twenty times weaker than response-scale C; Claude, the most scripted model within responses, is also the most conversation-scheduled, and Llama the least.

B.2.2 Cross-turn structure: tables and figures

Boundary momentum is null for every model on 646 turn boundaries, with power to detect a response-sized effect; conversation choreography is significant but about twenty times weaker than its response-scale counterpart. Table B.9 and Figure B.1 support the closing paragraph of §5.2; Figure B.2 shows what the within-response shuffle removes.

B.3 Existing evaluation does not see the blueprint (§5.3)

An expert quality judge, a published anti-template intervention, and a hand-written template all pass systems that remain heavily scripted, and the mechanism is measurable: at matched behaviour mix, the judge rates a turn that follows the blueprint no differently from one that does not, to within 0.02 points on a five-point scale (Table B.10). Testbed 2 changes the domain, annotation scheme, annotator, model scale, and number of systems, while holding the prompts and quality judge fixed within the testbed. On the released tagger layer, several tactics can share a sentence, so M is outside its operating range and we read C; on the one-label-per-sentence surface layer, the full score is estimable. The two layers order the 22 systems similarly (Spearman $\rho = 0.71$). Table B.12 gives the 4B systems and the human gold responses; Table B.13 gives all 22.

At matched behaviour mix, conformity to the blueprint does not move the rating. Over $1{,}011$ matched groups and $25{,}919$ pairs of Testbed 2 turns from the same system with the same multiset of labels in a different arrangement, the more scripted turn of each pair is rated $+0.004$ higher

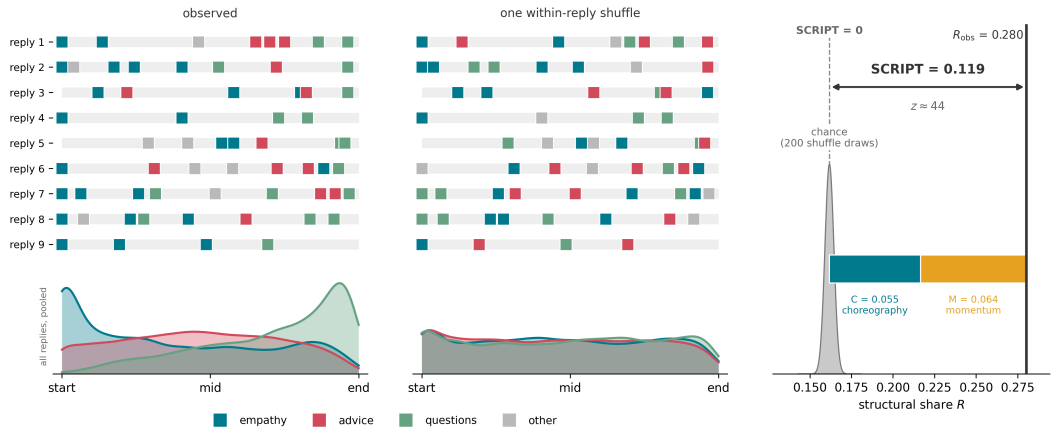


Figure B.2: **What the metric sees.** *Left:* nine real annotated Claude responses to nine different people, each span drawn at its position, and below them the position density of each behaviour pooled over all of Claude's responses: empathy clusters at the opening, advice in the middle, questions at the end. *Middle:* the same responses and the same pooled densities after one draw of the within-response shuffle null of Eq. 2. Span counts, positions, slot structure and behaviour mix are all preserved, so none of them can move the score, yet the three densities collapse onto one shape: position no longer predicts behaviour, which is exactly what the shuffle removes and the metric isolates. *Right:* the score on all of Claude's pooled span events: 200 shuffle draws form the chance spike, and the dashed line through it is the score's meaningful zero. The observed R sits far outside it; the gap is SCRIPT $= 0.119$ ($z \approx 44$), split into choreography ($C = 0.055$, fixed positions) and momentum ($M = 0.064$, chained moves).

on the 1–5 scale, with a cluster-bootstrap 95% interval of $[-0.015, +0.022]$ (Table B.10). Any arrangement effect larger than 0.02 points, a sixteenth of the within-system spread of the rating, is excluded. Conformity is measured per response: Eq. 4 scores a response under its own system's profile built from the system's other responses, and the excess over within-response shuffles says how closely the response follows the blueprint with its label mix held fixed (Appendix C.6). That excess is positive for 93.5% of Testbed 1 responses and 89% of Testbed 2 turns, so the corpus-level score is carried by almost every response, not by a few. Before matching, the judge does appear to reward conformity within 13 of the 22 systems (median $\rho = 0.26$ among the prompting systems; pooled within-system $\rho = 0.15$); after matching the preference is gone. The pairs are not close calls: they differ in conformity by one within-system standard deviation on average, and the result holds among the half of the pairs that differ most and when every turn is scored against the pooled profile of all 22 systems rather than its own. The judge's apparent sensitivity to conformity is a sensitivity to which behaviours a turn contains, not to where they sit; that is how a per-response rating passes a templated system. On Testbed 1 the same design with the clinicians' ten judgments is consistent but less conclusive, because only short responses find a match (693 pairs, 4.3 events per response): no difference is significant, and the intervals exclude effects above 0.1 for six of the ten judgments (Table C.10).

Positive control: the judge does respond to length at matched mix. The matched-mix design would be uninformative if the judge could not respond to anything within a matched group. Within the same groups, the longer of two same-mix turns (characters of the normalised response) is rated $+0.039$ higher, 95% CI $[+0.021, +0.058]$, and $+0.048$ $[+0.025, +0.069]$ among the half of pairs with the larger length gap (Table B.11). The design therefore detects an effect of a few hundredths of a point; the arrangement has none. The surface layer leaves 32–63% of sentences unlabelled ("other"), so part of the arrangement varied within a pair is where those sentences sit; the tactic-multiset version under the released taggers is left for future work.

Across systems, the rating rises with the score. Across the 22 systems the expert-protocol empathy rating correlates positively with surface-layer SCRIPT ($\rho = 0.72$, the same at matched sample size) and with tagger-layer C ($\rho = 0.37$): the systems the judge rates highest are among the most

Table B.10: **At matched behaviour mix, conformity to the blueprint does not move the rating.** Pairs of responses from the same system with the same label multiset and a different arrangement; *diff* is the outcome of the higher-conformity response minus the lower, averaged within matched groups, with a cluster-bootstrap 95% interval. Testbed 2: the judge's empathy rating (1–5; within-system SD 0.36). Testbed 1: the clinicians' ten flags (0–1) and scores (0–2), summarised. *Raw* ρ: the pooled within-system correlation of conformity with the outcome before matching. Full table in Appendix C.6.

	groups	pairs	diff	95% CI	raw ρ
Testbed 2, all 22 systems, own profile	1,011	25,919	$+0.004$	$[-0.015, +0.022]$	$+0.15$
top half of the conformity gap	737	12,960	0.000	$[-0.023, +0.024]$	
pooled profile of all 22 systems	1,011	25,910	0.000	$[-0.019, +0.019]$	$+0.11$
Testbed 1, ten clinician judgments	198	693	$\lvert\text{diff}\rvert \leq 0.06$	bounds ≤ 0.14	$\lvert\rho\rvert \leq 0.14$

Table B.11: **Matched-mix contrasts on Testbed 2, with the length positive control.** Pairs of turns from the same system with the same surface-layer label multiset; *diff* is the judge's rating of the higher-conformity (or longer) turn minus the lower, averaged within matched groups, cluster-bootstrap 95% interval over groups (2,000 resamples); the arrangement contrast in the same groups is in Table B.10.

Contrast	groups	pairs	diff	95% CI	median gap
Response length (characters)	1,013	26,191	$+0.039$	$[+0.021, +0.058]$	49
Response length, top half of the gap	837	13,142	$+0.048$	$[+0.025, +0.069]$	85

choreographed. Response length does not explain it: the rating's correlation with events per turn is $\rho = 0.40$ (not significant), and the correlation with surface-layer SCRIPT is 0.65 after controlling for length (0.60 for surface C, 0.28 for tagger C). What the judge rewards is the mix: its rating rises with the share of question sentences ($\rho = 0.71$) and falls with the share of unlabelled ones (-0.69). The judge rates the real human supporters below every one of the 22 systems (2.90 against 3.22–4.67; Table B.12), so its scale is not anchored to human support. The reading that SCRIPT simply tracks competence does not hold: human speakers read low across the board, trained counsellors and expert authors among them (Table B.3). On Testbed 1, stratifying each model's responses by clinician-rated empathy accuracy or by the harmful-content flag gives small gaps of either sign; the only gaps that resolve at matched stratum size run the other way, with the better-rated stratum slightly more scripted (Gemini on both axes, $z = 2.8$ and 3.9; Qwen on empathy accuracy, $z = 2.3$; Appendix B.5.1). SCRIPT measures structure, not quality, and the two should be reported together.

No intervention both lowers the blueprint and raises the rating. MINT, the anti-template intervention of Zhan et al. (2026), lowers cross-turn stickiness by the 26% they report (0.571 → 0.422 at 4B) and tagger-layer C by the same proportion (0.150 → 0.112, −25%), to the level a tactic-list prompt reaches alone (0.110), twice the verbalized-sampling floor (0.050) and eight times the human level (0.014); at 1.7B neither statistic moves (0.514 → 0.510; 0.160 → 0.157). Under the surface layer the blueprint after MINT is higher than before it (C 0.093 → 0.195; surface SCRIPT 0.112 → 0.263), though its momentum term is read at a higher density (5.6 → 8.1 events per turn) and is not compared. Varying the tactics across turns did not remove the within-response blueprint. The pattern holds across the table (Table B.12): every RL recipe reads above vanilla prompting on the surface layer and within 0.04 of it on the tagger layer, and the recipes rated most empathic (4.6–4.7) are among the most scripted. Prompting is what moves the score, and the one system at the floor, verbalized sampling with the vanilla prompt ($C = 0.050$, SCRIPT $= 0.000$), is also the lowest-rated prompting system (3.51). A stickiness statistic and a quality judge both certified MINT; neither could see what it left in place.

A hand-written template, read against its chance level. Gueorguieva et al. (2026) report that their hand-built five-block template matches 83–90% of LLM responses. Re-dealing each response's tactics at random still matches 36–46%, and the within-response coverage they also report is reproduced by the null almost exactly, so it carries no information about order. What survives the null is the LLM–human gap (46–52 points of excess against 17–28 for human writers), which is the

Table B.12: **Testbed 2 at 4B: eleven interventions on one base model, and the human gold responses.** All systems answer the same 315 Lend an Ear supporter turns and are rated by the same expert-protocol judge (Zhan et al., 2026). *Tagger C:* choreography under their released ten-tactic taggers (several tactics per sentence, so C is the comparable term). *Surface* SCRIPT*:* the full score under the one-label-per-sentence surface layer, z in parentheses. *Stickiness* is their statistic recomputed with their definition; *empathy* their aggregated rating; events per turn on the tagger layer. Prompting lowers the blueprint and RL does not; only verbalized sampling with the vanilla prompt reaches the floor on the tagger layer, and it has the lowest empathy rating of the prompting systems. The 1.7B systems, in Table B.13, give the same picture.

System (4B)	Tagger C	Surface SCRIPT (z)	Stickiness	Empathy	Events/turn
Prompting					
vanilla prompting	0.150	0.112 (25)	0.571	3.75	7.1
tactic-list prompt	0.110	0.127 (22)	0.515	3.88	6.7
tactic-history prompt	0.108	0.109 (18)	0.402	3.78	6.8
verbalized sampling (vanilla prompt)	**0.050**	**0.050 (6)**	0.395	3.51	4.6
verbalized sampling (tactic prompt)	0.065	0.083 (17)	0.537	3.67	7.5
verbalized sampling (history prompt)	0.056	0.044 (9)	0.420	3.61	6.8
Reinforcement learning					
quality-only RL	0.155	0.206 (110)	0.671	4.62	16.3
quality RL + token diversity	0.120	0.239 (90)	0.566	4.58	12.8
MINT (quality + KL novelty)	0.112	0.263 (81)	0.422	4.67	10.7
MINT (quality + entropy)	0.136	0.180 (97)	0.858	3.49	19.3
MINT (quality + KL + entropy)	0.120	0.173 (81)	0.718	4.08	16.7
Human gold responses	0.014	0.002 (0.1)	0.266	2.90	2.4

finding; the raw coverage could not say how much of it was chance. The position half of the script profile, computed on the same human-coded data with no search, is their template written as a table with its chance level known (Appendix B.3.1). Their stickiness statistic decomposes the same way (Table B.14).

B.3.1 THE 22 SYSTEMS, THE HAND-WRITTEN TEMPLATE, AND STICKINESS

Under the within-response null, about half of the coverage of the hand-written template and about half of the LLM stickiness statistic are available to chance; the LLM–human gap survives in both, and the position half of the script profile recovers the template without a search. Table B.13 gives the full 22-system table summarised in Table B.12. All three checks use the public release of Zhan et al. (2026). *Systems.* Each system's 315 supporter turns are split into sentences with the same NLTK tokenizer their tagger used (sentence and tag counts agree for 89–100% of turns; the rest fall back to sentence rank), and one (tactic, position) event is formed per tactic on a sentence. SCRIPT uses 10 bins and 200 shuffles. Because the systems differ in response length (2.4 to 21.6 events per turn) and the score is downward-biased at small samples, we also rescore each system on 20 random subsamples of turns holding ≈1,150 events, the size of the smallest LLM system, and report the subsample-extrapolated estimate of Appendix B.5.1. Stickiness follows their evaluation code exactly: for each of the ten tactics, the probability that it appears in the system's turn t given that it appeared in the gold turn $t-1$ (the systems answer the gold history), averaged over tactics; our values reproduce their Table 2 (vanilla 4B 0.571, MINT Q+KL 4B 0.422, human 0.266). Empathy is the mean of their aggregated rating over 315 turns. *Regex.* Tactic letters follow their Table 3 (X emotional expression, V validation, P paraphrasing, A advice, I information, E empowerment, R reappraisal; other tactics receive letters that break a match), tactics within a sentence are ordered by their column order, consecutive repeats are collapsed, and a response at level k counts as matched if pattern k matches from the start or a shorter pattern matches to the end of the string. The null re-deals each response's letters within the response and re-collapses; 200 draws. *Stickiness null.* Within a speaker, the turns at each turn index are permuted across conversations ($S = 300$), preserving the per-index composition and destroying within-conversation continuity.

The two studies that reported the template by hand released their annotations, so their own statistics can be re-read under the instrument's null. §5.3 did this for the MINT intervention; three checks

Table B.13: **SCRIPT on the 22 systems of Zhan et al. (2026), under two annotation layers.** All systems answer the same 315 supporter turns of the Lend an Ear test set (gold conversation history) and are rated by the same expert-protocol judge. *Tagger layer:* their released taggers under the ten-tactic scheme, one event per tactic marked on a sentence; several tactics share a sentence, so transitions are few, M is outside its operating range (negative for most systems) and C is the comparable column; *matched* n rescores each system on turn subsamples holding ≈1,150 events (mean of 20 draws). *Surface layer:* the three-label sentence annotator of §3.4 with unmatched sentences kept as a fourth label, one event per sentence, so both terms are estimable and SCRIPT is read in full. Stickiness is their statistic recomputed with their definition; empathy is their aggregated rating; events per turn are given for both layers.

		Tagger layer (10 tactics, multi-label)					Surface layer (1 label / sentence)						
System	size	C	M	SCRIPT	z	matched n	C	M	SCRIPT	z	stickiness	empathy	events/turn
vanilla prompting	1.7B	0.160	−0.034	0.125	17	0.089	0.185	0.004	0.188	30	0.514	3.60	6.5 / 5.1
vanilla prompting	4B	0.150	−0.023	0.127	19	0.092	0.093	0.019	0.112	25	0.571	3.75	7.1 / 5.6
tactic-list prompt	1.7B	0.135	−0.049	0.085	10	0.054	0.158	0.013	0.171	25	0.449	3.90	5.6 / 4.4
tactic-list prompt	4B	0.110	−0.002	0.108	14	0.077	0.104	0.023	0.127	22	0.515	3.88	6.7 / 4.9
tactic-history prompt	1.7B	0.115	−0.060	0.054	6	0.031	0.155	0.005	0.160	23	0.403	3.89	5.7 / 4.3
tactic-history prompt	4B	0.108	−0.042	0.065	9	0.030	0.086	0.023	0.109	18	0.402	3.78	6.8 / 4.7
verbalized sampling (vanilla prompt)	1.7B	0.067	−0.052	0.014	1	0.013	0.060	−0.010	0.051	6	0.410	3.22	3.8 / 3.0
verbalized sampling (vanilla prompt)	4B	0.050	−0.050	0.000	0	−0.008	0.036	0.014	0.050	6	0.395	3.51	4.6 / 3.4
verbalized sampling (tactic prompt)	1.7B	0.114	−0.064	0.050	5	0.035	0.152	0.002	0.154	20	0.454	3.71	4.7 / 3.6
verbalized sampling (tactic prompt)	4B	0.065	−0.018	0.048	7	0.017	0.073	0.010	0.083	17	0.537	3.67	7.5 / 5.3
verbalized sampling (history prompt)	1.7B	0.095	−0.060	0.034	3	0.022	0.125	0.008	0.134	16	0.377	3.65	4.7 / 3.3
verbalized sampling (history prompt)	4B	0.056	−0.036	0.019	3	−0.005	0.048	−0.003	0.044	9	0.420	3.61	6.8 / 4.8
quality-only RL (PsychoCounsel)	1.7B	0.152	0.014	0.166	51	0.114	0.132	0.083	0.215	108	0.649	4.42	14.6 / 11.8
quality-only RL (PsychoCounsel)	4B	0.155	0.023	0.178	53	0.118	0.138	0.067	0.206	110	0.671	4.62	16.3 / 12.9
quality RL + token diversity (R1-Zero-Div)	1.7B	0.108	0.013	0.121	35	0.081	0.142	0.052	0.194	84	0.631	4.38	13.2 / 10.2
quality RL + token diversity (R1-Zero-Div)	4B	0.120	−0.004	0.116	31	0.075	0.202	0.037	0.239	90	0.566	4.58	12.8 / 10.1
MINT (quality + KL novelty)	1.7B	0.157	0.043	0.200	59	0.160	0.183	0.098	0.281	104	0.510	4.54	13.3 / 10.7
MINT (quality + KL novelty)	4B	0.112	0.018	0.129	27	0.099	0.195	0.068	0.263	81	0.422	4.67	10.7 / 8.1
MINT (quality + entropy)	1.7B	0.140	0.024	0.164	66	0.087	0.129	0.054	0.183	110	0.858	3.56	21.6 / 18.3
MINT (quality + entropy)	4B	0.136	0.037	0.173	69	0.083	0.133	0.047	0.180	97	0.858	3.49	19.3 / 16.7
MINT (quality + KL + entropy)	1.7B	0.146	0.008	0.155	45	0.092	0.174	0.030	0.204	80	0.642	4.25	15.3 / 11.6
MINT (quality + KL + entropy)	4B	0.120	0.018	0.138	42	0.068	0.117	0.056	0.173	81	0.718	4.08	16.7 / 12.8
Human gold responses	—	0.014	−0.061	−0.047	−2.5	−0.047	0.005	−0.004	0.002	0.1	0.266	2.90	2.4 / 1.9

remain: their template under our null, their stickiness statistic decomposed, and their template recovered from position alone.

Their template under our null. The template of Gueorguieva et al. (2026) is a regular expression over tactic letters that matches 83–90% of LLM responses and far fewer human ones. We ran their patterns on their Study-1 coding and on the same tactic strings with each response's own tactics redealt at random, our within-response null. Our reproduction of their coverage is within a few points of theirs. Under the null, their opening pattern still matches 38% of GPT-4-turbo responses and 46% of Llama-3-70B responses, and their full five-block pattern 36% and 40%; for the human writers, 29% and 18%. Roughly half of the reported LLM coverage, and half of the human coverage, is available to a random arrangement of the same tactics, and the within-response coverage they also report (81–92%) is reproduced by the null almost exactly (83–95%), so it carries no information about order at all. The LLM–human gap survives the null (excess 46–52 points against 17–28), which is the finding; the raw coverage does not say how much of it was chance, which is the reason a template needs a meaningful zero.

Stickiness decomposed. Their second statistic, the probability that a behaviour present in one turn is present in the next, can be given the same treatment: permute, at each turn index, which conversation a speaker's turn belongs to, so that how often each behaviour is used at each point of a conversation is preserved and only within-conversation continuity is destroyed (Table B.14). On their 322 conversations about half of the LLM stickiness is composition ($0.46 \rightarrow$ null 0.25 for GPT-3.5; $0.51 \rightarrow 0.38$ for GPT-4), and what remains is still two to three times the human supporters' excess (0.21 and 0.13 against 0.07). On our multi-turn corpora the same statistic shows how much a null can change a reading: under the clinician layer every model has a large excess (0.25–0.29, $z \approx 28$), under the single-annotator LLM layer almost none (0.01–0.05). One reviewer codes a whole conversation in Testbed 1, so what persists across its turns is the coder's hand. Under the single-annotator layer the models, each continuing its own ten-turn conversation with its previous responses in context, carry almost no behaviour from one turn to the next; the therapist, a real person continuing a real conversation, is the one speaker with continuity (0.08, $z = 8.4$).

Table B.14: **Tactic stickiness decomposed.** Stickiness as defined by Zhan et al. (2026): the probability that a behaviour present in one supporter turn is present in the next, averaged over behaviours, within conversations. *Null*: the same statistic after permuting, at each turn index, which conversation a speaker's turn belongs to ($S = 300$), which preserves how often each behaviour is used at each point of a conversation and destroys only within-conversation continuity. *Excess* is what remains. On their data about half of the LLM stickiness is composition, and the remaining excess is two to three times the human supporters' for GPT-3.5 and GPT-4. On our corpora each model continued its own ten-turn conversation with its previous responses in context, and its excess is nonetheless small; the therapist, a real person continuing a real conversation, shows more. Under the clinician layer the excess is large for every model because one reviewer codes a whole conversation: what persists across turns there is the annotator's hand, which the single-annotator LLM layer removes.

Speaker	stickiness	null	excess	z	turn pairs
Their data: 322 real support conversations and Lend an Ear, their taggers, ten tactics					
GPT-4	0.509	0.380	+0.130	8.4	228
Llama-2-70B	0.548	0.485	+0.063	3.1	103
GPT-3.5-turbo	0.457	0.247	+0.210	32.1	1022
Human supporters	0.266	0.198	+0.068	4.3	265
Our multi-turn corpora, LLM annotation layer, five behaviour groups					
Qwen	0.689	0.673	+0.016	2.1	647
Llama	0.680	0.635	+0.045	5.2	645
GPT	0.674	0.666	+0.008	0.8	647
Claude	0.778	0.747	+0.031	4.0	647
Gemini	0.700	0.674	+0.026	3.0	647
Therapist	0.574	0.492	+0.082	8.4	618
Our multi-turn corpora, clinician layer, five behaviour groups (one reviewer per conversation)					
Qwen	0.749	0.485	+0.264	27.2	646
Llama	0.789	0.521	+0.268	27.9	647
GPT	0.770	0.511	+0.259	29.0	646
Claude	0.793	0.545	+0.248	27.9	646
Gemini	0.767	0.480	+0.287	28.9	647

The template found without looking. Their template was found by inspecting tactic strings with the help of a beam search. The position half of the script profile, computed on their human-coded data with no search, is the same template written as a table: paraphrasing, validation and emotional expression occupy the opening decile of GPT-4-turbo's responses, advice and information the middle and end, questions a narrow band in the middle; the human writers' rows are spread over the response. A profile is a template with its position information kept and its chance level known, and it does not need to be rediscovered for the next taxonomy.

B.3.2 Human writers and supporters under the ten-tactic scheme

On the released annotations of both empathy-tactic studies, under their own human coder and their own automatic taggers, the human writers and supporters read a choreography term two to seven times lower than the LLM systems, and the random-label controls read zero. These are the two independent-data rows of Table B.3.

Gueorguieva et al. (2026) release, with the code of Zhan et al. (2026), three annotation layers under their ten-tactic scheme (validation, paraphrasing, advice, information, emotional expression, empowerment, reappraisal, questioning, assistance, self-disclosure, plus a few rare codes). *Study 1 human coding*: 3,164 sentences from 303 responses to Reddit support posts, coded sentence by sentence by one trained coder, multi-label; the responses were written by GPT-4-turbo (179), Llama-3-70B (50) and human writers with a psychology background (74). *Real conversations*: 322 human–LLM emotional-support conversations from WildChat and SENSE-7 (1,719 supporter turns; GPT-3.5-turbo, three GPT-4 variants, Llama-2-70B), tagged with their released Llama-3.1-8B tactic taggers; ten conversations of unknown source are excluded. *Lend an Ear*: 50 human–human support conversations, 315 human supporter turns, tagged with the same taggers.

Table B.15: **SCRIPT on the empathy-tactic annotations of Gueorguieva et al. (2026) and Zhan et al. (2026).** Same estimator, same 10 bins, same within-response shuffle null as everywhere else in the paper; one event per tactic marked on a sentence, at the sentence's normalised start offset. Control is the score of the same events with labels re-dealt within each response. Because several tactics share a sentence, transitions are few and M is outside its operating range at these densities; the comparable column is C, which reads 0.050–0.126 for the LLM systems and 0.009–0.051 for the human writers and supporters under both the human coding and the automatic taggers.

System	C	M	SCRIPT	z	control	events / responses	events/response
Study 1 of Gueorguieva et al. (2026): sentence-level human coding, Reddit support posts							
GPT-4-turbo	0.101	0.026	0.127	29.8	0.003	3,079 / 179	17.2
Llama-3-70B	0.126	−0.086	0.040	2.8	0.010	541 / 50	10.8
Human writers (psychology background)	0.051	−0.069	−0.018	−1.6	−0.013	952 / 74	12.9
Real support conversations (Zhan et al., 2026): WildChat/SENSE-7 tagged with their released taggers; Lend an Ear human supporters							
GPT-3.5-turbo	0.050	0.025	0.076	24.9	0.003	12,026 / 1,203	10.0
GPT-4 (three variants pooled)	0.052	0.002	0.053	13.4	0.004	5,468 / 293	18.7
Llama-2-70B	0.066	−0.028	0.038	6.2	−0.003	2,542 / 118	21.5
Human supporters (Lend an Ear)	0.009	−0.081	−0.072	−3.8	−0.010	776 / 296	2.6

Table B.16: **Identification baselines and transfer conditions.** Rank-1 accuracy, five models, chance 0.20; 200 probe draws per model and held-out slice.

Probe scored with	k=1	k=5	k=10	k=20
Full profile, held-out corpus	0.38	0.57	0.65	0.74
position table only	0.37	0.55	—	0.73
behaviour mix only	0.35	0.48	—	0.60
full profile, probe re-dealt	0.34	0.50	—	0.62
Full profile, within corpus (odd/even split)	0.34	0.53	0.59	0.65
Full profile, held-out annotator	0.38	0.43	0.45	0.44
Full profile, corpus and annotator held out	0.36	0.43	0.45	0.46
Label-permutation control	0.19	0.18	0.18	0.18

We reduce each layer to events exactly as for Testbed 1: one (tactic, position) event per tactic marked on a sentence, with position the sentence's start offset in the response divided by the response's length (94% of sentences located verbatim; the rest fall back to sentence rank). SCRIPT is then computed with the same 10 bins, the same 200 within-response shuffles, and the same estimator. Table B.15 gives the result. These layers are sentence-level and multi-label: several tactics are marked on one sentence and become co-located events, so under the slot convention a response of ten to twenty events carries only a few transitions and M is outside its operating range (§5); it is negative for four of the seven rows, including both human rows. The comparable quantity is the choreography term C, and it tells one story in both layers. Under their own human coding, GPT-4-turbo and Llama-3-70B read $C = 0.101$ and 0.126 against 0.051 for the human writers; in real conversations under their taggers, GPT-3.5, GPT-4 and Llama-2-70B read 0.050–0.066 against 0.009 for the human supporters. The human direction of the ruler therefore holds under a third and a fourth annotator and a third scheme, and the random-label controls stay at zero. The pooled SCRIPT values are reported for completeness; where M is out of range they should not be compared across rows.

B.4 THE PROFILE NAMES THE MODEL (§5.3)

What the profile reads: baselines. Table B.16 scores the same held-out-corpus probes four ways. The behaviour mix alone (label frequencies, no position or order) reaches 0.60 at $k = 20$ against 0.74 for the full profile; the position table alone reaches 0.73, so the transition table adds about one point; and re-dealing the probe's labels within each response, the paper's own null applied to the probe, brings the full profile down to 0.62. The arrangement therefore carries identity, but the larger share of the above-chance accuracy comes from which behaviours a model uses. Holding out the annotator keeps accuracy at about twice chance whatever k (0.38–0.44), as does holding out corpus and annotator together (0.36–0.46).

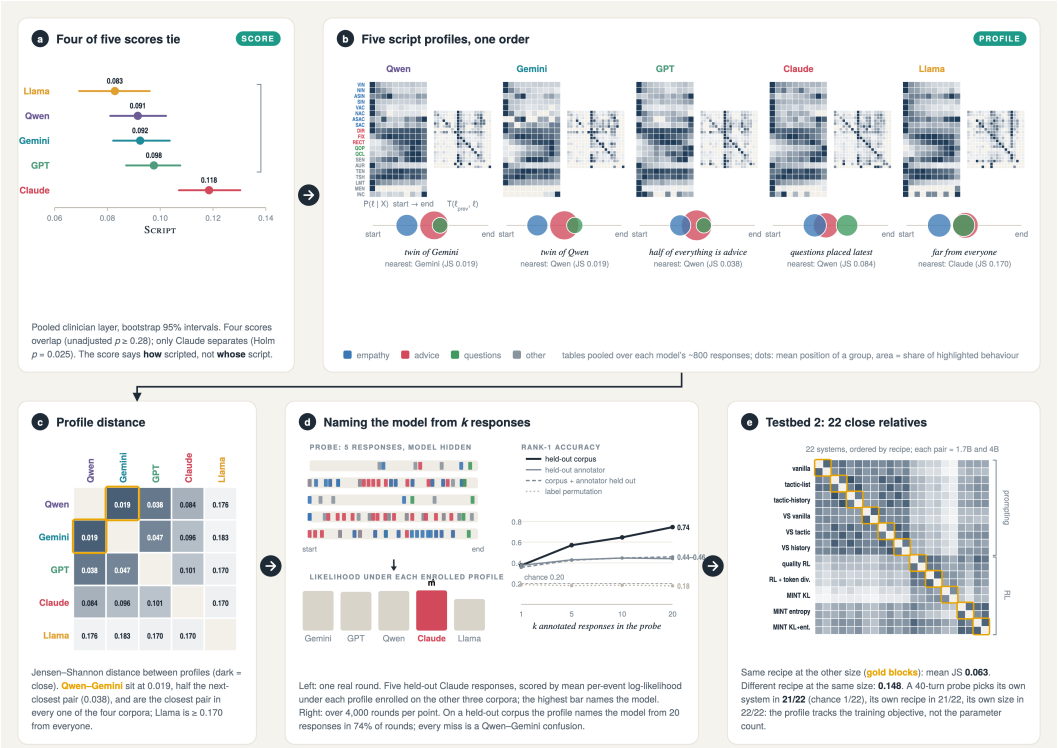


Figure B.3: **Tied scores, five profiles: the script profile names the model.** **(a)** Four of the five SCRIPT scores overlap within their bootstrap intervals; only Claude separates. **(b)** The five script profiles behind those scores, the position table $\hat{P}(\ell \mid X)$ and the transition table $\hat{T}(\ell_{\text{prev}}, \ell)$ of each model, with the mean position of empathy, advice and question spans below: one shared order, with a model-specific accent. **(c)** Jensen–Shannon distance between profiles; Qwen and Gemini are near twins, Llama is far from everyone. **(d)** One real round of the identification game: five held-out responses of one model, model name hidden, scored under each enrolled profile; the highest likelihood names the model. Right: rank-1 accuracy by probe size, on a held-out corpus, a held-out annotator, both, and under a label-permutation control. **(e)** Testbed 2: the profile distance between the 22 systems, ordered by recipe; the two sizes of one recipe (gold blocks) sit closer than any two recipes.

Four of the five models are indistinguishable on the scalar score, but the profile behind the score separates all five: on a held-out corpus it names the generating model from twenty annotated responses three times out of four (chance one in five), and on Testbed 2 a forty-turn probe picks its own system among 22 close relatives in 21 of 22 cases, with no text, no training, and no model access (Fig. B.3).

The profile names an unknown sample. Enrolling each model's profile on three corpora and scoring probes of k annotated responses from the held-out fourth by Eq. 4, rank-1 accuracy is 0.38 at $k = 1$, 0.57 at $k = 5$ and 0.74 at $k = 20$ (chance 0.20; Table B.17). With whole held-out corpora as probes, 17 of 20 (model, corpus) cases are identified, and every miss is a Qwen–Gemini confusion (Fig. B.4 walks through one round). Held out to an unseen annotator the curve is lower (0.38 at $k = 1$ to 0.44 at $k = 20$), and with corpus and annotator both unseen it settles at about twice chance (0.36–0.46); a label permutation run through the identical pipeline sits at chance throughout (0.18–0.19). The script profile therefore transfers across corpora nearly intact and partially across annotators (Appendix B.4.1).

One order, five signatures. What the classifier reads is visible by eye. Computing C and M builds the two count tables of §3.4, and read side by side the five profiles show a shared order with model-specific accents (Fig. B.3; Table B.17). Empathy opens and advice fills the middle in every

Table B.17: **The script profile identifies the system.** Top: Testbed 1. Left, the behaviour geometry read from each model's profile, the mean position and share of the three headline groups ("questions" includes hedging: QOP+QCL+TEN; shares are of all highlighted behaviour). Right, profile distance and identification: JS = Jensen–Shannon distance to the nearest and to the farthest other model; ID@k=5 = rank-1 identification of the model from 5 annotated responses, enrolled on three corpora and tested on the held-out fourth, pooled over the four choices of held-out corpus (200 draws each; chance 0.20). Middle: the identification curve by probe size and transfer condition (rank-1 accuracy; chance 0.20). Bottom: Testbed 2, eleven recipes at two sizes of one base model; JS between profiles and identification of a 40-turn probe against all 22 profiles (Appendix C.3).

	Empathy		Advice		Questions			Profile	
Model	pos.	share	pos.	share	pos.	share	One-line signature	JS min / max	ID@k=5
Qwen	0.29	0.221	0.53	0.429	0.58	0.086	twin of Gemini	0.019 / 0.176	0.42
Gemini	0.27	0.200	0.50	0.443	0.59	0.097	twin of Qwen	0.019 / 0.183	0.29
GPT	0.35	0.200	0.49	0.501	0.56	0.081	half of everything is advice	0.038 / 0.170	0.60
Claude	0.36	0.253	0.46	0.312	0.64	0.212	questions placed latest	0.084 / 0.170	0.66
Llama	0.29	0.263	0.51	0.327	0.50	0.226	far from everyone	0.170 / 0.183	0.86

Identification by probe size k (rank-1 accuracy, five models, chance 0.20)

Transfer condition	k=1	k=5	k=10	k=20
held-out corpus, same annotator	0.38	0.57	0.65	0.74
held-out annotator, same corpora	0.38	0.43	0.45	0.44
corpus and annotator both held out	0.36	0.43	0.45	0.46
label-permutation control	0.19	0.18	0.18	0.18

Testbed 2: 22 systems, eleven recipes × two sizes of one base model

mean JS: same recipe, other size / other recipe, same size / both differ	0.063 / 0.148 / 0.161
40-turn probe picks own system / own recipe / own size (chance 1/22, 2/22, 11/22)	21/22 / 21/22 / 22/22

model; questions close in four of the five (Llama places them beside its advice). Within that order, GPT spends half of its highlighted behaviour on advice, Claude places its questions latest, Llama sits farthest from every other model (Jensen–Shannon distance ≥ 0.170), and Qwen and Gemini are near twins (0.019, half the distance of the next-closest pair). These relations hold corpus by corpus: Qwen–Gemini is the closest of the ten pairs and Llama the most distant model in every one of the four corpora (Table B.18), so the profiles are properties of the systems, not of one data source. The transition half shows the same grammar: the strongest chains are behaviours sustaining themselves, empathy after empathy and question after question, while advice self-chains only weakly because advice is already common (Appendix B.4.1, Fig. B.5). The one systematic failure of identification is a result in its own right: Qwen and Gemini, from different labs, are behaviourally interchangeable at the resolution of the clinicians' scheme. Table B.16 shows that most of the identification accuracy comes from the behaviour mix and the rest from the arrangement.

The profile tracks the training objective, not the parameter count. Testbed 2 shows what moves a profile. Across eleven recipes applied to two sizes of one base model, profiles of the same recipe at the two sizes sit close (mean JS 0.063) while different recipes at one size sit far apart (0.148); prompt-level interventions move the vanilla profile by 0.06–0.10 and RL recipes by 0.15–0.44. A 40-turn probe scored against all 22 profiles picks its own system in $21/22$ cases (chance $1/22$), the same recipe in $21/22$, and the same size in $22/22$ (Table B.17, bottom; Appendix C.3). The score says how scripted; the profile says whose script.

B.4.1 Profile structure and its stability across corpora

The score answers "how scripted"; everything model-vs-model lives in the profile. Here we examine the profile itself: what it contains, and whether its model distinctions are properties of the models or accidents of one data source.

Profile structure. The position half of every profile (Fig. B.3b) makes the template legible directly. The entire empathy block (accurate and inaccurate alike) is slammed against the opening decile for every model; recommendations (RECT) occupy the mid-late middle; open and closed questions concentrate in the final two deciles — with model-specific signatures visible at code level: Claude's open questions are the most extreme end-loading in the data, Llama's SAC (accurate sup-

Table B.18: **Model vs. model, dataset by dataset.** Left block: the twin structure recomputed independently inside each corpus — JS distance of the Qwen–Gemini pair, the next-closest pair, the separation ratio (next-closest / twin), and Llama's mean distance to the other four models. The twins are the closest pair in every corpus; Llama is the farthest model in every corpus. Right block: leave-one-corpus-out identification (enroll on three corpora, identify all of each model's responses on the held-out corpus) by mean per-event log-likelihood and by corpus-scale JS matching; 5 = all five models correctly identified. All 6 misses across the two scoring rules are Qwen↔Gemini confusions.

	Twin structure within corpus					Held-out ID (of 5)	
Corpus	Qwen–Gem.	next closest		ratio	Llama dist.	loglik	JS
counselchat	0.056	GPT–Gemini	0.075	1.4	0.161	4	4
pair	0.060	GPT–Gemini	0.069	1.1	0.129	5	4
carebench	0.024	Qwen–GPT	0.038	1.6	0.225	4	5
hope	0.038	Qwen–GPT	0.065	1.7	0.156	4	4
POOLED	0.019	Qwen–GPT	0.038	2.0	0.175	17/20	17/20

port) spreads across the whole response, GPT ends on SAC (its sign-off encouragement habit). Figure B.5 plots the transition half as *lift*, $P(\text{current} \mid \text{previous})/P(\text{current})$: the strongest chains everywhere are self-chains — accurate empathy begets accurate empathy (lift 2.8 for Llama up to 14.4 for Gemini) and questions beget questions (3.1–10.5) — while advice self-chains only weakly (1.2–1.7) because advice is already the majority behaviour. Two individual habits stand out: Gemini is the only model whose accurate empathy launches directly into a question (lift 2.2, versus ≤ 1.2 for everyone else), and Claude's question chain (3.1) is the *weakest* of the five — its many questions are distributed through the response's end rather than fired in one burst.

Cross-dataset stability. If the profile distinctions were corpus artefacts, profiles of the same model computed on different corpora would scatter. They do not. Embedding all twenty (model $\times$ corpus) profiles by their pairwise JS distances, the points organise by model rather than by corpus. Table B.18 checks the two headline model-vs-model facts corpus by corpus. First, the twin structure: Qwen–Gemini is the closest pair in *every single corpus* (JS 0.024–0.060, always rank 1 of 10 pairs), and Llama is always the most distant model (mean JS 0.129–0.225). Second, identification: enrolling profiles on three corpora and identifying each model's responses on the held-out fourth — repeated for all four choices of held-out corpus — yields 17/20 correct by likelihood and 17/20 by JS matching (chance 4/20), and every single miss is a twin-pair confusion (Gemini↔Qwen); Llama, GPT, and Claude are identified 24/24 times across all corpora and both scoring rules. The profile separates the models, it does so on every dataset, and its one systematic failure is itself a finding: two models trained by different laboratories are behaviourally interchangeable at the resolution of the clinician's coding scheme.

Identification and its transfer cost. Fig. 3a and Table B.16 report rank-1 accuracy against probe size k under four conditions of increasing difficulty. Enrolling on three corpora and probing the held-out fourth, identification rises from 38% at $k{=}1$ to 74% at $k{=}20$; restricted to a single corpus with disjoint enrol and probe responses it tracks slightly lower (34% to 65%), simply because the reference profiles are then built from a third of the data — the fingerprint itself transfers across corpora essentially intact. Held out to an unseen *annotator* it costs more (38% to 44%), and when corpus and annotator are both new it settles at roughly twice chance (36–46%) — still informative, and still growing with k, but the honest ceiling for this instrument: a fingerprint read through a different clinician's hand is a partially different fingerprint. A label permutation through the identical pipeline sits at chance throughout (rank-1 0.18–0.19), which rules out the possibility that any of this is an artefact of the matching procedure. Residual confusions concentrate on Qwen↔Gemini in every condition: the twin structure survives transfer. Figure B.4 walks through one round of the identification game, including the twin case.

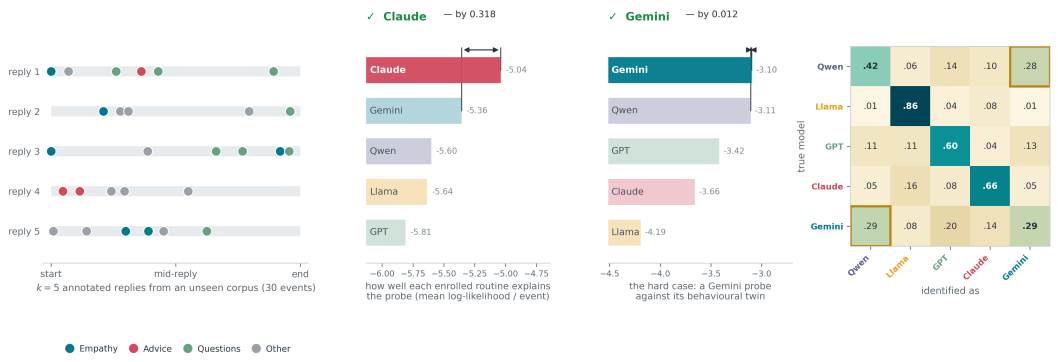


Figure B.4: **One round of the identification game.** *First panel:* the probe, five real annotated responses from the held-out corpus (carebench) with the model name hidden; each span is drawn at its position, coloured by behaviour group. *Second:* each enrolled reference profile, built on the other three corpora, scores the probe by mean per-event log-likelihood (Eq. 4); Claude wins, correctly, by a wide margin, and the signature is visible in the first panel: empathy opens the responses, questions sit at their ends. *Third:* the hard case, a Gemini probe identified correctly by 0.01 over its behavioural twin Qwen, with the other three models far behind. *Fourth:* the game repeated, 200 draws of k=5 responses per (model, held-out corpus), pooled over the four held-out choices (4,000 games; mean rank-1 accuracy 0.57, chance 0.20); the only systematic confusion is the boxed twin pair.

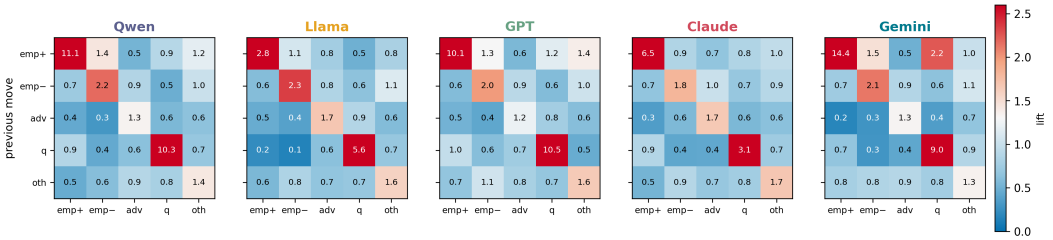


Figure B.5: **What follows what.** The transition half of each profile as lift, $P(\text{current} \mid \text{previous})/P(\text{current})$, over five behaviour groups (emp+ = accurate empathy, emp− = inaccurate empathy, adv = advice, q = questions). Red > 1: the move chains after the previous one; blue < 1: avoided. Self-chains dominate everywhere (the shared grammar); the individual accents are visible off-diagonal — e.g., Gemini is the only model whose accurate empathy launches directly into a question (lift 2.5), and Claude's question self-chain (3.2) is the weakest because its questions are spread through the response's end rather than fired in a burst.

B.5 The level survives the annotator, the estimator, and the task (§5.3)

The level, human direction, and choreography ordering survive replacing the clinicians with a blind LLM annotator. A sequential test certifies a blueprint from a few dozen annotated responses, and the level survives every estimation choice we varied. The fragile quantity is the fine ordering among models with similar scores, which is why model identity rests on the profile rather than on the scalar score alone.

The finding does not depend on the clinicians. The decisive annotator-independence check is the blind LLM layer, which re-annotates every response, the therapist's included, under the same twenty codes (74,526 spans, 96.7% located verbatim). It reproduces the level for all five models (0.085–0.118, z = 52–73), the human direction (0.053 for the therapist), and the two most seated models (Llama and GPT under both annotators), but not the fine ordering of SCRIPT ($\rho = 0.0$). The reason is visible span by span: the LLM highlights 90% of the text the clinicians highlighted and agrees on observable moves (open questions 73%, direct advice 50%) but almost never on the clinical judgments the manual asks for (the inaccurate-empathy codes $\leq 1\%$; Table A.3). It reproduces the

Table B.19: **Ablations and sensitivity checks on Testbed 1.** Each row summarises one check of Appendix B.5.1; ranges are across the five models unless a model is named, and every excess is measured over the appropriate within-response null. No claim about the existence or the level of the blueprint changes under any of these variations. The last two rows are why SCRIPT is reported beside a quality score rather than in place of one: at equal stratum size the sign of the quality gap is model-specific, and the few gaps that resolve run in favour of the better-rated stratum being more scripted.

Check	Result	z	Reading
Circular-shift null	excess 0.035–0.065	9–24	tracks C_{excess} (0.052–0.063): isolates choreography
Within-bin permutation null	excess 0.001–0.029	0.3–11	conservative lower bound on momentum; Llama's is at the floor
Midpoints instead of starts	C changes by ≤ 0.003	—	placement signal lives at span starts
Adding a duration decile	gain ≤ 0.0014	≤ 2.2	span length carries no structural information
Label granularity 20 / 13 / 5	0.083–0.118 / 0.102–0.127 / 0.095–0.114	≥ 21	level stable; held-out ID 4–5/5 at every granularity
Skip-gram momentum, lags 1/2/3	Claude $0.038 \to 0.004 \to -0.002$; Llama $0.042 \to 0.014 \to 0.001$	lag 1: 13–19	adjacent structure carries the momentum; lag 2 is 0.001–0.014 and lag 3 at zero for every model
Response-length terciles	13 of 15 cells 0.039–0.099; Llama short -0.065, mid 0.018	6.0–21.6	C flat in all 15 cells (0.052–0.067); M below range at 4.6 events/response
Accurate vs. inaccurate empathy, Δ raw / at matched n	Qwen $+0.008$ / $+0.057$; Llama -0.002 / -0.021; GPT -0.012 / $+0.031$; Claude -0.048 / -0.021; Gemini $+0.033$ / $+0.081$	—	matched sign model-specific; Gemini ($z = 2.8$) and Qwen ($z = 2.3$) resolvable
Unflagged vs. harmful-flagged, Δ raw / at matched n	Qwen -0.008 / $+0.025$; Llama -0.002 / -0.006; GPT $+0.001$ / $+0.021$; Claude -0.035 / -0.021; Gemini $+0.040$ / $+0.072$	—	same signs per model on both axes ($p = 0.03$); only Gemini resolvable ($z = 3.9$)

structural finding and not the clinical one, so the level and C transfer and the fine ordering does not. The other checks agree (Appendix B.5.2): re-dealing the corpus's own labels reads zero on both layers; within each clinician jury with enough events the level exists and Claude is the most scripted model in four of five; a fine-grained expert scheme on human text reads at the human level under experts and LLM annotators alike (Table B.23); and on Testbed 2 a position-blind labeller that sees one sentence at a time reproduces the human direction and the system ordering ($\rho = 0.71$ with tagger-layer C). Expert annotation is what made the finding visible; it is not what the measure depends on.

A few dozen responses certify a blueprint. Recomputing the fixed-sample z as annotated responses arrive and stopping at the first crossing of 1.96 rejected in 14 of 60 null streams (23%); SCRIPT-Seq rejected in none of 200 (guarantee ≤ 0.05). On the five models it certifies a blueprint in a median of 34–43 annotated responses at $\alpha = 0.05$ (50–61 at $\alpha = 0.01$; every stream by $t = 68$), against the ≈ 800 responses per model the fixed-sample analysis consumed. A template at three-quarters of the Testbed 1 strength needs a median of about 60 responses, at half about 130, at a quarter about 350, and at a tenth is certified in only 2 of 30 streams within a corpus-scale budget (Appendix C.5, Table C.9). The transfer pattern follows the score: Claude's counselling responses certify at $t = 32$ and the AnnoMI counsellors at $t = 93$, data-to-text error spans at $t = 311$ and RAGTruth hallucination spans at $t = 385$, and WMT24 never rejects. Anyone who already has an annotated evaluation set has what is needed to measure scriptedness.

Table B.20: **Uncertainty and estimator robustness.** Left: full-sample point estimate with 95% CI (centred paired cluster bootstrap over prompt items, 400 replicates, per-replicate nulls). Middle: estimator variants — choreography re-estimated with continuous-position kNN MI (Ross 2014, k=3, excess over the same null), and the order-2 momentum gain at group level. Right: SCRIPT recomputed within each within-model response-length tercile (z in parentheses). Bottom: the ten Holm-corrected paired tests summarised.

			Estimator variants		Length terciles		
Model	SCRIPT	95% CI	C^{kNN} (z)	$M^{(2)}$ gain (z)	short	mid	long
Qwen	0.091	[0.081, 0.102]	0.032 (10.2)	−0.011 (−4.5)	0.045 (7.8)	0.046 (8.0)	0.072 (12.1)
Llama	0.083	[0.069, 0.096]	0.051 (15.2)	−0.009 (−2.3)	−0.065 (−5.7)	0.018 (2.5)	0.071 (15.1)
GPT	0.098	[0.087, 0.108]	0.042 (15.2)	−0.003 (−1.3)	0.055 (8.0)	0.064 (11.0)	0.062 (11.0)
Claude	0.118	[0.107, 0.130]	0.045 (15.8)	−0.010 (−3.3)	0.045 (6.9)	0.092 (20.6)	0.099 (21.6)
Gemini	0.092	[0.082, 0.104]	0.038 (11.0)	−0.017 (−5.0)	0.039 (6.0)	0.060 (9.5)	0.063 (10.4)

Paired tests (10 pairs, Holm): Claude > each of {Qwen, Llama, GPT, Gemini}, $\Delta = 0.027$–0.033, $p_{\mathrm{Holm}} = 0.025$;
the six pairs among the other four have raw $p \geq 0.28$. 13 of 15 tercile cells $z \geq 6.0$; Llama's short tercile (4.6 events per response) is outside the operating range.

The level survives every estimation choice. Table B.19 summarises one check per row; Appendix B.5.1 gives each in full. Bootstrap intervals have half-width 0.010–0.013, so each score is known to about ±10–16% of its value, and the level holds under alternative nulls, span placements, label granularities, bin counts, both conventions for co-located spans (Appendix C.1), and within response-length terciles, where the only exception is Llama's shortest tercile, below the density threshold for M while C stays flat. No claim rests on a second-order term.

B.5.1 STATISTICAL ROBUSTNESS AND ABLATIONS

Every score is known to about ±10–16% of its value; the only pairwise ordering the score resolves is Claude above the other four; and no claim about the existence or level of the blueprint changes under any estimation choice varied here.

Because plug-in mutual information on discretised bins is a biased estimator and the shuffle null removes only its *mean* inflation, we quantify the residual uncertainty directly and stress-test every estimation choice. Table B.20 collects the headline estimates and tests; Table B.19 summarises the ablations, one row per check.

Confidence intervals and pairwise tests. We use a paired cluster bootstrap: prompt items are resampled with replacement (400 replicates), all five models are re-scored on the *same* resampled items (each replicate re-runs the full pipeline including its own 20-shuffle null), and 95% intervals are read from the centred bootstrap distribution around the full-sample point estimate. Per-model intervals have half-width 0.010–0.013, so SCRIPT is estimated to about ±10–16% of its value. Cluster resampling itself inflates the plug-in level by a common-mode +0.03–0.04 (duplicated responses sharpen the pooled tables faster than the within-response null can absorb), which is why inference uses the *paired differences*, where the common-mode bias cancels: for each replicate all models see identical items, so $\Delta^{(b)} = \mathrm{SCRIPT}^{(b)}_{m_1} - \mathrm{SCRIPT}^{(b)}_{m_2}$ is a clean paired statistic. Testing all ten pairs with Holm correction yields exactly one structure: Claude sits above each of the other four ($\Delta = 0.027$–0.033, 95% CIs excluding zero, Holm-adjusted $p = 0.025$, the resolution floor of 400 replicates), while no pair among Qwen, Llama, GPT and Gemini is distinguishable (Holm-adjusted $p = 1$; raw $p = 0.28$–0.96). The score resolves one model and a field-wide level; naming the other four requires the profile.

Per-corpus scores. Table B.21 scores each model inside each of the four corpora on its own. The two multi-turn corpora (2,200–4,200 events per model) are inside the operating range and reproduce the pooled picture, with Claude the most scripted model in both. The two single-turn corpora hold 400–650 events per model, below the ≈ 1,000-event floor: C is unchanged there (0.053–0.092, as everywhere), but M_{excess} is strongly negative (−0.05 to −0.12) because a few hundred transitions spread over ten bins and a 20×20 table leave the observed table more concentrated than its shuffles, so SCRIPT reads at or below zero for four of the five models. These cells are the density limit of M (Appendix C.3) and are not evidence of a different blueprint in single-turn prompts; Claude is nonetheless the highest-reading model in both.

Table B.21: **SCRIPT inside each corpus.** Left: SCRIPT (z) computed on each corpus alone and pooled (10 bins, 200 shuffles). Right: the choreography term C per corpus, which is unaffected by the event count. The single-turn corpora (counselchat, pair) hold 400–650 events per model and are below the operating range of M; read them through z and C.

	SCRIPT (z)					C			
Model	counselchat	pair	carebench	hope	pooled	counselchat	pair	carebench	hope
Qwen	−0.021 (−1.3)	−0.045 (−2.7)	0.076 (15.0)	0.055 (10.0)	0.091 (28.5)	0.068	0.064	0.057	0.047
Llama	−0.055 (−4.0)	−0.038 (−1.9)	0.059 (11.2)	0.025 (4.2)	0.083 (22.1)	0.061	0.075	0.064	0.073
GPT	−0.032 (−2.3)	−0.032 (−2.0)	0.080 (16.6)	0.069 (13.9)	0.098 (34.9)	0.061	0.067	0.054	0.070
Claude	0.015 (1.3)	0.040 (3.4)	0.101 (25.2)	0.082 (14.9)	0.118 (43.9)	0.066	0.092	0.053	0.059
Gemini	−0.031 (−2.2)	−0.047 (−2.6)	0.077 (15.7)	0.056 (9.7)	0.092 (28.9)	0.053	0.054	0.063	0.055

Annotator-level analysis. Repeating the paired bootstrap on six leave-one-annotator-out subsets, the gap $\bar{\Delta} = \text{SCRIPT}_{\text{Claude}} - \overline{\{\text{Qwen}, \text{Llama}, \text{GPT}, \text{Gemini}\}}$ survives the removal of any single annotator ($\bar{\Delta} = 0.028$–0.035, every 95% CI excluding zero, $p = 0.008$, the floor of 120 replicates), including R3, whose items carry 40% of all span events; Claude's margin over the runner-up is 0.017–0.026 in every subset. The strictest analysis treats the annotator as a random effect: each jury's gap is estimated by a within-jury paired bootstrap and the six estimates are pooled by DerSimonian–Laird random-effects meta-analysis. The pooled gap is 0.042 ± 0.008 ($z = 5.3$, $\tau = 0.011$, $I^2 = 37\%$); R3, R4 and R6 are individually significant ($z = 4.6$, 4.7, 4.0), R1 and R2 are positive but not ($z = 1.2$, 1.8), and R5 is too small to read. We therefore state the ordering as established at both the item and the annotator level, while noting that it is the one fine ordering the score resolves; the other four models tie, and model identity rests on the profile.

Estimator variants. Two alternative estimators probe the two structural terms (Table B.20, middle block). Replacing the binned $I(L; X)$ with the Ross (2014) k-nearest-neighbour estimator for discrete–continuous mixtures (k=3, positions unbinned, same null) gives $C^{\text{kNN}}_{\text{excess}} = 0.032$–$0.051$ at $z = 10$–16 — somewhat below the binned values, since the continuous estimator conditions more finely, but with the same conclusion and the same extremes (Llama highest, Qwen lowest under both estimators). Choreography is not a binning artefact. The order-2 momentum gain $I(L; L_{\text{prev2}} \mid X, L_{\text{prev}})/H(L)$, computed at the 5-group level where the tables are dense enough to support it, is -0.017 to -0.003 ($z = -5.0$ to -1.3): no second-order predictability is resolvable, and the small negative values are the bias mismatch described in Appendix C.3. Seat plus one step of memory is the whole of the measurable routine.

Alternative null models. The default null permutes labels freely within each response; two stricter nulls that *preserve* part of the structure isolate the two terms from the other side (Table B.19, rows 1–2). A within-response *circular shift* keeps the label sequence's adjacency structure while destroying its alignment with position, so its excess should approximate C alone — and it does for four models (0.049–0.065 against $C = 0.052$–0.060); for Llama it reads 0.035 against $C = 0.063$, the one model whose sequence structure is weak enough for the shift to remove part of it. A *within-bin* permutation preserves the label–position joint exactly and scrambles only within-bin order, giving a conservative lower bound on M: 0.012–0.029 ($z = 3.5$–11.4) for four models and 0.001 for Llama, whose momentum under the slot convention is the smallest of the five. Under the reference null and the circular shift, every model exceeds chance by double-digit z; and the agreement between null-based isolation and the chain-rule decomposition is itself a validation of the estimator.

Span-length ablation. The metric deliberately uses only span *starts*, because span extent is the most annotator-dependent choice in span data. Rows 3–4 of Table B.19 confirm the choice costs nothing: re-estimating with midpoints, or augmenting position with a duration decile, changes essentially nothing. Span length carries no structural label information beyond the start position.

Response-length dependence. Splitting each model's responses into within-model length terciles and re-scoring each tercile separately leaves SCRIPT significant in 13 of the 15 model×tercile cells (Table B.20, right block; 0.039–0.099, $z = 6.0$–21.6). The one failure is Llama's shortest tercile; its middle tercile is marginal but positive (0.018, $z = 2.5$), and the shortest, at 4.6 located spans per response and $1{,}263$ events, reads -0.065. This is the density limit of M described in Appendix C.3

and not absent structure: C is flat across all fifteen cells (0.052–0.067), and Llama's long tercile reads 0.071 ($z = 15$). Within the operating range, the template is not a mixture artefact of pooling short and long responses, and Claude is the most scripted model in every tercile but the shortest.

Implementation specification. For exact reproducibility of the statistical choices: entropies are plug-in in bits with $0 \log 0 = 0$ (zero cells contribute nothing; no smoothing is applied anywhere in the *metric* — smoothing, $\alpha = 0.5$, exists only at profile-*comparison* and probe-scoring time); rare labels are retained, never merged — their small-sample noise is absorbed by the null, which preserves each response's exact label multiset; a label set is the sorted set of labels observed in the system being scored; the position bin is $\min(\lfloor xB \rfloor, B-1)$, so x exactly at the end of the response falls in the last bin; events are sorted by (response, position) with a stable sort; co-located events form a slot and no transition is formed between them (Appendix C.1); a bin with fewer than two transition events contributes zero to M; $H(L) = 0$ (a single label) returns all three terms as zero; duplicate (label, position) pairs count once; the null uses `numpy`'s default generator with seed 0 and permutes labels independently within each response; the estimator is deterministic given the seed.

Relation to expert quality judgments (and why we do not claim it). Stratifying each model's responses by the expert ratings and re-scoring each stratum as is (Table B.19, bottom rows) gives gaps of either sign, and the comparison is confounded by our own operating range: the better-rated stratum is in almost every case the rarer one (393–1,432 events for accurate empathy against 1,993–6,469 for inaccurate), and SCRIPT is downward-biased below $\approx 1{,}000$ events (Table B.22), so the smaller stratum is expected to read lower whatever its behaviour.

We therefore re-scored both strata at a matched response count: 20 resamples per side at $\min(n_{\text{good}}, n_{\text{bad}})$ responses, 100-shuffle null each, the larger stratum subsampled without replacement (this is what removes the bias from the contrast) and the smaller bootstrapped, so both sides carry a sampling spread. At matched size the accurate-empathy stratum is *more* scripted for Qwen (+0.057), GPT (+0.031) and Gemini (+0.081) and less scripted for Llama (−0.021) and Claude (−0.021); Gemini's gap ($z = 2.8$) and, more weakly, Qwen's ($z = 2.3$) clear the resampling spread. The harmful-flag axis gives the same five signs (Qwen +0.025, GPT +0.021, Gemini +0.072; Llama −0.006, Claude −0.021), again with Gemini the only resolvable model ($z = 3.9$).

The sign agreement reported in §5.3 is what remains. Every model keeps the sign of its matched gap when the rating axis changes from empathy accuracy to harmful content, and the two axes were rated separately and measure different things, so under independent signs the agreement of all five would occur with probability $2^{-5} = 0.03$. That is the only part of the quality relation we treat as established; its magnitude we cannot resolve for four of the five models at this annotation volume, which is why no claim in the paper rests on it.

Sensitivity to label granularity. Re-running everything under three granularities of the label scheme — the full 20 codes, 13 codes (the eight empathy codes merged into accurate/inaccurate), and 5 coarse groups — leaves the score's level and significance stable and held-out identification essentially unaffected (Table B.19, row 5), with the expected estimator movements: merging empathy codes raises M by removing co-label churn, and coarsening to 5 groups raises it further for the model with the sparsest transition table (Llama, $0.020 \to 0.050$). One caution follows: fine-grained score *orderings* are not ontology-invariant (at 5 groups Llama and Claude tie at the top), one more reason the paper rests model identity on the profile, never on the score ordering.

Higher-order sequential structure. Beyond the order-2 conditional gain, skip-gram probes at the group level — $I(L; L_{\text{prev}k} \mid X)$ for lag $k = 1, 2, 3$, each in excess of the within-response null — show structure decaying with lag for every model (Table B.19, row 6): lag-1 dependence of 0.025–0.042 ($z = 13$–19), lag-2 of 0.001–0.014 (resolvable for Llama, GPT and Qwen), and lag-3 at zero. Under the slot convention the lag is counted in distinct positions, so the lag-2 dependence is a real if small longer-range organisation, strongest for Llama; a variable-order model would sharpen these tails but cannot change the headline: adjacent structure carries the bulk of M.

Operating for small samples. Subsampling experiments give concrete operating guidance, collected in Table B.22. The one behaviour worth flagging in prose is the direction of the small-sample error: the null-subtracted point estimate is *downward* biased at small n, because the signal has not

Table B.22: **Operating guidance from the subsampling experiments.** Observed estimator behaviour and the practical recommendation that follows, for each free choice a user of the metric must make.

Choice	Observed behaviour	Recommendation
Sample size n	near zero at 25 responses; ≈half the asymptotic value at 100 responses (≈1,000 events, $z \approx 4$–9); stable (SD < 0.005) from ≈400 responses / 4,000 events	below 1,000 events report z only and read SCRIPT as a lower bound; from ≈2,000 events point values are comparable; subsample extrapolation recovers the asymptote from ≈400 responses; cluster-bootstrap CIs whenever n permits
Position bins B	score plateaus for $B = 5$–10, declines for $B \geq 20$ (sparsity pushes real structure into the null)	default $B = 10$; $B = 5$ below ≈1,000 events; never compare systems scored at different B
Shuffles S	point estimate essentially independent of S (SD ≤ 0.0014 even at S=25); z noisy at small S (SD ≈ 7 at S=25, 2.7 at S=200, 1.1 at S=500, pooled corpus)	$S = 200$ default; $S \geq 500$ when a borderline $z \in [2, 4]$ is itself the claim
Estimator	binned plug-in is deterministic, fast, exactly matched to its null; kNN systematically more conservative (finer position conditioning)	binned for reporting; kNN as cross-check near decision thresholds, for genuinely continuous positions, or small alphabets; publish both when they disagree in rank
Probe size k (identification)	cross-corpus transfer: $k = 5$ gives 57% rank-1, $k = 20$ gives 74% (chance 20%), full corpus 4–5/5	for verification-grade decisions (≥95%), $k \leq 20$ is insufficient — annotate more or restrict to corpus-scale JS matching

yet risen above the noise floor the null removes (Fig. B.6). An under-sampled SCRIPT therefore understates scriptedness; it never invents it. The bias can also be corrected directly, in the style of entropy-extrapolation estimators: subsample the corpus at several fractions, fit SCRIPT against $1/n$, and report the extrapolated intercept. Validated against the known full-sample scores on Testbed 1 (five draws at each size), the extrapolated estimate recovers most of the asymptote from ≈ 400 responses (Claude 0.115 ± 0.007 against a truth of 0.119; Llama 0.070 ± 0.005 against 0.083), roughly halves the bias at 200 responses (Claude 0.091 against 0.119), and remains conservative — it reduces the bias, never inflates the score — at 100. Below that the subsample curve is not identifiable, and the correct instrument is z, which stays valid at any n.

Cross-scheme calibration cards. Because the label ontology fixes $H(L)$, the alphabet size, and the null level, raw SCRIPT values are comparable only *within* a scheme. For cross-scheme interpretation we recommend publishing, with every application, a *calibration card*: scheme, $|\mathcal{L}|$, $H(L)$, event count, null mean $\bar{R}_{\text{null}}$, SCRIPT, and z, together with the position of the value on the measured anchor ruler (random-label control ≈ 0; WMT24 MT spans ≈ 0; RAGTruth hallucination spans 0–0.03; data-to-text error spans ≈ 0.05; human-written news under the 18-technique propaganda scheme ≈ 0.02; human counsellors under the AnnoMI MI scheme ≈ 0.05; counseling behaviour codes ≈ 0.10). The card makes the non-transferability visible at a glance: in our own data the null level alone ranges from 0.074 (Mistral, RAGTruth: 4 labels, 3,562 events) to 0.353 (GPT-4, RAGTruth: 485 events) to 0.162 (Claude, counseling: 20 labels, 8,502 events) — differences a raw score comparison would silently absorb. New anchor tasks extend the ruler; they never replace the within-scheme comparison.

Matched-ceiling normalisation. One derived quantity does travel. Re-arranging each corpus's own events into their most scripted achievable order — same responses, same positions, same per-response label multisets, labels re-dealt in a single fixed position order — and scoring that arrangement gives the highest SCRIPT the corpus's composition and density permit. This *matched ceiling* is the maximum of exactly the invariance class the null averages over, so the null mean (0) and the ceiling are the two ends of one scale, and the observed score divided by the ceiling is a *fraction of achievable rigidity* in $[0, 1]$, comparable across schemes. Measured on every corpus in this paper:

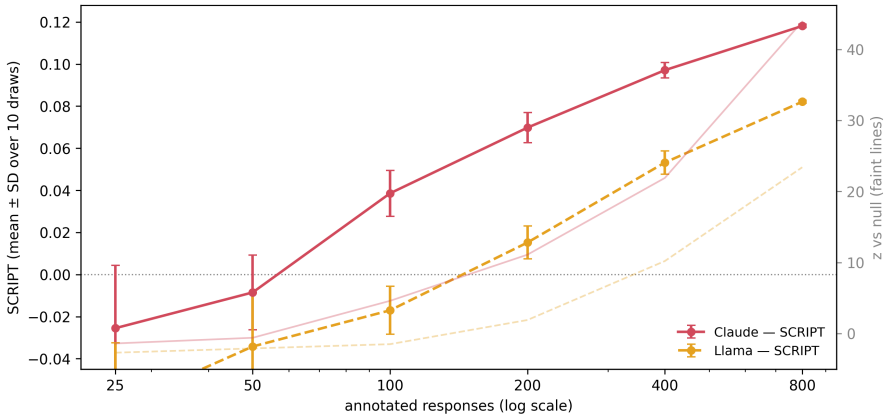


Figure B.6: **The evidence ladder.** SCRIPT (mean ± SD over 10 subsample draws) and its z (faint lines, right axis) as a function of the number of annotated responses, for the strongest- and weakest-scripted models. The point estimate is *downward* biased at small n and stabilises from ≈400 responses; significance arrives much earlier ($z \approx 3$ by 50–100 responses). Practical rule: below ∼1,000 events quote z and read the value as a lower bound.

the five counseling models use 19–25% of achievable rigidity (ceilings 0.39–0.49); data-to-text error spans 15%; WMT24 MT spans 7%; RAGTruth hallucination spans 4%; the human authors of the propaganda corpus 6% (ceiling 0.40, close to the counseling models' 0.39–0.49, so here the raw scores and the fractions tell the same story). The most instructive case is AnnoMI: its raw score (0.047) reads at half the counseling level, but its session-scale 4-label scheme permits near-perfect arrangement (ceiling 0.82), so the human counsellors use only 6% of the rigidity available to them — normalisation does not merely repair cross-scheme comparability, it sharpens the human–model contrast the raw ruler understates.

Token-level labelling. Nothing in the definition requires character-offset spans. Token-level BIO tagging maps onto the same estimator by collapsing each contiguous `B–I` run to a single event at the run's opening token, with position the token index normalised by response length — which is exactly the (label, start) reduction used here. The one non-negotiable adaptation is the null: permuting *tokens* would break intra-run cohesion and manufacture spurious momentum (every `I` follows a `B` by construction), so the permutation must operate on runs, just as ours operates on spans. With run-level events and a run-level null, estimator and interpretation carry over unchanged.

Procedure summary. *Score:* events → bins → $R = [I(L;X) + I(L;L_{\text{prev}}|X)]/H(L)$; repeat on S within-response label shuffles; report $R - \bar{R}_{\text{null}}$ and z. *Profile:* the two count tables $P(\ell, x)$ and $T(\ell_{\text{prev}}, \ell)$, stored unsmoothed. *Distance* (two corpus-scale systems): add $\alpha = 0.5$ to every cell, normalise, average the Jensen–Shannon distances of P and T. *Identify* (small probe vs. enrolled references): score each reference by $\frac{1}{n}\sum_i[\log \hat{P}(\ell_i \mid x_i) + \log \hat{T}(\ell_i \mid \ell_{i-1})]$ (rows α-smoothed, renormalised) and pick the argmax — never JS on a sparse probe.

Generality across annotation schemes. To push scheme-agnostic generality beyond the two schemes already reported, we ran the identical implementation on RAGTruth — human-annotated hallucination spans (4 label types, character offsets) over six generators' outputs (Table C.2). The metric runs unchanged and reads sensibly: GPT-4's hallucination spans show no arrangement structure (SCRIPT = 0.002, $z = 0.1$), the Llama-2 family shows mild structure (0.009–0.017, $z = 1.7$–4.7), and Mistral-7B the most (0.030, $z = 8.0$). Weaker generators do not only hallucinate *more*; their hallucinations follow positional and sequential habits. With counseling codes (≈ 0.10), data-to-text error spans (≈ 0.05), human counsellors under the AnnoMI motivational-interviewing scheme (≈ 0.05, and momentum-dominated where the models are choreography-dominated), human-written news under an 18-technique expert scheme (≈ 0.02), MT error spans (≈ 0), and hallucination spans in between, the instrument has now been exercised on six unrelated annotation schemes across its full range.

Table B.23: **SCRIPT on the propaganda-technique spans of the Propaganda Techniques Corpus (Da San Martino et al., 2019), as released by Kasner et al. (2025).** 101 human-written news articles; 18 techniques; position = span start over article length; 10 bins, 200 within-article shuffles. Top: the expert annotation and its random-label control. Bottom: the release's LLM-annotator layers on the identical articles (setup / annotator), ordered by event count; below ∼1,000 events the value is a lower bound and z is the quantity to read.

Layer	SCRIPT	z	C	M	raw R	null $\bar{R}$	events
Human experts (Da San Martino et al.)	0.022	3.4	0.008	0.014	0.294	0.272	1,439
Random-label control (same events)	0.002	0.3	0.005	−0.003	0.274	0.272	1,439
zeroshot / Gemini 2.0 Flash Thinking	0.031	5.8	0.013	0.018	0.257	0.226	1,864
5-shot / DeepSeek-R1	0.007	0.9	0.011	−0.004	0.326	0.319	983
zeroshot / Claude 3.7 Sonnet	0.026	2.4	0.012	0.014	0.379	0.353	803
5-shot / Llama 3.3	0.006	0.6	0.010	−0.004	0.330	0.324	778
zeroshot / Llama 3.3	0.024	1.8	0.024	−0.000	0.502	0.478	574
CoT / DeepSeek-R1	0.017	1.4	0.017	0.000	0.537	0.520	478
zeroshot / DeepSeek-R1	0.015	1.2	0.015	−0.000	0.528	0.513	459
CoT / Llama 3.3	0.016	1.1	0.044	−0.029	0.587	0.571	411
zeroshot / o3-mini	0.015	0.8	0.017	−0.002	0.412	0.397	376
zeroshot / GPT-4o	0.059	2.5	0.013	0.046	0.627	0.568	246

Human-written news under an expert 18-technique scheme. The propaganda anchor (Table B.23) is the one corpus on the ruler whose instrument resembles Testbed 1's clinician layer — many nuanced, partly overlapping codes, trained expert annotators, long texts — and whose authors are human. It comes from the release of Kasner et al. (2025), who used the test split of the Propaganda Techniques Corpus (Da San Martino et al., 2019) to evaluate LLMs as span annotators; we use the expert spans, and their LLM-annotator layers, as data about the authors. Each article is one response; each expert span is one (technique, start offset / article length) event; 10 bins, 200 within-article shuffles, the same estimator as every other row. Two results. First, the level: 0.022 ($z = 3.4$) against a random-label control of 0.002 ($z = 0.3$), with a matched ceiling of 0.40 and hence 6% of achievable rigidity. A fine-grained expert scheme does not by itself produce the counseling models' 0.10 and 19–25%; on human authors it produces the human level. The reading is a lower bound: at $1{,}439$ events the corpus sits near the floor of the operating range (Fig. B.6), so the point estimate is downward-biased and z is the quantity to trust. Second, the composition: momentum (0.014) exceeds choreography (0.008), so the techniques chain — repetition follows repetition, name-calling clusters with loaded language — without settling into fixed positions. The plug-in values make the null's role visible: raw $R = 0.294$ against a null mean of 0.272, i.e. 92% of the uncalibrated mutual information is sparse-table inflation from eighteen labels spread over ten bins and eighteen possible predecessors.

B.5.2 INSTRUMENT VALIDATION

Because the metric is itself a claim, we verified it beyond reruns. *(i) Independent reproduction:* a from-scratch re-implementation of the event extraction (own normalisation, own span location, own deduplication) re-derived all 39,696 events from the raw released CSVs (100% of spans located verbatim) and reproduced every pooled score in Table 2 and every JS distance to within ± 0.001. *(ii) Synthetic ground truth:* on constructed corpora with known generating processes the metric recovers exactly the intended quantity — a perfect slot template scores SCRIPT $= 0.853$ with the signal entirely in C; label chains with uniform positions score 0.238 with the signal entirely in M ($C = 0.001$); matched random data scores 0.0008 ($z = 0.7$). The C/M decomposition does not leak. Known-answer sweeps sharpen the negative control (Fig. C.2): on content-driven generators (labels i.i.d., true $R = 0$), raw R climbs from 0.05 to 0.09 as the label skew grows from uniform to $90/10$ at fixed N, and from 0.01 to 0.33 as the corpus shrinks to ten responses at fixed skew, while SCRIPT stays within ± 0.011 of zero throughout, with C_{excess} and M_{excess} individually flat at the same tolerance — the two artefacts of §3.2 are cancelled separately, not merely in the sum. Under the same i.i.d. generator the permutation-null mean reproduces the analytic Miller–Madow plug-in bias term by term (tracking $(r-1)(c-1)/(2N \ln 2)$ per table to within a few percent by 320 responses, and sitting above it exactly where the asymptotic formula under-corrects sparse tables): the null's

mean is the known small-sample bias, not an ad hoc offset. *(iii) Estimator robustness:* halving or doubling the bin count moves pooled scores by at most 0.017 ($B = 5$) and 0.025 ($B = 20$), with M the term that moves (Llama, the model with the sparsest transition table, moves most); Claude stays highest at every B, and the ordering of the other four is not bin-invariant, consistent with the paired tests above. Recomputing within each annotator jury separately: R4 (2,300–3,200 events per model) reads 0.056–0.123 ($z = 8$–23), R1 and R2 (550–800 events per model) read 0.062–0.127 ($z = 5.6$–11), and R3, the heaviest annotator (2,100–3,800 events per model, 40% of all events), reads 0.023–0.056 ($z = 3.9$–10.7) for four models and null for Llama (-0.004). The two lightest juries behave as the evidence ladder predicts: R6 (260–380 events per model) reads $z = 4.1$–7.3 for four models and 1.1 for Llama, and R5 ($\approx$135 events per model, below the $\sim$250-event floor of Fig. B.6) reads null throughout. The level therefore exists in every jury with enough events but its size is the coder's as well as the model's, and Llama, the model with the least momentum, is the one whose reading a sparse jury can push to the floor; the exact ordering of the four tied models is not stable across juries, which is why we claim only the ordering the pooled tests and the annotator-level meta-analysis agree on (Claude most scripted). *(iv) Negative controls:* the random-label control and the WMT24 null (Table C.2) show the metric stays silent where structure is absent.

C THE ESTIMATOR IN DETAIL

This appendix supports §3: the conventions the momentum term needs, the calibration of the null, the sensitivity of the estimator, the construction and matching of the profile, and the sequential test.

C.1 TRANSITION CONVENTIONS FOR THE MOMENTUM TERM

The level of the blueprint holds under both conventions for co-located spans ($z \geq 22$ for every system); what the convention changes is the momentum term of individual systems, and with it their fine ordering.

The momentum term (Eq. 1) needs a predecessor for each event, and two features of span data make the predecessor a matter of convention: the first span of a response has none, and spans that start at the same character have no order. Both conventions are fixed by positions alone, so they are applied identically to the observed data and to every null draw; observed and null statistics are always produced by the same estimator, and the conventions cannot bias the calibrated score. They do decide what the momentum term reads, which is why they are stated in §3.2 and tested here.

Opening moves. The first slot of a response enters no transition. M is therefore averaged over *transition events*, the events that have a predecessor, and its position weights $p_T(x)$ range over those events, while $H(L)$ and C use every event. On a common event set the chain rule gives $C + M = I(L; X, L_{\text{prev}})/H(L)$ exactly; with the first slots excluded from M, $R = C + M$ is that quantity up to the predictability of opening moves, which it does not count. The alternative of giving the first span a virtual start-of-response predecessor would make the identity exact but would let M read the choice of opening move, which is a positional fact already captured by C; we do not use it.

Co-located spans. When the same text is highlighted under two codes, the two events share a start position. The estimator treats them as one multi-label *slot*: a transition is a pair of adjacent events at strictly increasing positions, and no transition is formed inside a slot. The legacy alternative chains co-located codes in the stable order of the input, which in Testbed 1 is the order of the codes in the annotation export. The difference matters because co-location is frequent: 33% of adjacent pairs on the clinician layer and 19% on the LLM layer share a start (Table C.1). Under the legacy convention the most frequent chained pairs are the same sentence coded directive and fix-it, directive and topic-shift, or directive and inaccurate validation, in every case in export order, so part of what that convention reports as sequence is co-annotation, and the direction of the "transition" is the interface's. Table C.1 recomputes every system under both conventions. C is unchanged by construction. The level survives for every system under both ($z \geq 22$ on the clinician layer, ≥ 52 for the models on the LLM layer), so no claim about the existence or the field-wide level of the blueprint depends on the convention. The momentum component of individual systems does move: on the clinician layer Llama's M halves ($0.042 \to 0.020$) and Claude's rises ($0.056 \to 0.064$), and

Table C.1: **The momentum term under the two conventions for co-located spans.** Every system on the clinician layer (20 codes, pooled corpora) and on the blind LLM layer (§3.4, six speakers), 10 bins, 200 shuffles, seed 0. *Tied share*: fraction of adjacent same-response pairs that share a start position. *Slot* is the estimator of §3.2 (no transition inside a slot), used throughout the paper; *chain* is the legacy convention that orders co-located codes by the export order. C is identical under both by construction; the level survives everywhere; the M of individual systems, and hence their C/M split and fine ordering, depends on the convention.

					Slot convention (paper)			Chain convention (legacy)		
Layer	System	events	tied share	C	SCRIPT	M	z	SCRIPT	M	z
clinician	Qwen	8,292	0.35	0.052	0.091	0.040	28.5	0.098	0.046	40.5
clinician	Llama	6,108	0.35	0.063	0.083	0.020	22.1	0.105	0.042	37.9
clinician	GPT	8,843	0.33	0.060	0.098	0.038	34.9	0.098	0.038	44.4
clinician	Claude	8,502	0.27	0.055	0.118	0.064	43.9	0.110	0.056	56.2
clinician	Gemini	7,951	0.36	0.057	0.092	0.035	28.9	0.103	0.045	44.9
LLM	Qwen	14,686	0.23	0.063	0.085	0.022	52.4	0.108	0.045	76.2
LLM	Llama	9,842	0.19	0.086	0.118	0.032	61.3	0.145	0.059	85.1
LLM	GPT	15,198	0.17	0.083	0.108	0.025	73.4	0.124	0.041	102.7
LLM	Claude	13,771	0.15	0.063	0.114	0.050	69.0	0.123	0.060	86.6
LLM	Gemini	14,341	0.22	0.063	0.095	0.032	56.9	0.117	0.053	82.2
LLM	Human therapist	4,130	0.21	0.045	0.053	0.008	12.8	0.094	0.049	27.8

Table C.2: **The ruler: controls that fix the scale.** The same computation as Table 2 on constructed corpora with a known answer, on the random-label control, and on three annotation schemes from unrelated tasks (sizes and provenance in Appendix A.4; per-generator values in Appendix B.5.1). †The eighth WMT24 system (140 events) reads 0.20 ($z = 2.6$) and is the one reading on this row outside noise.

Corpus	SCRIPT	C	M	z	expected	events
Known-answer ends, and the random-label control						
Pure slot template (synthetic)	0.853	≈0.85	≈0	—	all in C	
Pure label chain, uniform positions (synthetic)	0.238	0.001	0.237	—	all in M	
Matched random data (synthetic)	0.001			0.7	zero	
Random-label control (Claude's events, re-dealt)	−0.004	−0.001	−0.003	−1.8	zero	8,502
Annotation schemes from unrelated tasks						
MT error spans, WMT24 (7 of 8 systems†)	−0.028–0.042			$\lvert z\rvert \leq 1.2$	near zero	1,696
Hallucination spans, RAGTruth (6 generators)	0.002–0.047			≤ 8.0		14,289
Data-to-text error spans (4 generators)	0.037–0.060			≤ 10.7		6,119

on the LLM layer the human therapist's M almost vanishes ($0.049 \rightarrow 0.008$) while every model keeps half to four-fifths of its own.

What the term picks up in the data. Figure C.1 makes the momentum term concrete on Testbed 1's most chained model. M is a sum of pulls — pairs of codes where knowing the current span changes the odds of the next one beyond what its position already implies. Most of those pulls are a code sustaining itself across consecutive spans; the strongest *cross*-code pull is the one the figure follows end to end: after a validation span, direct advice arrives at nearly twice the rate that position alone predicts. *Validate, then advise* is a habit of order, invisible to any frequency count and to the choreography term, and it is exactly what the conditioning on X in Eq. equation 1 isolates.

C.2 VALIDATION OF THE NULL CALIBRATION

On content-driven generators with true $R = 0$, raw R climbs with label skew and with shrinking sample size, while SCRIPT stays flat at zero (Fig. C.2), confirming that the within-response permutation cancels both composition and small-sample bias. The same draws calibrate the decomposition, not only its sum: C and M are reported in excess of their own null means,

$$C_{\text{excess}} = C_{\text{obs}} - \overline{C^{(s)}}, \qquad M_{\text{excess}} = M_{\text{obs}} - \overline{M^{(s)}},$$

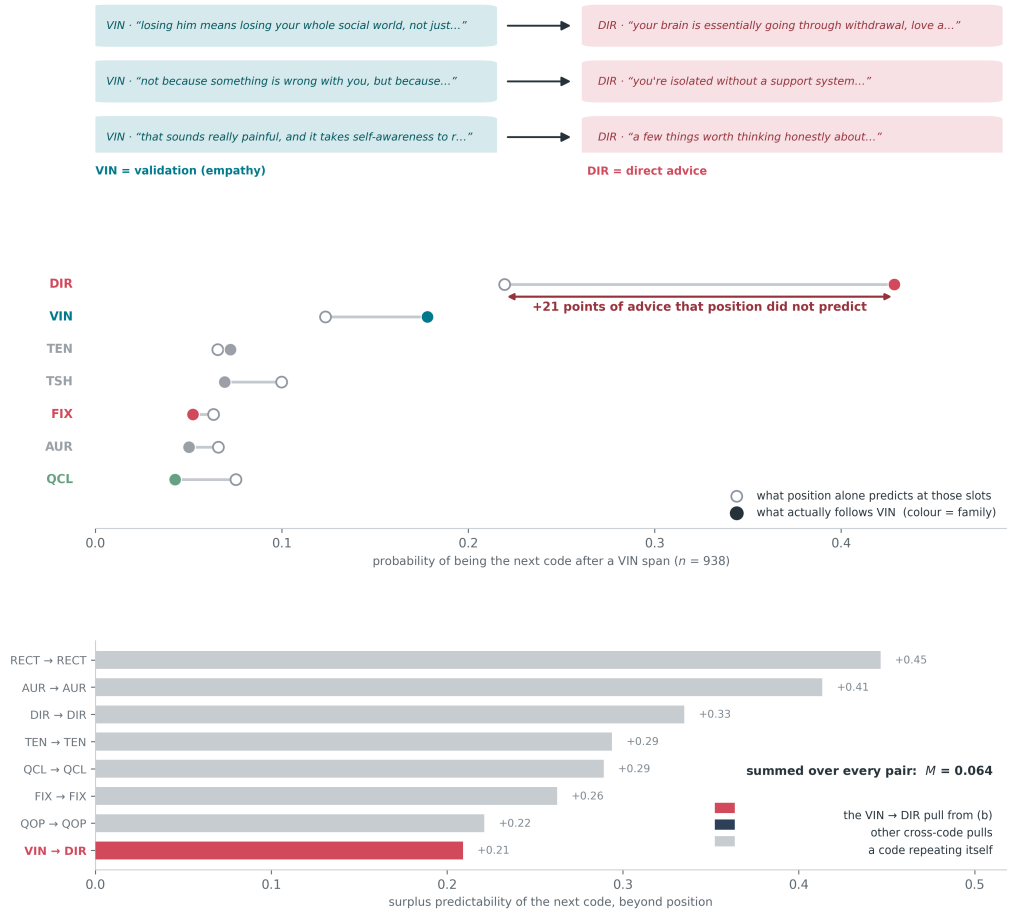


Figure C.1: **What the momentum term measures.** *Top:* three verbatim pairs from Claude's responses: a validation span (VIN) followed at the next distinct position by direct advice (DIR). *Middle:* the same habit as the metric sees it. Across all 938 VIN spans that have a successor, the code that actually follows (filled) is set against what position alone would predict at those same slots (open): advice follows validation 43% of the time where position predicts 22%, twenty-one points of advice that no fixed seating plan explains. *Bottom:* that pull in context, the largest surpluses in Claude's transition table. Grey bars are a code sustaining itself; the coloured bars are cross-code moves. Summed over every pair of codes, these surpluses are the momentum term, $M = 0.064$.

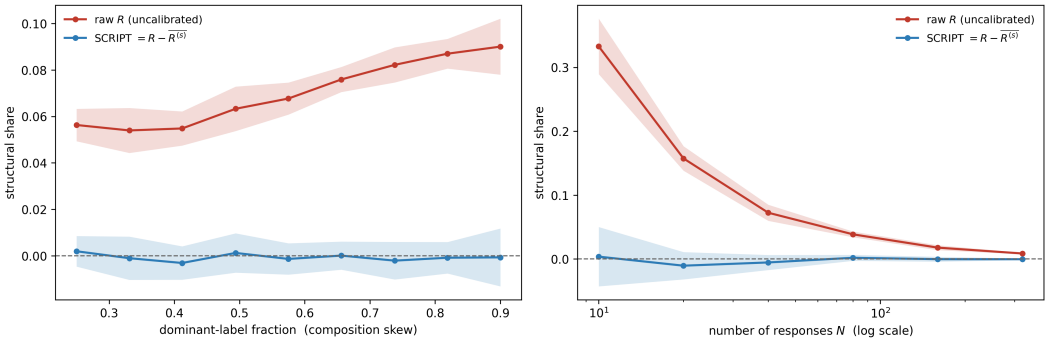


Figure C.2: **Known-answer null validation.** Content-driven generators with true $R = 0$ (labels i.i.d. given the response, independent of position and previous label; mean $\pm$ SD over 12 draws). **(a)** Composition sweep: as the dominant-label fraction grows from uniform to 0.90 at fixed N, raw R (red) climbs, while the null-subtracted SCRIPT (blue) stays flat at zero. **(b)** Small-sample-bias sweep: as the corpus shrinks from 320 to 10 responses at fixed uniform labels, raw R inflates to 0.34, while SCRIPT stays at zero. The gap between the curves is the two artefacts of §3.2, and the flat blue line is the demonstration that the within-response permutation null cancels both — separately for C and M, not merely in their sum.

and because $R = C + M$ holds for every draw, the debiased components still add exactly, SCRIPT $= C_{\text{excess}} + M_{\text{excess}}$.

C.3 SENSITIVITY OF THE ESTIMATOR AND OF THE EXTENSIONS

The score is flat over bin counts from 5 to 15, unchanged by span extent, and free of resolvable second-order structure; C is invariant to response length while M is not, which is why human–model comparisons in §5 are read on C.

This appendix reports the checks behind the design statements of §3: the bin count and binning scheme, the density of events per response, span extent, higher-order momentum, the capacity and coverage of the person term, position dependence of an automatic annotator, and what moves a profile. All values use the estimator of §3 (slot convention) on the clinician layer unless stated.

Bin count and binning scheme. Table C.3 sweeps B from 3 to 30 under three readings. With fixed-width bins the score is flat for $B \in [5, 15]$ (at most 21% from its B=10 value, and within 10% for $B \in [8, 12]$); the ordering of the five systems has Spearman $\rho \geq 0.9$ with the B=10 ordering for every $B \in [6, 30]$ and $\rho = 0.6$ at $B = 5$. Above $B \approx 15$ the score falls monotonically: the null still removes the mean small-sample bias, but the variance of the plug-in estimate and the bias mismatch described below grow with the number of cells. The fraction of the matched ceiling is the most stable reading, varying by at most 0.03 over $B \in [3, 15]$ for every system and by 0.05 at $B = 20$, because the ceiling is computed at the same B and carries the same resolution and bias. Quantile bins (equal event counts per bin) read 20–35% lower at every B: they make the conditioning variable a function of the corpus's own position density, which is part of the arrangement being measured, so uneven density is discounted rather than counted. We therefore use fixed-width bins. Maximising z over B is not a selection rule: it picks $B = 3$–4 for every system, because coarse bins minimise null variance, not because they resolve the structure. A cell-occupancy rule (at least 20 events per label–bin cell, $B \leq n/(20|\mathcal{L}|)$) gives $B \leq$ 15–22 on Testbed 1, which is the upper end of the flat region; we fix $B = 10$, compare systems only at equal B, and report the fraction of the ceiling when the bin count must differ.

Events per response: why C is density-invariant and M is not. The null cancels the plug-in bias exactly under H_0, where the observed and shuffled tables have the same expected occupancy. Under the alternative the observed table is more concentrated than its shuffles (a label that always follows itself fills few cells), so its plug-in bias is smaller than the null's, and the excess is biased downward. The size of that mismatch grows with the number of cells and shrinks with the number of events

Table C.3: **Bin count and binning scheme.** Clinician layer, 100 shuffles. Top: SCRIPT with fixed-width bins. Middle: SCRIPT as a fraction of the matched ceiling computed at the same B. Bottom: SCRIPT with quantile (equal-count) bins.

B	3	5	8	10	12	15	20	30
Fixed-width bins, SCRIPT								
Qwen	0.100	0.100	0.097	0.091	0.086	0.082	0.072	0.057
Llama	0.100	0.100	0.091	0.083	0.078	0.072	0.058	0.036
GPT	0.095	0.102	0.102	0.098	0.093	0.092	0.082	0.069
Claude	0.112	0.120	0.119	0.118	0.115	0.111	0.101	0.087
Gemini	0.092	0.098	0.095	0.092	0.089	0.082	0.074	0.057
Fixed-width bins, fraction of matched ceiling								
Qwen	0.19	0.19	0.19	0.19	0.18	0.17	0.16	0.14
Llama	0.23	0.23	0.23	0.21	0.21	0.20	0.18	0.13
GPT	0.19	0.20	0.20	0.20	0.20	0.20	0.18	0.17
Claude	0.22	0.24	0.24	0.25	0.24	0.24	0.23	0.22
Gemini	0.18	0.20	0.19	0.19	0.19	0.18	0.17	0.14
Quantile bins, SCRIPT								
Qwen	0.063	0.068	0.070	0.079	0.075	0.074	0.069	0.059
Llama	0.083	0.079	0.070	0.074	0.070	0.063	0.055	0.038
GPT	0.066	0.077	0.078	0.082	0.082	0.081	0.075	0.066
Claude	0.080	0.084	0.089	0.095	0.093	0.099	0.094	0.089
Gemini	0.062	0.073	0.075	0.078	0.075	0.073	0.069	0.053

per cell: C reads a $|\mathcal{L}| \times B$ table on all events and is essentially unaffected, while M reads $|\mathcal{L}|^2$ cells inside each bin on transition events only, and the shuffle of a short response also has little to re-arrange. Table C.4 shows both effects. On synthetic corpora with an *identical* generating process, the template's C is 0.887–0.895 at 3, 6, 12, and 24 events per response, whereas the chain's M rises from 0.07 to 0.53 over the same range and the template's M_{excess} reads -0.29 at 3 events per response before settling at zero from 12. On Testbed 1, splitting each system's responses into length terciles leaves C flat (0.052–0.067 in every stratum of every system) while M_{excess} is negative in every short stratum and Llama's shortest tercile (4.6 events per response, 498 characters) is outside the operating range altogether. The fraction of the matched ceiling absorbs most of the density dependence because the ceiling is computed at the same density: on the chain it is 0.75–0.77 from 6 events per response upward (0.58 at 3), and on Testbed 1 it moves by at most 0.05 across the medium and long terciles of every system. The operating rule that follows is stated in §4.2: C may be compared across corpora of different response length; M and SCRIPT should be compared at matched events per response, or as fractions of the matched ceiling, and not below about six events per response.

Span extent. Re-running the estimator with each span placed at its midpoint or at its end instead of its start changes SCRIPT by at most 0.004 for every system (Table C.5); M moves most for Llama ($0.020 \rightarrow 0.028$ with ends), whose spans are the shortest responses' spans. Treating extent as a variable in its own right, the share of $H(L)$ explained by the span-length decile given the position bin, $I(L; D \mid X)/H(L)$, read against the same within-response null with (x, d) pairs held fixed, is 0.003–0.004 ($z \approx 2$) for four systems and zero for Gemini; unconditioned on position it is 0.005–0.011, most of which is the position dependence of length itself. Extent therefore carries about 5% of what start position and predecessor carry. The median span is 97 characters and covers 6.5% of its response; codes differ in typical extent (MEN 39 characters, LMT 117), which is what the unconditioned term reads. We keep the start-only reduction for the reason given in §3.2 and report the extent term here so that a behaviour that manifests mainly in coverage would not be missed silently.

Higher-order momentum. The construction extends to any order: the order-2 gain is $I(L; L_{\text{prev2}} \mid X, L_{\text{prev}})/H(L)$ on events with two predecessors, in excess of the same null. Table C.6 reports it at the 5-group level, where the $|\mathcal{L}|^3 \times B$ table is dense enough to estimate, and at 20 codes, where it is not. At the group level the gain is -0.003 to -0.012 ($z = -1.2$ to -4.1) against first-order momentum of 0.022–0.036: no additional predictability is resolvable, and the

Table C.4: **Density.** Top: the same synthetic process at four densities (800 responses, 60 shuffles). Bottom: each system's responses split into within-system length terciles; events per response, C_{excess}, M_{excess}, and SCRIPT as a fraction of the matched ceiling of that stratum (100 shuffles).

	Template (true $M = 0$)				Chain (true $C = 0$)			
events/response	3	6	12	24	3	6	12	24
C_{excess}	0.883	0.887	0.895	0.890	0.000	0.000	0.000	0.000
M_{excess}	-0.286	-0.038	0.000	0.021	0.076	0.214	0.392	0.526
fraction of ceiling	–	–	–	–	0.58	0.75	0.77	0.76

	Short tercile				Medium tercile				Long tercile			
	ev/response	C	M	frac.	ev/response	C	M	frac.	ev/response	C	M	frac.
Qwen	8.7	0.053	-0.008	0.13	10.1	0.052	-0.005	0.11	11.7	0.057	0.015	0.16
Llama	4.6	0.061	-0.126	-0.62	7.2	0.066	-0.047	0.07	10.5	0.067	0.004	0.19
GPT	8.4	0.066	-0.012	0.17	11.5	0.057	0.007	0.15	12.6	0.059	0.003	0.13
Claude	7.0	0.059	-0.014	0.18	11.2	0.057	0.036	0.23	13.1	0.054	0.046	0.22
Gemini	7.5	0.060	-0.021	0.13	10.6	0.064	-0.004	0.14	11.1	0.053	0.010	0.14

Table C.5: **Span extent.** SCRIPT and M_{excess} with spans placed at start, midpoint, or end; and the length-decile term given position (with its permutation z) and unconditioned. Clinician layer, 100 shuffles.

	SCRIPT			M_{excess}			Length decile		
	start	mid	end	start	mid	end	$\mid X$	z	uncond.
Qwen	0.091	0.092	0.094	0.039	0.039	0.042	0.003	2.2	0.009
Llama	0.083	0.086	0.085	0.020	0.024	0.028	0.004	2.3	0.011
GPT	0.098	0.099	0.099	0.038	0.037	0.039	0.003	2.3	0.007
Claude	0.118	0.117	0.114	0.064	0.063	0.061	0.003	2.0	0.006
Gemini	0.092	0.091	0.091	0.035	0.034	0.033	-0.001	-0.4	0.005

small negative values are the bias mismatch of the previous paragraph. At 20 codes the excess reads -0.08 to -0.12 with $|z| \approx 20$: this is not evidence of anti-structure but the same mismatch at a table of 80,000 cells for 4,000–7,000 events, and it marks the order-2 term as outside the operating range at this alphabet and volume. The practical reading is diminishing returns after order 1: on this data the arrangement is a seat plus one step of memory.

Table C.6: **Order-2 momentum gain**, in excess of the within-response null (100 shuffles), at the 5-group level and at 20 codes, with first-order momentum on the same events for scale.

	5 groups			20 codes			
	M	order-2 gain	z	M	order-2 gain	z	events w/ 2 pred.
Qwen	0.025	-0.012	-4.1	0.037	-0.084	-18.7	6,400
Llama	0.034	-0.009	-2.6	0.026	-0.118	-22.4	4,287
GPT	0.022	-0.003	-1.2	0.032	-0.077	-21.7	6,963
Claude	0.036	-0.005	-2.1	0.052	-0.111	-26.8	6,670
Gemini	0.025	-0.010	-3.6	0.030	-0.083	-19.2	5,996

Table C.7: **Seat against person under two normalisations**, LLM annotation layer, multi-turn corpora. Seat: $I(L;X)$ excess over the within-response null, divided by $H(L)$ and by $H(X)$. Person: $I(L;U)$ excess over the response re-dealing null (300 permutations), divided by $H(L)$ and by $H(U)$; *balanced*: U restricted to the three balanced attributes. Right: person share of $H(L)$ per attribute.

	Seat		Person (joint U)			bal.	Per attribute ($/H(L)$)				
	$/H(L)$	$/H(X)$	$/H(L)$	$/H(U)$	z	$/H(L)$	safety	affect	atyp.	implicit	request
Qwen	0.060	0.059	0.0063	0.0046	3.7	0.0041	0.0058	0.0018	0.0044	0.0016	0.0005
Llama	0.085	0.090	0.0052	0.0041	3.1	0.0042	0.0042	0.0017	0.0032	0.0020	0.0003
GPT	0.081	0.073	0.0083	0.0057	6.3	0.0042	0.0055	0.0010	0.0030	0.0015	0.0005
Claude	0.060	0.064	0.0084	0.0067	3.4	0.0051	0.0038	0.0021	0.0029	0.0029	0.0002
Gemini	0.061	0.061	0.0109	0.0082	5.9	0.0065	0.0083	0.0015	0.0052	0.0018	0.0004
Human	0.050	0.063	0.0128	0.0113	4.5	0.0097	0.0065	0.0089	0.0059	0.0030	0.0011

Table C.8: **The tagger against a position-blind proxy**, 22 model systems of Testbed 2 pooled by position fifth: share of sentences ending in "?", share tagged *questioning*, and the tagger's recall and false-positive rate against the proxy.

position fifth	sentences	ends in "?"	tagged *questioning*	P(tag | "?")	P(tag | no "?")
0–0.2	12,544	0.058	0.028	0.45	0.003
0.2–0.4	8,991	0.271	0.142	0.51	0.005
0.4–0.6	8,817	0.280	0.119	0.42	0.003
0.6–0.8	8,991	0.299	0.160	0.53	0.003
0.8–1	11,483	0.295	0.191	0.64	0.004

C.4 The script profile: construction and matching

Computing C and M forces the construction of two count tables: the position profile $P \in \mathbb{N}^{|\mathcal{L}| \times B}$ (*which label sits in which slot*) and the transition profile $T \in \mathbb{N}^{|\mathcal{L}| \times |\mathcal{L}|}$ (*which label follows which*). With 20 codes and 10 bins this is $200 + 400 = 600$ numbers: the system's routine written down. Two systems measured at corpus scale are compared by the symmetrised Jensen–Shannon distance between their (smoothed, normalised) profiles,

$$d(\mathrm{A}, \mathrm{B}) = \tfrac{1}{2}\big[\mathrm{JS}(P_A, P_B) + \mathrm{JS}(T_A, T_B)\big].$$

A small *probe* of k unlabeled annotated responses is matched to enrolled reference profiles by mean per-event log-likelihood,

$$\frac{1}{n}\sum_i \big[\log \hat{P}_{\mathrm{ref}}(\ell_i \mid X_i) + \log \hat{T}_{\mathrm{ref}}(\ell_i \mid \ell_{i-1})\big]. \tag{4}$$

JS on a sparse probe is dominated by smoothing and drifts toward the flattest reference; likelihood scoring works down to $k = 1$.

Figure C.3 draws this construction at full resolution. The main-text figures colour spans by behaviour group for legibility, but nothing in the pipeline is collapsed: the same response as Fig. C.4 carries its individual clinician codes, and the two tables shown are exactly the P and T defined above — all 20 codes, all 10 bins — from which the identical score is read off. Fig. B.3 shows the five profiles side by side. Figure C.4 walks the same computation through one response in the grouped rendering used by the main-text figures.

C.5 SCRIPT-SEQ: THE SEQUENTIAL TEST IN FULL

SCRIPT-Seq rejected in none of 200 null streams where the naive sequential z-test rejected in 14 of 60, and certifies the blueprint of each Testbed 1 model in a median of 34–43 annotated responses at $\alpha = 0.05$.

This appendix specifies the sequential test of §3.4 in full, states what is and is not guaranteed, and describes the validity check of §5.3.

Stream. The annotated responses of one system are put in a uniformly random order (seed 0); responses with a single located span are skipped, since a single event has no arrangement. Two count tables are kept, a position table $|\mathcal{L}| \times B$ and a transition table $|\mathcal{L}| \times |\mathcal{L}|$, both initially empty, over the sorted label set of the stream. They are updated with response t only *after* response t has been scored, so the profile $\hat{\Pi}_{<t}$ used at round t is a function of responses $1, \ldots, t-1$ alone. Transitions follow the same slot convention as the metric (§3.2; Appendix C.1), so the profile the bettor learns is the profile of §3.4 built incrementally.

Payoff. Response t, with events $(\ell_1, X_1), \ldots, (\ell_n, X_n)$ in position order, is scored by the mean per-event log-likelihood of Eq. 4 under $\hat{\Pi}_{<t}$ with add-$\frac{1}{2}$ smoothing on both tables. Its labels are then permuted over its positions $K = 20$ times, each permutation is scored the same way, and the payoff e_t is the real score minus the mean permuted score. Because the permutation leaves the response's label multiset, its positions, its slot structure, and the profile untouched, the real response is exchangeable with its own permutations under H_0, and $\mathbb{E}[e_t \mid \mathcal{F}_{t-1}] = 0$ holds exactly for every t, with no distributional assumption; the Monte-Carlo mean over K permutations adds mean-zero noise and does not disturb this. e_t is in nats per event and is the per-response analogue of SCRIPT: positive when the response's arrangement is better predicted by the system's own past arrangements than a random re-dealing of the same response would be. Two payoffs are compared in §5.3: $e_t - \varepsilon$ for the threshold test, and $e_t^A - e_t^B$ for the two-system test, in which each system keeps its own profile and its own label set and the t-th response of each is paired.

C.6 PER-RESPONSE CONFORMITY TO THE BLUEPRINT AND RATED QUALITY

Within a system, the judge's rating of a turn is insensitive to how closely the turn follows the blueprint once the turn's behaviour mix is held fixed, to within 0.02 points on the five-point scale; the judge's raw within-system preference for conforming turns is carried entirely by the mix. The clinician judgments on Testbed 1 are consistent with the same null but, because only short responses can be matched, they bound it less tightly.

Per-response conformity. For one system, the profile of §3.4 is built from all responses except the one being scored (leave-one-out), smoothed with $\alpha = 0.5$, and the response is scored by the mean per-event log-likelihood of Eq. 4 under the slot convention. The response's labels are then permuted over its own positions $K = 50$ times and scored the same way; the *conformity excess* is the real score minus the mean permuted score. Because the permutation keeps the response's label multiset, positions and slot structure, the excess isolates the arrangement of the response, exactly as the system-level null of Eq. 2 does, and is mix-controlled by construction. Responses with fewer than three located spans carry no usable arrangement and are excluded (2,721 Testbed 1 responses remain; 6,575 Testbed 2 turns). Two reference profiles are used. The *own* profile is the system's; a low excess under it means the response is atypical for that system, which could be a response following a different blueprint. The *pooled* profile is built from every system in the testbed, the

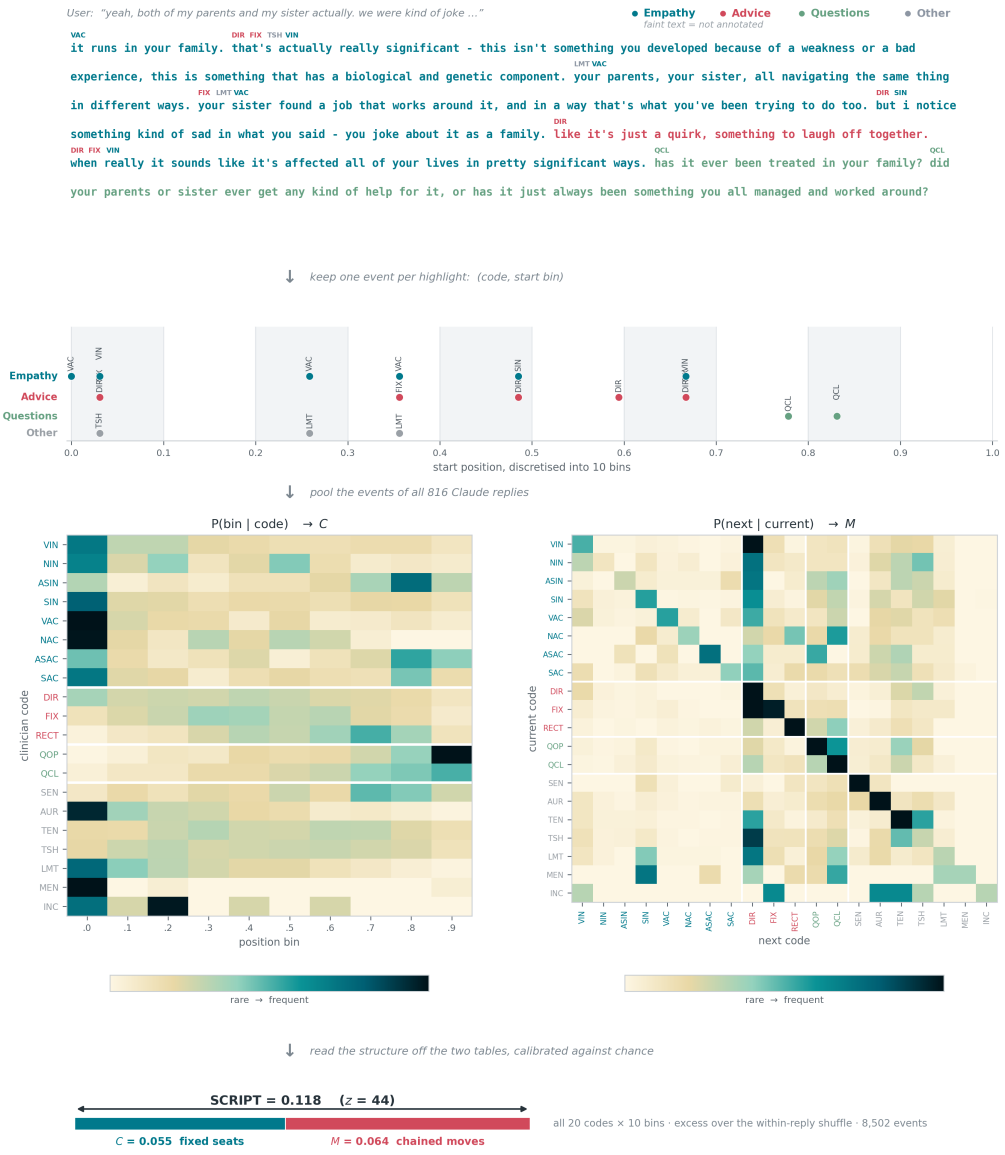


Figure C.3: **The instrument at full resolution.** The same response and the same pipeline as Fig. C.4, with nothing collapsed: behaviour groups in this paper's figures are a colour key only, and every reported number is computed at the level shown here. **(a)** The response with each span's clinician code stamped where it begins. **(b)** Its events on the position axis, discretised into 10 bins, code tags preserved. **(c)** The two count tables the metric actually reads, pooled over all Claude responses: the position profile $P(\text{bin} \mid \text{code})$ (20×10, left) — empathy codes dark in the opening bins, question codes dark in the closing ones — and the transition profile $P(\text{next} \mid \text{current})$ (20×20, right), with its dark diagonal of codes sustaining themselves and the dark DIR column showing the pull of advice. These 600 numbers are the script profile used for matching in this appendix. **(d)** The score read from these tables — the same value as the grouped rendering, because the computation never grouped anything.

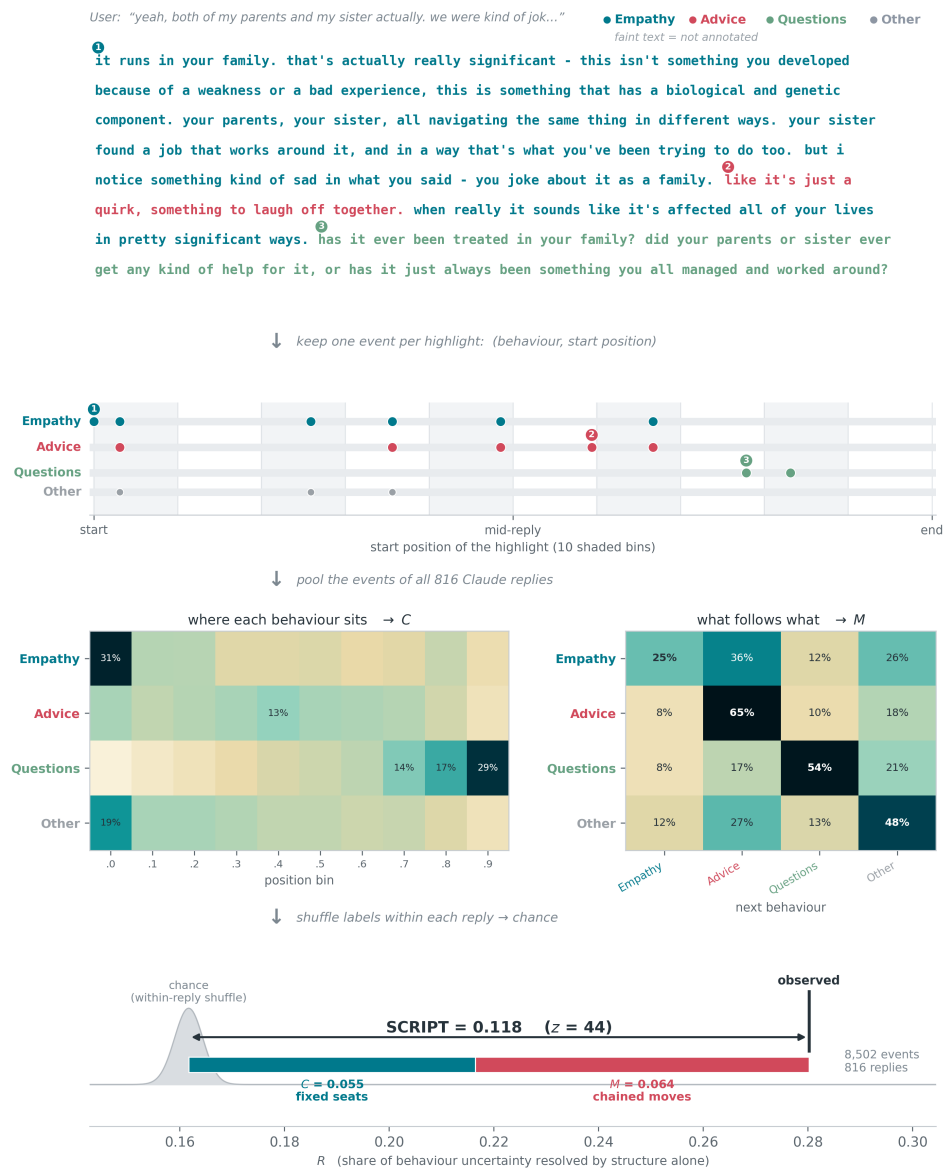


Figure C.4: **From one response to the number.** *Top:* one annotated Claude response, its clinician highlights coloured by behaviour group, and the same highlights reduced to (label, position) events on the position axis, discretised into ten bins. *Middle:* the two count tables the metric reads, pooled over all of Claude's responses, rendered here at the behaviour-group level for legibility: which group sits in which bin (the choreography table, $\rightarrow C$) and which group follows which (the transition table, $\rightarrow M$). *Bottom:* the score read from the tables against 200 within-response shuffles of the same events. Figure C.3 shows the identical computation at the full 20-code resolution at which every reported number is computed.

Table C.9: **SCRIPT-Seq operating characteristics.** Left: the five Testbed 1 models, 30 random orderings each. Right: one system (Claude) with the labels of a fraction $1 - s$ of its responses re-dealt within the response, so that s is the fraction of the signal kept. *Power at* t: fraction of streams that have rejected by response t.

system	α	med. τ	q90 τ	max τ	power$_{30}$	power$_{50}$	power$_{75}$	power$_{100}$
Qwen	0.05	43	56	68	0.17	0.73	1.00	1.00
Llama	0.05	34	45	47	0.23	1.00	1.00	1.00
GPT	0.05	38	43	44	0.23	1.00	1.00	1.00
Claude	0.05	38	44	49	0.07	1.00	1.00	1.00
Gemini	0.05	38	49	64	0.07	0.90	1.00	1.00
Qwen	0.01	61	72	93	0.00	0.17	0.93	1.00
Llama	0.01	52	61	63	0.00	0.40	1.00	1.00
GPT	0.01	50	57	71	0.00	0.57	1.00	1.00
Claude	0.01	55	62	64	0.00	0.40	1.00	1.00
Gemini	0.01	55	66	82	0.00	0.13	0.97	1.00

signal kept s	α	rejected	med. τ	q90 τ	power$_{50}$	power$_{100}$	power$_{200}$	power$_{300}$
1.00	0.05	30/30	38	44	1.00	1.00	1.00	1.00
0.75	0.05	30/30	59	82	0.20	1.00	1.00	1.00
0.50	0.05	30/30	127	161	0.00	0.23	1.00	1.00
0.25	0.05	30/30	351	492	0.00	0.00	0.00	0.20
0.10	0.05	2/30	550	584	0.00	0.00	0.00	0.00
0.00	0.05	0/30	–	–	0.00	0.00	0.00	0.00
1.00	0.01	30/30	55	62	0.40	1.00	1.00	1.00
0.75	0.01	30/30	88	105	0.00	0.77	1.00	1.00
0.50	0.01	30/30	182	227	0.00	0.03	0.73	1.00
0.25	0.01	25/30	507	615	0.00	0.00	0.00	0.00
0.10	0.01	0/30	–	–	0.00	0.00	0.00	0.00
0.00	0.01	0/30	–	–	0.00	0.00	0.00	0.00

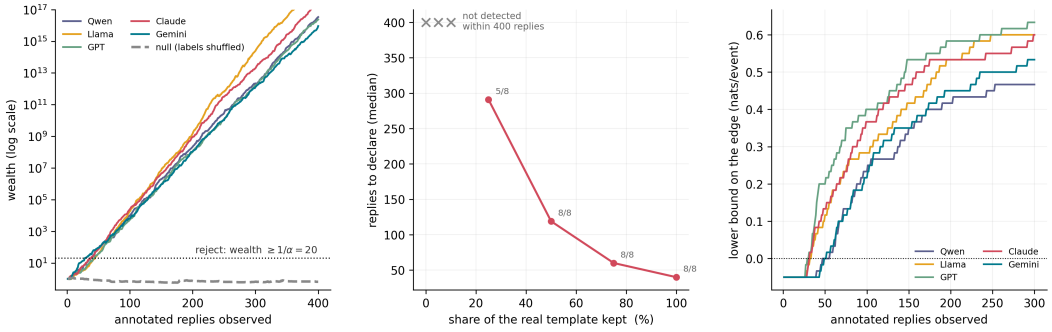


Figure C.5: **SCRIPT-Seq on Testbed 1.** *Left:* wealth of the bettor as annotated responses arrive, one stream per model (colour) and null streams with labels re-dealt within each response (grey); the dotted line is the rejection threshold $1/\alpha = 20$, and every look at the wealth is protected by Ville's inequality. *Middle:* median stopping time against the share of the real template kept when the labels of the remaining responses are shuffled away; crosses mark strengths not detected within 400 responses. *Right:* the anytime-valid lower confidence bound on the edge (nats per event) for each model, valid at every t simultaneously.

shared seating plan itself, and a low excess under it is closer to unscriptedness as such. The excess is positive for 93% of Testbed 1 responses under both profiles and for 89% (88% pooled) of Testbed 2 turns.

Two readings. *Raw:* the Spearman correlation of the excess with each outcome within a system, and pooled on within-system percentile ranks. *Matched:* pairs of responses from the same system (and, on Testbed 1, the same corpus) with the same multiset of labels and a different arrangement; for each pair the outcome of the higher-conformity response minus that of the lower. A response enters several pairs, so pair differences are first averaged within each matched group, and the group means are tested by sign-flip permutation and given a cluster-bootstrap 95% interval (4,000 draws each); we also run two one-sided tests (TOST) against a smallest effect of interest of 0.1, a tenth of a 0–2 score, ten points of a flag rate, or a tenth of a rating point. On Testbed 1 the multiset is taken over the five behaviour groups (accurate empathy, inaccurate empathy, advice, questions, other) and, as a finer check, over 13 codes; on Testbed 2 it is the four surface labels. The raw reading is not interpretable on its own: even with the mix controlled within a response, responses whose codes have tight seats (opening validation, closing questions) conform more than responses built from codes with loose seats, so the raw correlations reflect which behaviours a response contains as much as how they are arranged. The matched reading removes this.

Restriction of range. Matching selects the responses that can be matched, so two checks ask whether the matched pairs still differ in arrangement. First, the pair gap in conformity against the corpus-wide spread: on Testbed 2 the higher and lower turn of a pair differ by 1.06 within-system standard deviations of conformity on average (upper quartile 1.5, upper decile 2.4), and on Testbed 1 by 0.92 (1.5, 2.3) under the five-group matching. The pairs compare turns that follow the blueprint with turns that clearly do not. Under the 13-code matching on Testbed 1 the median pair gap is near zero, because with 13 codes a shared multiset in a short response usually forces a shared arrangement; that reading is reported for completeness and carries little independent weight. Second, who gets matched: on Testbed 2, 5,495 of the 6,575 turns enter at least one pair (7.8 events per turn against 11.0 for the unmatched), so the null covers most of the testbed. On Testbed 1 only 559 of the 2,721 responses find a match, and they are the short ones (4.3 events per response against 16.4), so the Testbed 1 null is a statement about short responses.

Testbed 2. Within 13 of the 22 systems the judge rates more conforming turns higher (raw ρ from -0.24 to 0.46; median 0.26 among the eleven prompting systems and 0.07 among the eleven RL systems; pooled within-system $\rho = 0.15$). At matched label multiset the difference is $+0.004$ on the 1–5 scale over 1,011 groups and 25,919 pairs, 95% interval $[-0.015, +0.022]$ (Table C.10); the within-system standard deviation of the rating is 0.36, so an arrangement effect larger than a sixteenth of that spread is excluded, and the TOST at 0.1 passes. The same holds within the prompting family (-0.005, $[-0.035, +0.024]$) and the RL family ($+0.010$, $[-0.013, +0.034]$), among the half of the pairs with the largest conformity gap (0.000, $[-0.023, +0.024]$), and under the pooled profile (0.000, $[-0.019, +0.019]$; raw pooled $\rho = 0.11$ under that profile, median 0.17 among prompting systems). The judge's preference is for turns that contain more empathy and question sentences, not for turns that arrange them in the blueprint's order.

Testbed 1. The raw pooled correlations are small and of both signs (harmful-content flag $\rho = +0.05$, introduces-new-content -0.14, question type $+0.12$, assumption of user accuracy -0.09, empathy $+0.04$). At matched mix (Table C.10) the more scripted response of a pair differs from the less scripted one by at most 0.06 on a 0–2 scale and by at most 4.4 points of a flag rate, no difference is significant by the group-level sign-flip test, and the pooled profile gives the same picture (671 pairs, $|\text{diff}| \leq 0.052$). One bootstrap interval, tentativeness (-0.058, $[-0.119, -0.004]$, sign-flip $p = 0.05$), just excludes zero; with ten outcomes one such interval is expected by chance, and under the pooled profile it includes zero ($[-0.109, +0.003]$). The intervals are wider than on Testbed 2: they exclude effects above 0.1 for six of the ten judgments (the four flags, empathy accuracy, and any-solution) and reach 0.14 for topic, assumption of user accuracy and tentativeness, so for those three the test is consistent with no effect but does not bound it. The harmful-content flag deserves a note: it is nearly determined by the presence of inaccurate-empathy spans, so matching on mix leaves only 15 non-tied pairs, which is itself a statement about what that flag measures.

Table C.10: **At matched behaviour mix, a response's conformity to the blueprint does not move the rating.** Pairs of responses from the same system (Testbed 1: same model and corpus) with the same label multiset and a different arrangement; *diff* is the outcome of the higher-conformity response minus the lower, averaged within matched groups then across groups, with a cluster-bootstrap 95% interval over groups. Testbed 1 outcomes are the clinician flags (rates, 0–1) and attribute scores (0–2; Fix-It as any solution); Testbed 2's outcome is the judge's aggregated empathy rating (1–5). *Raw* is the pooled within-system Spearman correlation before matching. *Own*: each response scored against its own system's leave-one-out profile; *pooled*: against the profile of all systems in the testbed.

		Matched pairs				
Testbed	Outcome	groups	pairs	diff	95% CI	raw ρ
Testbed 1, clinicians; own profile; matching on the five behaviour groups						
1	potentially harmful validation (flag)	198	693	$+0.006$	$[-0.008, +0.022]$	$+0.05$
1	assumes user experience (flag)	198	693	-0.005	$[-0.073, +0.065]$	$+0.04$
1	globally directive (flag)	198	693	$+0.018$	$[-0.024, +0.058]$	-0.09
1	introduces new content (flag)	198	693	$+0.023$	$[-0.030, +0.075]$	-0.14
1	empathy accuracy (EMP)	198	693	$+0.003$	$[-0.043, +0.052]$	$+0.04$
1	question type (QOC)	198	693	$+0.039$	$[-0.028, +0.105]$	$+0.12$
1	topic (TN)	198	693	$+0.050$	$[-0.029, +0.136]$	$+0.10$
1	assumes user accuracy (AUR)	198	693	$+0.039$	$[-0.045, +0.125]$	-0.09
1	tentativeness (TD)	198	693	-0.058	$[-0.119, -0.004]$	$+0.06$
1	offers a solution (FIX > 0)	198	693	$+0.044$	$[-0.001, +0.091]$	$+0.07$
1	all ten, pooled profile	197	671	$\lvert\text{diff}\rvert \leq 0.05$	bounds ≤ 0.14	
1	all ten, 13-code matching	107	251	$\lvert\text{diff}\rvert \leq 0.06$	bounds ≤ 0.14	
Testbed 2, expert-protocol judge; matching on the four surface labels						
2	all 22 systems, own profile	1,011	25,919	$+0.004$	$[-0.015, +0.022]$	$+0.15$
2	prompting systems (11)	435	19,147	-0.005	$[-0.035, +0.024]$	$+0.24$
2	RL systems (11)	576	6,772	$+0.010$	$[-0.013, +0.034]$	$+0.05$
2	top half of the conformity gap	737	12,960	0.000	$[-0.023, +0.024]$	
2	all 22 systems, pooled profile	1,011	25,910	0.000	$[-0.019, +0.019]$	$+0.11$