Generating Clinical Vignettes that Preserve Cognitive Formulations

Amit Oren[1] **Nimrod Hertz-Palmor**[2] **Dean Ariel**[3,4] **Guy Laban**[1,5,6*]

[1]Department of Industrial Engineering and Management,
Ben-Gurion University of the Negev, Beer Sheva, Israel
[2]MRC Cognition and Brain Sciences Unit, University of Cambridge, United Kingdom
[3]Clalit Health Services, Israel [4]School of Public Health, Tel Aviv University, Israel
[5]School of Brain Sciences and Cognition,
Ben-Gurion University of the Negev, Beer Sheva, Israel
[6]Azrieli National Center for Autism and Neurodevelopment Research, Beer Sheva, Israel

Abstract

Large language models can generate fluent clinical case vignettes, but fluency alone does not ensure fidelity to a specifiable clinical structure. We introduce FORMA, a theory-grounded framework that compiles a cognitive model of a disorder into a directed weighted graph, samples a person-specific configuration of that graph, and validates whether the generated vignette preserves the specified components and causal links. We instantiate FORMA on Posttraumatic Stress Disorder using the Ehlers and Clark cognitive model, generating 16,500 vignettes across 500 personas, 11 generation models, and three ablation conditions. Evaluation combines an external edge-recovery probe, two clinical experts, a scaled LLM judge, and a clinician user study with 100 licensed practitioners. The cognitive graph is recoverable from full-condition vignettes (MCC $= +0.41$, AUC $= 0.70$) but not from zero-shot generation (MCC $= +0.01$, AUC $= 0.50$). Experts rate full vignettes substantially higher than zero-shot alternatives, and clinicians perceive them to be human-written 85% of the time, compared with 22% for zero-shot. FORMA also reduces demographic disparity in perceived quality by 1.5–7×. These results show that cognitive formulation can serve as an auditable specification for scalable synthetic clinical text generation. A repository with the data and code is available online: https://github.com/Amit-Oren/FORMA.

1 Introduction

Clinical case vignettes, short narrative descriptions of a patient's presentation, are a workhorse of mental-health training, research, and assessment (Evans et al., 2015; Sheringham et al., 2021), used to teach diagnostic reasoning (Rosenquist et al., 2000; St. Marie et al., 2021), standardise experimental stimuli, and benchmark the clinical capabilities of language models (Benoit, 2023; Yanagita et al., 2024). Across all three uses the same constraint binds: vignettes must look like real clinical material, in numbers and along axes of variation that a hand-curated bank cannot deliver. Curated banks are small and skewed toward prototypical cases (Evans et al., 2015; Heverly et al., 1984); real clinical records resist redistribution at scale and rarely come annotated with the clinical structure researchers want to vary. Synthetic generation with large language models (LLMs) is the natural alternative (Frayling et al., 2024; Reichenpfader and Denecke, 2024; Bhate et al., 2023), but a generated vignette is only useful insofar as it is *faithful* to a specifiable clinical picture rather than fluent in the average. Given a diagnostic label, a competent LLM produces text that reads as plausible while averaging over the population of patients with the diagnosis and clustering around trauma-type templates rather than the individual processes that maintain the disorder.

We address this by grounding generation in a *cognitive formulation* of the disorder: a standard tool in cognitive-behavioural therapy for case conceptualisation (Persons, 2008) specifying a disorder by a directed, person-specific network of psychological components whose interactions explain how the disorder is maintained in this patient, rather than merely using a symptom checklist. They are mechanistic, individualised, and clinically standard, and formal cognitive models with this structure exist for many disorders including depression (Beck et al., 1979), OCD (Salkovskis, 1985), social anxiety (Clark and Wells, 1995), and Posttraumatic Stress Disorder (PTSD) (Ehlers and Clark, 2000), with recent work showing these models can be empirically recovered from patient data (Hertz-Palmor et al., 2025). Concretely, we compile the cognitive model of a disorder into a directed, weighted graph over its components, sample a concrete graph configuration for each persona, and use the configuration as a structural specification that the generating LLM must satisfy. A closed-loop validator then verifies that each specified component and directed causal link is realised in the text, inserting targeted patches where it is not. The resulting framework,

*Corresponding author: laban@bgu.ac.il

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29423–29472
October 24-29, 2026 ©2026 Association for Computational Linguistics

The trigger could also bring back the event whole and without warning, so that a smell of heat or the sight of flames was enough for her to hear the shouting in her head again as if she were back there.

[...]

Because she was certain the world was dangerous and because she felt flawed for reacting so strongly, she tried to avoid anything that might be stressful. She also checked her surroundings constantly, in the office, on base, in parking lots, noticing exits, listening for changes in sound, looking for anything that might signal threat.

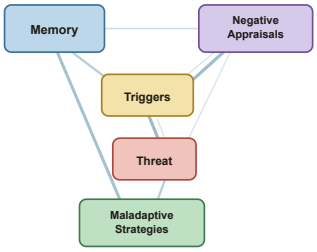


Figure 1: Excerpt from a full-condition vignette (Persona 14, `gpt-5.4`) and the persona's specified cognitive graph. Highlighted spans mark clinician-identified realizations of each Ehlers and Clark component. The complete, unabridged vignette is given in Appendix I.

FORMA[1], is disorder-agnostic.

We demonstrate it on PTSD, which has a well-validated cognitive theory (Ehlers and Clark, 2000) operationalising the disorder as a graph over five components (traumatic memory, negative appraisals, current threat, re-experiencing triggers, maladaptive strategies) linked by directed causal edges. PTSD's symptom heterogeneity is enormous: the DSM-5 criteria admit over 600,000 distinct profiles (Galatzer-Levy and Bryant, 2013), making it a stress test for whether a generation framework can produce *individuated* rather than *averaged* cases. Figure 1 shows an excerpt beside the graph it was generated from, clinician-marked by component; the full case is in Appendix I.

This paper makes three contributions.

1. **A theory-grounded generation framework** (FORMA) that compiles a cognitive model into a directed weighted graph, samples configurations as personas, and uses a closed-loop validator to enforce structural fidelity (§3), together with a three-condition ablation design (full / no-formulation / zero-shot) that isolates the marginal contribution of the graph from the persona's self-report items and the demographic specification (§3.4).

2. **A multi-tier evaluation protocol** triangulating structural validity (judge versus specified graph, corroborated by two clinical raters annotating all 20 directed edges on 150 vignettes), geometric structure of the generated set (cross-model convergence and trauma-type stereotyping in embedding space), expert validity (two clinical experts), scale (an LLM judge on 16,500 vignettes from 11 models), ecological validity (a clinician user study with N=100 licensed practitioners), and demographic fairness (§4.2–5.7).

3. **An empirical case that the cognitive graph is a methodological intervention, not a stylistic input variable.** Across the evaluation suite, full-formulation vignettes: (i) preserve the specified directed edges (MCC $+0.41$ vs $+0.01$ in zero-shot; AUC of edge weight as a predictor of edge presence 0.70 vs 0.50); (ii) in embedding space, make different generation models converge on a shared portrayal of the same persona ($d_z = 1.52$) while breaking trauma-type stereotyping; (iii) are rated higher by clinical experts (Cohen's d_z up to $+1.98$) and pass as human-written to clinicians 85% of the time (vs 22% in zero-shot); (iv) elicit monotonically higher inter-rater agreement. Effects replicate across 11 models from 5 families (§5).

The released dataset[2] comprises 16,500 vignettes (500 personas $\times$ 11 models $\times$ 3 conditions), a 330-vignette evaluation subsample with 2 expert and 3 LLM-judge ratings, and 1,600 ratings from the clinician user study. Full release details in §7.

2 Related Work

Case vignettes have been used in clinical training, research, and assessment for decades (Heverly et al., 1984; Evans et al., 2015; Sheringham et al., 2021), and Mohan et al. (2014) document how unvalidated vignettes distort clinical reasoning when used in training. Recent LLM-based generation efforts produce fluent but structurally shallow output: Benoit (2023) and Yanagita et al. (2024) use zero-shot chat prompts on diagnostic labels; Bhate et al. (2023) and Frayling et al. (2024) report similar surface-fluency-without-faithfulness on clinical notes; Reichenpfader and Denecke (2024) use persona conditioning for demographic diversity; Suhas et al. (2025) release a large-scale dataset of Prolonged-Exposure therapy conversations from

[1]From Latin *forma* ("shape, structure"): the cognitive graph supplies the form, the LLM the realisation.

[2]Data + code: `https://github.com/Amit-Oren/FORMA`; interactive data browser: `https://clinical-vignettes.streamlit.app/`.

synthetic clients. None of this work uses a formal cognitive model of the disorder as the generation specification or evaluates whether the generated text preserves directed causal structure.

The theoretical move that would close this gap is already standard in clinical psychology. Case formulation represents a disorder not as a symptom checklist but as a person-specific directed network of psychological components whose interactions explain how the disorder is maintained (Persons, 2008; Beck et al., 1979; Salkovskis, 1985; Clark and Wells, 1995; Ehlers and Clark, 2000). Generating text conditioned on a formal structure is a recognised problem family with established graph-to-text (Koncel-Kedziorski et al., 2019) and knowledge-grounded generation (Yu et al., 2022) machinery, but it has not been applied to clinical case generation where the "*knowledge*" is a mechanistic cognitive model of the disorder. Recent NLP work has begun to bring theoretic structure into LLM generation of mental-health text but stops short of a formal graph. Wang et al. (2024) simulate patients for CBT training by conditioning generation on a free-form cognitive conceptualisation diagram; Louie et al. (2024) elicit domain-expert principles as the scaffold for client simulation; Sharma et al. (2023) use LLMs to rewrite negative thoughts via cognitive-reframing operators, and Laban et al. (2026) used an LLM-based architecture for cognitive reappraisal in on-going interventions. More specifically related, Tu et al. (2025) built an LLM dialogue system for structured PTSD diagnostic assessment; Wang et al. (2025) simulate counselling seekers with an emotion-modulator and multi-session memory; and Wu et al. (2026) ground counselling-dialogue generation in multiple psychological theories simultaneously. The structure these systems use is either a free-form natural-language summary, a single-operator transformation, or a dialogue-act schema for assessment rather than for content generation and there is no formal directed graph of the disorder's components that the generated text can be audited against, edge by edge. We test this trajectory by compiling a cognitive model into a directed weighted graph, sample concrete graph configurations as personas, and audit the generated text against the specified edges with an external structural probe (§5.1), treating the cognitive formulation as an auditable computational specification rather than a free-form text scaffold.

3 Methodology

Following Hertz-Palmor et al. (2025), we operationalise the Ehlers and Clark (2000) cognitive model of PTSD as a directed graph over five components: trauma memory, re-experiencing triggers, sense of current threat, negative appraisals, and maladaptive cognitive/behavioural strategies (definitions and operationalisation in Appendix A). For each persona we sample a graph configuration and matched self-report items from a 403-item clinical pool derived from real-patient disclosures to the five EC components in a clinical sample (Hertz-Palmor et al., 2026); these were parameterised into per-component item distributions, and each persona's items are drawn at random from these distributions and then verified by the Vignette Validator (§3.3). Full pool and sampling procedure in Appendix B (pool released in the repository).

3.1 Persona Construction

Each persona is built in three sampled-and-validated stages (Figure 2). *Demographics* (age, gender, ethnicity, occupation, trauma type, PCL-5 severity, relationship status; schema following U.S. Census Bureau, 2025) are sampled from category-uniform distributions (relationship status conditioned on age) and checked by a *Persona Validator Agent* for clinical plausibility (e.g., a 19-year-old is unlikely to be widowed); failed fields are resampled, up to six attempts. *Cognitive model.* Each of the five Ehlers & Clark (EC) nodes is independently activated with probability $p_{\text{node}} = 0.7$ and each directed edge between active nodes with $p_{\text{edge}} = 0.7$; active edges receive a continuous weight from $\text{Uniform}(0.01, 1.0)$, inactive edges weight zero. *Self-report items.* For each active component we sample three items from the clinical item pool. A *Persona Validator Agent* checks internal coherence, trauma-type consistency, and cross-node contradictions; failed items are replaced by a *Persona Crafter Agent*, again up to six attempts. Both persona-level agents use `g-5.4`; the same 500 personas are reused across every (model, condition) cell in §4.1.

3.2 Vignette Generation

The *Vignette Crafter Agent* receives the persona's demographics, the sampled self-report items, and the cognitive graph (active components and weighted causal links). Its task is to write a 500–700 word third-person clinical narrative reflecting

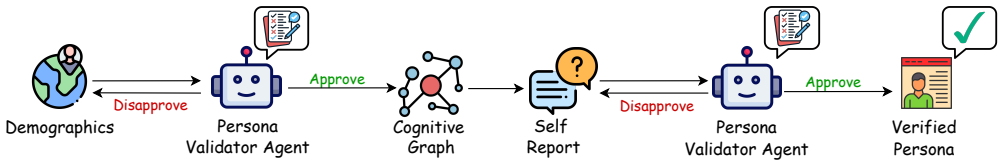


Figure 2: Persona construction. Demographics and a cognitive formulation graph are sampled independently; three self-report items are drawn per active component. A Persona Validator checks plausibility; failures are routed to a Persona Crafter for targeted repair.

the persona faithfully. Edge weights map to four narrative-prominence levels: > 0.5 explicit, 0.1–0.5 implicit, < 0.1 passing, zero-weight components if they appear at all are described in isolation with no causal language.

3.3 Vignette Validation

The Vignette Validator and Crafter share an LLM and form a closed feedback loop (Figure 3). The Validator checks each active component and each weight-positive edge $A \rightarrow B$: an edge is satisfied if some paragraph mentions both endpoints (or synonyms) and conveys that A influences B. For each unsatisfied edge the Crafter writes a targeted *patch* (anchor sentence copied verbatim plus a new sentence inserted after it) instead of rewriting; up to five retries. Most vignettes pass on the first attempt (Appendix C); the sole exception is `gpt-4o-mini`, whose runs retain a median of 4 unsatisfied edges because it consistently fails to emit valid patch-block markers (a markup-compliance issue, not content), so we exclude it from per-condition scaled-judge and GT-recovery analyses but retain it for rating-schema-structure analyses.

3.4 Ablations

We run every persona under three conditions to isolate each layer of structure. *Full*: demographics + self-report items + cognitive graph, structural validation on active edges. *No-formulation*: demographics + self-report items only, no graph; validator checks consistency with self-report. *Zero-shot*: demographics only, no validation.

4 Experimental Setup

4.1 Persona and Vignette Set

We generated 500 personas spanning ages 18–80 ($M = 48.5$, $SD = 18.2$), approximately balanced gender, ethnicity uniform across eight world regions, and 29 trauma categories (interpersonal violence, childhood trauma, war and collective violence, accidental and life-threatening events, loss/medical). PCL-5 severity ($M = 57.8$, $SD = 13.9$) placed every persona above the clinical threshold for probable PTSD. Chi-square tests confirm that demographic and clinical attributes are mutually independent in the realised set (Appendix D); full distributional breakdowns are in Appendix E. Each persona is generated under the three conditions of §3.4 with eleven generation models from five families (Anthropic, DeepSeek, Google, OpenAI, Alibaba; full model list and per-model token counts in Appendix F), yielding 16,500 vignettes. Tokens per vignette range from $\sim$ 900 (small open-weight, zero-shot) to $\sim$ 13,000 (`gemini-2.5-pro`, full). For inter-judge analysis we drew a balanced subsample of 330 vignettes (10 personas $\times$ 11 models $\times$ 3 conditions), stratified by trauma category, gender, age, and PCL-5; the same 10 personas appear under every (model, condition) cell to hold persona-level variance constant.

4.2 Rating Schema, Raters, and Metrics

Every vignette is scored on 18 items in four groups with distinct roles: three CVI items (clarity, relevance, importance; 1–3; St. Marie et al. 2021) for perceived quality, four construction guidelines (grounding, narrativity, diagnostic salience, clinical economy; 1–3; Evans et al. 2015) for narrative quality, six DSM-5 PTSD criteria (American Psychiatric Association, 2013) (binary) as a diagnostic-floor sanity check, and five EC components (binary) as the locus of the formulation effect. Five raters score the 330-vignette subsample: two clinical experts (R1, R2; both MD general-practice residents trained to read clinical cognitive formulations, blind to model and condition) and three LLM judges (`claude-opus-4-5`, `deepseek-v4-pro`, `deepseek-v4-flash`); Flash additionally scores the full 16,500-vignette set. R1 and R2 first calibrated against six training vignettes (not part of the evaluation subsample), discussed any disagreements with the authors until convergence on the rating standard, then completed the 330-vignette scoring blind over approximately four

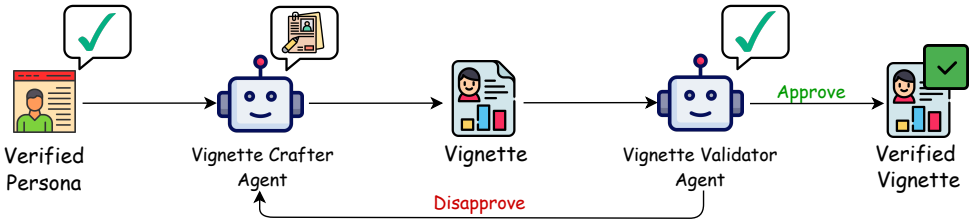


Figure 3: Vignette generation and validation loop. For each unsatisfied edge, the Validator passes violations to the Crafter, which retries up to five times.

days each. We report Gwet's AC1 (binary) and AC2 with quadratic weights (ordinal) for inter-rater agreement (six items have base rate ≥ 0.85 in at least one condition. Condition contrasts use omnibus RM-ANOVA (ordinal, $df = (2, 218)$) or Cochran's Q (Cochran, 1950) (binary), then Bonferroni-corrected paired t-tests / proportion contrasts with paired Cohen's d_z. For ground-truth recovery (rater's binary EC labels vs the persona's specified `active_nodes`, §5.1) we report precision, recall, F1, balanced accuracy, MCC, and Gwet's AC1. Generation and judging hyperparameters, total compute, and the full system prompts for the Crafter, Validator, Edge Probe, and Judge agents are in Appendices G and H; Appendix I reproduces a representative full-condition vignette with its specified graph (Figure 6) and an expert annotation of each EC-relevant span.

5 Results

5.1 Structural Fidelity

We test whether the cognitive graph leaves a structural imprint on the text, starting from the strongest gold (two clinical experts) and scaling out.

Component recovery. The two experts (R1, R2) score each persona's vignette for the presence of each EC component on the 330-subsample. Treating the persona's specified `active_nodes` as gold, in full all five components recover above chance (R1: AC1 0.64–0.86; R2: 0.66–0.85), and recovery collapses to chance under zero-shot (e.g., R1 on *Threat* -0.52 AC1; full per-EC per-rater in Appendix J). The three LLM judges show the same direction and rank-order across components (Appendix J). An expert qualitative annotation of a representative full-condition vignette, in which a licensed clinical psychologist tags each EC-relevant span span-by-span, is in Appendix I. Restricted to the 15,000-vignette generation set, the scaled judge (`deepseek-v4-flash`) reproduces the pattern at scale: all five EC components recover above chance in full (AC1 0.51–0.75); the gap between full and zero-shot is the size of the formulation's contribution (Appraisals MCC .59 $\rightarrow$ .01, Strategies .41 $\rightarrow -.01$; Appendix K).

Edge recovery. We then ask whether the specified *directed edges* between components also appear. An external probe (the scaled judge, blind to the persona's active set and condition) reads every vignette in the 330-subsample and returns a binary judgement on each of the twenty directed EC pairs $A \rightarrow B$, yielding 6,600 paired cells. Table 1 reports the metrics: in full, the judge recovers specified edges above chance (MCC $= +0.41$, AUC $= 0.70$); in no-formulation, recovery is at chance (MCC $= +0.03$); in zero-shot, at chance (MCC $= +0.01$). The judge's ability to read the specified graph off the text is created by the formulation.

The probe cannot read back an input it never receives, being given only the vignette prose, never the graph, the self-report items, or the condition (§4.2). The stronger control is no-formulation, which passes through the identical generation-and-validation loop with the same self-report content and differs from full only in whether the graph is supplied, yet recovers at chance. Recovery therefore tracks the specification rather than the pipeline or the validator loop, whose own limits we note in the Limitations.

In full, specified-active edges are judged present 75% of the time vs 35% for specified-absent (40.8-pp gap; +3.0/+0.5 pp in no-form/zero-shot). Within the active set recovery is flat across weight bands (0.74/0.75/0.84), so weight matters as a presence cue more than as a prominence dial.

Every directed edge type recovers above chance under the formulation. All 20 directed EC edges have positive MCC against the specified gold in full ($+0.25$ to $+0.66$); the strongest are the directed pairs the cognitive model itself emphasises as maintaining the disorder (Maladaptive Strategies $\rightarrow$ Threat: MCC $+0.66$; Triggers $\rightarrow$ Negative

Table 1: External edge-recovery probe (`deepseek-v4-flash`) on the 330-vignette evaluation subsample. Top block: classification metrics treating "specified active" (weight > 0) as the gold label and the judge's per-edge binary as the prediction. Bottom block: P(judge says edge present) by the persona's edge-weight band. The 40.8-percentage-point gap between specified-active and specified-absent rates in the full condition (vs 0.5 pp in zero-shot) is the signature of the formulation doing structural work. Each column aggregates 20 directed edges across 110 vignettes ($n = 2{,}200$).

	Full	**No-form.**	**Zero-shot**
Classification metrics (active vs absent)			
Precision	**0.67**	0.51	0.49
Recall	**0.75**	0.34	0.30
F1	**0.71**	0.41	0.37
MCC	**+0.41**	+0.03	+0.01
AC1	**+0.41**	+0.07	+0.06
AUC	**0.70**	0.52	0.50
P(judge says edge present), by weight band			
high (> 0.5)	0.74	0.35	0.30
med (0.1–0.5)	0.75	0.30	0.27
low (< 0.1)	0.84	0.47	0.44
zero ($= 0$)	**0.35**	0.31	0.30

Appraisals: +0.63; Threat → Memory: +0.57; Threat → Negative Appraisals: +0.56), and the weakest are edges with high judge-marginal rates such as Triggers → Threat (recall 1.00, precision 0.47; the judge tends to read this connection from any narrative that mentions both triggers and threat). The full per-edge table is in Appendix L. Figure 4 draws the same pattern as the EC graph once per condition, each edge weighted and coloured by the judge's recovery probability. The structure is clearly visible in the full panel, while no-formulation and zero-shot are uniformly washed-out, reflecting near-flat recovery across the graph.

5.2 Human Validation of Directed-Edge Recovery

The edge-recovery result above is produced by an LLM probe, and the two experts validate EC *components* rather than the directed links themselves. We therefore replicated the edge analysis under human annotation at comparable scale. Two clinical raters independently annotated all 20 directed EC edges on a 150-vignette sample (50 personas × 3 conditions, balanced across the ten retained generation models), blind to condition and to the persona's specified graph, yielding 3,000 directed-edge judgements per rater. Each edge was scored both for presence and on a graded 0–1 prominence scale. The identical schema was applied to the same 150 vignettes by two of the paper's own LLM judges (`deepseek-v4-flash`, `deepseek-v4-pro`), giving a four-rater panel of two humans and two models. Protocol and full pairwise statistics are in Appendix M.

The probe is indistinguishable from an additional human annotator. On the presence scale every rater pair agrees at Gwet's AC1 between 0.63 and 0.75, and human-to-LLM agreement (mean AC1 = 0.67) is at least as high as agreement between the two humans themselves (AC1 = 0.63). The graded scale behaves the same way (human-to-LLM ICC 0.67–0.72; human-to-human ICC = 0.67). Agreement with the probe is therefore not lower than the agreement two clinicians reach with each other.

Humans recover the specified graph. All four raters independently reproduce the recovery gradient reported above (Table 2). In full, MCC against the specified graph ranges from +0.32 to +0.41 and the AUC of the specified weight predicting detection from 0.65 to 0.69, against approximately zero MCC and approximately 0.50 AUC in both no-formulation and zero-shot. The graded strength ratings carry the same formulation-dependent signal, correlating with the specified weight under full (Spearman 0.34–0.43 across raters) and at chance in both ablations, so the weights leave a trace that appears only when the graph is supplied. Because the human raters apply the causal standard by hand, without sight of the graph, the recovery reported in §5.1 is a property of the generated text rather than an artefact of the probe and the in-loop validator sharing a rubric.

5.3 Vignettes' Geometric Structure

Two embedding-space analyses test whether the formulation acts as a shared structural specification across models. To match the structural and clinical-content analyses (§5.1, §5.7), the geometric analysis uses the 10 generation models that successfully validate; `gpt-4o-mini` is excluded for the validator-loop pathology documented in Appendix C. To guard against an encoder-specific artefact, we re-embed these 15,000 vignettes with two sentence encoders trained on different objectives and corpora: MPNet (Song et al., 2020; Reimers and Gurevych, 2019) (`all-mpnet-base-v2`), used as the primary encoder in the body, and BGE-base (Xiao et al., 2024) (`BAAI/bge-base-en-v1.5`) used for replication.

Cross-model convergence. The within-persona cross-model cosine similarity (MPNet) is highest

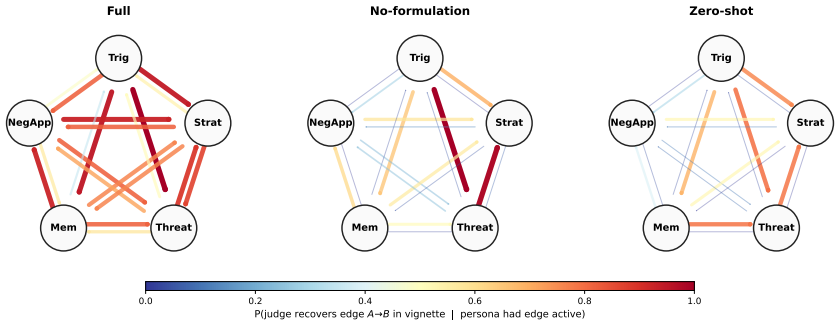


Figure 4: Edge-level recovery of the specified cognitive graph by condition ($n = 110$ vignettes per panel, 1,980–2,300 directed-edge judgements aggregated). Each panel shows the five-node EC graph (Trig=Triggers, NegApp=Negative Appraisals, Mem=Memory, Threat=Sense of threat, Strat=Maladaptive Strategies). Each directed edge $A \to B$ is coloured and thickness-weighted by $P(\textit{judge marks } A \to B \textit{ present} \mid \textit{persona had this edge active})$. Edge structure is clearly visible only under Full; no-formulation and zero-shot panels show uniform near-flat recovery, indicating that without the graph in the input the text contains no signal about which edges the persona was constructed with.

Table 2: Human validation of directed-edge recovery on the 150-vignette sample (50 personas × 3 conditions; 20 directed edges; 3,000 judgements per rater). H1 and H2 are the two clinical raters; FLASH and PRO are `deepseek-v4-flash` and `deepseek-v4-pro` scoring the identical schema, blind to condition and to the specified graph. All four raters recover the specified graph under full and sit at chance under both ablations, reproducing Table 1 under blind human annotation.

	Full	**No-form.**	**Zero-shot**
MCC vs specified graph			
H1	**+0.41**	+0.06	+0.04
H2	**+0.32**	+0.04	+0.01
FLASH	**+0.32**	+0.04	−0.02
PRO	**+0.38**	−0.01	−0.02
AUC (specified weight → detection)			
H1	**0.69**	0.54	0.53
H2	**0.65**	0.53	0.51
FLASH	**0.65**	0.52	0.49
PRO	**0.67**	0.50	0.49
Spearman (rated strength, specified weight)			
H1	**+0.43**	+0.08	+0.05
H2	**+0.36**	+0.06	+0.02
FLASH	**+0.34**	+0.03	−0.01
PRO	**+0.39**	+0.01	−0.02

in full ($\mu = 0.792$), intermediate in no-formulation (0.767), lowest in zero-shot (0.737); all three pairwise differences are significant (paired t, $n = 500$, $p < 10^{-48}$), full-vs-zero-shot is the largest contrast ($d_z = 1.52$). BGE preserves the ordering and the effect size (μ: 0.837/0.816/0.800; full-vs-zero-shot $d_z = 1.61$; Appendix N.3, Figure 10). The two encoders disagree only on absolute level (BGE places all vignettes closer in cosine space than MPNet), but the conditional contrast that the claim rests on is encoder-invariant. The formulation acts as a cross-model anchor: with the graph in the input, the ten retained generation models converge on a shared portrayal under either encoder.

Trauma-type stereotyping. A complementary t-SNE analysis (Appendix O) shows that under zero-shot, vignettes of the same trauma type cluster tightly in latent space: the generator writes from a trauma-type template rather than from the persona's individual content; the same coloured points are dispersed under full and no-formulation. Appendix N quantifies this shift (cognitive-signature/trauma-category cohesion ratio rises from 0.07 in zero-shot to 0.48 in full, a $\sim 7\times$ shift) and replicates the t-SNE pattern under BGE and across model capability tiers. Absolute cluster cohesion is lower under BGE than under MPNet, but the conditional gradient and the tier × condition interaction reproduce, ruling out an encoder-specific artefact.

5.4 Inter-Rater Agreement

Five raters score the 330-vignette subsample using the schema in §4.2: two clinical experts (R1, R2) and three LLM judges (Opus, DSpro, Flash). Table 3 reports inter-rater agreement (R1–R2, mean LLM–LLM, multi-rater All 5) per condition. Three patterns repeat across metric families: LLM–LLM agreement is uniformly high (median ≥ 0.93) in every condition; R1–R2 is near-perfect on CVI, Construction, and DSM items but only strong on the EC items in the full condition (0.82–0.95), dropping to moderate in no-formulation (Threat 0.43, Strategies 0.54) and zero-shot (Strategies 0.36); and the multi-rater All-5 collapses on the EC items in zero-shot (Threat 0.10) even though LLM–LLM remains near-ceiling there, indicating that the loss of agreement is driven by human–LLM divergence on EC content (the six cross pairs and full 10-pair matrix in Appendix P).

Table 3: Inter-rater agreement on the 330-vignette evaluation subsample, per condition ($n = 110$ each), aggregated to three cluster summaries. AC1 (Gwet's) for binary; AC2 (quadratic weights) for ordinal. R1–R2: the two human experts. LLM: mean over the three LLM–LLM pairs (Opus–Pro, Opus–Flash, Pro–Flash). All 5: multi-rater AC across R1, R2, and the three LLM judges. The full 10-pair matrix, including the six human–LLM cross pairs, is in Appendix P.

	Full			No-form.			Zero-shot		
Metric	R1–R2	LLM	All5	R1–R2	LLM	All5	R1–R2	LLM	All5
Ordinal (AC2, quadratic weights)									
Clarity	0.974	0.950	0.896	0.941	0.978	0.857	0.937	0.987	0.754
Relevance	0.973	0.976	0.945	0.956	0.980	0.901	0.923	0.979	0.775
Importance	0.977	0.963	0.951	0.903	0.947	0.888	0.902	0.863	0.783
Grounded	0.976	0.935	0.871	0.962	0.971	0.839	0.937	0.962	0.657
Narrative	0.983	0.976	0.951	0.960	0.945	0.859	0.944	0.848	0.707
Explicit	0.983	0.998	0.975	0.961	1.000	0.913	0.971	0.998	0.766
Focused	0.993	0.899	0.869	0.940	0.927	0.830	0.908	0.917	0.701
DSM-5 binary (AC1)									
Trauma exposure	1.000	0.895	0.925	1.000	0.797	0.842	0.855	0.927	0.867
Intrusion	1.000	0.937	0.887	0.981	0.935	0.869	0.933	0.975	0.916
Avoidance	1.000	0.931	0.902	0.991	0.852	0.861	0.831	0.915	0.803
Negative cognitions	1.000	0.975	0.983	1.000	0.969	0.982	0.952	0.797	0.792
Hyperarousal	1.000	0.973	0.966	1.000	0.952	0.945	0.961	0.988	0.966
Impairment	1.000	0.981	0.989	1.000	0.994	0.996	1.000	0.981	0.985
Ehlers & Clark binary (AC1)									
Threat	0.820	0.753	0.723	0.427	0.555	0.348	0.674	0.553	0.096
Appraisals	0.877	0.896	0.801	0.734	0.707	0.432	0.580	0.642	0.336
Memory	0.946	0.750	0.697	0.886	0.644	0.418	0.710	0.543	0.277
Strategies	0.952	0.946	0.882	0.535	0.580	0.426	0.356	0.253	0.108
Triggers	0.902	0.944	0.853	0.814	0.896	0.723	0.761	0.915	0.677

The conditional gradient is clearest on the EC items: *Threat* falls from All-5 AC = 0.72 in full to 0.10 in zero-shot, *Strategies* from 0.88 to 0.11, *Memory* from 0.70 to 0.28, and *Appraisals* from 0.80 to 0.34 (Table 3, rightmost columns within each block; visualised in Figure 19). This is the same conditional gradient the scaled judge reports for EC *coverage* in §5.5: the cognitive formulation does not merely make raters score higher on EC presence; it makes raters agree on what they see, because the components are actually there in the text. Without the formulation specification, EC content is sparse and inconsistent across vignettes, and raters disagree on whether it is present at all.

5.5 Clinical-Content Quality

Two experts rate the 330-vignette subsample (blind) and `deepseek-v4-flash` scales the schema to all 15,000 retained-model vignettes ($n = 5{,}000$ per condition). Pairing is at the (persona × generation model) unit. See Appendices Q, R.

Diagnostic validity. Before any condition contrast, the vignettes clear the diagnostic floor. Coverage of the six DSM-5 PTSD criteria is at ceiling in all three conditions (0.85–1.00), independently for the two clinical experts (Appendix Q) and for the scaled judge across all 15,000 vignettes (Appendix R). Every condition therefore yields recognisable PTSD presentations, and the effects reported below are differences in cognitive-formulation structure between vignettes that are already diagnostically valid, not differences in whether the text depicts PTSD at all.

Experts. Omnibus reaches $p < .001$ on all seven ordinal items, three of six DSM-5 items, four of five EC items. *Importance* shows the strongest contrast (full 2.94 vs zero-shot 2.23, $d_z = +1.98$); on EC items the gap is larger than the scaled judge's: Threat $0.86 \to 0.19$ ($d_z = +1.58$), Appraisals $0.85 \to 0.40$ ($+0.88$), Memory $0.91 \to 0.53$ ($+0.71$), Strategies $0.85 \to 0.44$ ($+0.76$). Non-significant omnibi cluster where the condition is structurally invariant (DSM-5 at ceiling, Triggers narratively familiar).

Scaled judge. (i) EC coverage is where the formulation pays: Strategies (full 0.91 vs zero-shot 0.49), Appraisals (0.85 vs 0.52), Memory (0.60 vs 0.42) show the largest gaps; Triggers saturates at ~ 0.93. (ii) Narrative (2.89 vs 2.47) and Importance (2.82 vs 2.54) follow. Clarity slightly favours the no-formulation condition ($d = -0.09$), an LLM-judge preference for unconstrained prose that the experts overturn on the same item ($d_z = +0.81$; per-rater means in Appendix S). Vignette lengths are comparable across conditions (full $\mu = 753$ vs zero-shot $\mu = 682$ words, $\sim 10\%$ grand-mean difference); the EC-Strategies and EC-Appraisals effects survive length-matched within-persona pairing ($d_z = +0.72$ and $+0.53$ at median 16-word length difference; Appendix N.4).

5.6 Clinician User Study

Expert and judge evaluations test content; they do not test whether practising clinicians find the vignettes plausible. We recruited $N = 100$ licensed clinicians with PTSD-patient experience (psychologists, psychiatrists, MD GPs, nurses; Appendix T) via an online panel, paid £7.50/hour pro-rated; ethics approval, informed consent, and debriefing in Appendix T. Each rated 17–18 randomly assigned vignettes plus an attention check, blind to model and condition (1,600 ratings, 43–57 per cell). They rated the three CVI items and perceived AI authorship (binary). We fit LMMs for the ordinal outcomes and GEE logistic for the binary (Liang and Zeger, 1986), clustering on participant.

Perceived quality. The condition effect is monotone on every item. Full vignettes are rated > 1 point higher on the 1–3 scale than zero-shot by the same clinician: Clarity b=+1.087, Relevance b=+1.109, Importance b=+1.111 (all $t > 34$, $p < 10^{-250}$); the no-formulation contrast is half the size on every metric ($p < 10^{-58}$). Par-

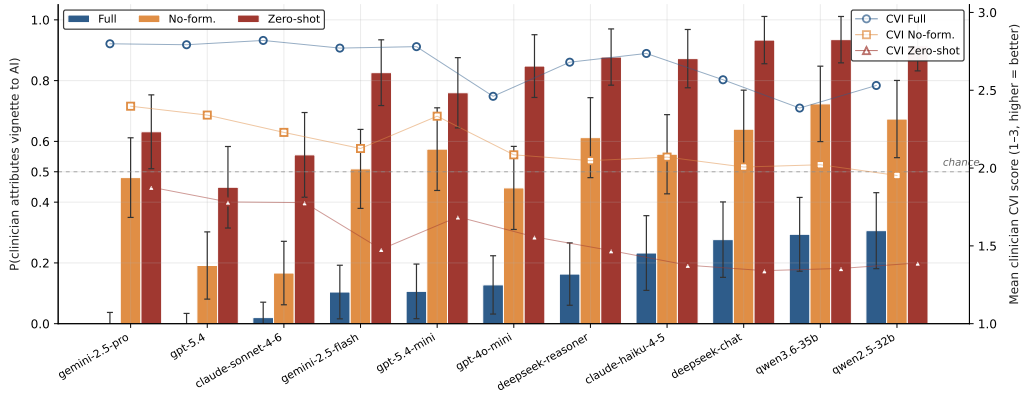


Figure 5: Clinician AI-authorship attribution rate (bars, left axis) and mean CVI quality score (markers, right axis) by generation model and condition. $N = 100$ clinicians. Models ordered by ascending full-condition AI rate. Bars are Wilson 95% CI. Two large proprietary models (`gpt-5.4`, `gemini-2.5-pro`) reach 0% AI-attribution in the full condition while scoring near-ceiling CVI; smaller open-weight models remain recognisable as AI (28–31%) but still receive higher CVI in full than in zero-shot.

ticipant ICC is $\sim$ 0.08; the condition effect is roughly an order of magnitude larger than between-participant variance (Appendix U). Clinicians spent 31% longer reading full than zero-shot vignettes.

Perceived AI authorship. Clinicians perceive AI authorship in 14.7% of full vignettes, 50.8% in no-formulation, 77.8% in zero-shot. GEE confirms all contrasts (full-vs-zero-shot OR $=$ 0.05, $z = -18.5$; full-vs-no-formulation OR $=$ 0.17; no-form-vs-zero-shot OR $=$ 0.30; all $p < 10^{-15}$; Appendix V). Per-model (Figure 5; Appendix W): `gpt-5.4` and `gemini-2.5-pro` reach 0% AI-attribution in full; open-weight models remain perceivable at 28–31%; the formulation lift is large for all models but does not equalise model quality.

5.7 Fairness by Demographic and Condition

We test if the framework's output is demographically equivariant across persona gender, ethnicity, age, occupation, and trauma category, using Statistical Parity Difference (SPD; Dwork et al., 2012), Disparate Impact Ratio (DIR; Feldman et al., 2015), and pairwise Cohen's d (Cohen, 1988) alongside omnibus F-tests and demographic-by-condition interactions (Appendix X). The formulation collapses demographic disparity in scaled-judge perceived quality: every demographic SPD falls 1.5–7$\times$ from zero-shot to full, full-condition DIRs exceed the four-fifths-rule threshold, $d_{\max} \leq 0.12$. EC coverage is invariant on gender, ethnicity, and age, with a residual content-fit disparity on trauma category (childhood-trauma narratives engage all five EC components more naturally than other narratives).

6 Discussion and Conclusion

We introduced FORMA, a theory-grounded framework that compiles a cognitive model of a disorder into a directed weighted graph, samples a configuration for each persona, and conditions LLM generation on the resulting structural specification. The four evaluation tiers converge on a single finding: this changes the structural content of the generated vignettes, not only their surface style. The specified directed graph is recoverable from full-condition vignettes but not from zero-shot or no-formulation vignettes; the formulation makes generation models converge on a shared persona portrayal and breaks zero-shot trauma-type stereotyping; experts rate full vignettes higher than zero-shot; clinicians read full vignettes as human-written far more often than zero-shot vignettes; and the framework reduces demographic disparity in perceived quality on every demographic axis tested.

Across our evaluation tiers, LLM judges mostly "*agreed*" on clinical text regardless of what the text contained, and on stylistic items their consensus ran counter to clinical experts. Inter-rater agreement on its own is therefore not evidence that an LLM-generated artefact is faithful to its specification. When the specification is itself structured, as a cognitive-formulation graph is here, it can serve as an external gold for the generated text; in this work the prompted graph is what makes the formulation effect identifiable. Compiling a cognitive model into an auditable directed graph turns synthetic case material into something a clinician can read for clinical plausibility and theoretical structure, and should generalise wherever a disorder admits a formal model.

Limitations

First, the in-loop Vignette Validator shares the same LLM model as the Vignette Crafter, so the per-model iteration counts in Appendix C reflect both generation ability and self-critique; the `qwen` rows, which pass on the first attempt with zero violations, illustrate the risk of fluent self-approval. We address this with an external edge probe (§5.1), but a stronger structural check with multiple independent validators is left to future work. Second, both demographics and cognitive formulation graphs are sampled programmatically from real patient data rather than representing real patients. However, working with cognitive formulations often calls for similar practices. The pipeline should be read as a controlled stress test of whether LLMs can populate a clinically structured template, not as a replacement for real clinical material; the framework's faithfulness to actual clinical heterogeneity remains an empirical question we cannot answer with this dataset alone. Third, all prompts, self-report items, and vignettes are in English, and the framework's behaviour under translation or culturally-specific persona sampling is untested. Fourth, the framework instantiates one cognitive theory (Ehlers and Clark, 2000). Cognitive theories of mental disorders are themselves contested; pinning vignette generation to one theory propagates that theory's assumptions and gaps to every vignette generated. We see this as a feature (the structural choices are explicit) but it should not be conflated with a generic claim about "what PTSD looks like". Fifth, the fairness analysis (§5.7) tests output equivariance only and does not test for stereotyping in content, representation harms in surface prose, or intersectional disparities; it also relies on a small number of unique persona profiles per demographic cell on the edge-probe subsample. Sixth, the clinician user study was conducted with a panel-recruited sample of 100 licensed practitioners, predominantly female, predominantly U.S./U.K./Canada/Australia. Generalisability to clinicians outside this profile and to other patient-care contexts is an empirical question.

7 Ethical Considerations

No real patients. The dataset contains no real patient material. All personas, demographics, self-report items, and vignettes are LLM-generated from the sampling distributions documented in §4.1. The Persona Validator and the cognitive-model parameters are drawn from published clinical theory and from prior work that re-derived the EC graph from ecological momentary assessment data on real patients (Hertz-Palmor et al., 2025); no individual-level data was ingested. This is the primary ethical advantage of synthetic case generation, and the basis on which we believe the dataset can be released openly. The clinician user study, the only component involving human participants, was reviewed and approved by the ethics committee of the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, and participants gave informed consent before rating (Appendix T).

Mis-attribution and downstream use. The clinician user study finding that full-condition vignettes pass as human-written 85% of the time is itself an ethical concern as well as a methodological one. If the dataset is incorporated into clinical training material or downstream LLM-evaluation benchmarks without clear provenance, end users may take it for material derived from real patients. We recommend that any release-time distribution of the dataset prominently and persistently label every vignette as synthetic and as not representing a real individual. A second risk runs the other way. Used as training or evaluation data for downstream clinical models, these vignettes propagate the framework's theoretical commitments and the generators' stylistic regularities into those models, amplifying the assumptions of a single cognitive theory and one generation pipeline; we take up that case below.

Limits of theory-grounded generation. A vignette that is faithful to a cognitive model is still only as accurate as that cognitive model. Pinning generation to one theory of a disorder propagates that theory's framings, its blind spots, and any populations the theory underrepresents. The EC model (Ehlers and Clark, 2000) is well-validated for the cognitive maintenance of PTSD in Western clinical samples; it is not a theory of culturally-specific trauma response, of complex PTSD, or of disorders co-morbid with PTSD. Our vignettes inherit this scope.

Bias and fairness. §5.7 reports an algorithmic-fairness analysis showing that the framework does not introduce demographic disparities in EC coverage and substantially reduces them in perceived quality. We emphasise the limitations of that analysis: it tests output equivariance on the variables the dataset records, not deeper representation harms or stereotyping in the textual content. We therefore caution against using the dataset to evaluate *clini-*

cal bias of downstream LLMs until those further axes have been audited.

Use in training mental-health LLMs. Synthetic case material is a tempting training corpus for clinical LLMs because it bypasses privacy constraints. We urge restraint. The model that generates the training data shapes the model that consumes it; downstream LLMs trained on FORMA vignettes will inherit our generators' rhetorical patterns and our framework's theoretical commitments. We see appropriate uses as standardised stimuli for clinician training (where the synthetic origin is disclosed), benchmark sets for clinical capabilities of LLMs (where the gold structure is the audited contribution, not the prose), and as a probe-set for inter-rater reliability research. We do not see appropriate uses as training data for clinical decision-support LLMs deployed in patient care.

Reproducibility and release. The full dataset and code are released at `https://github.com/Amit-Oren/FORMA`, under the MIT License for the code and CC BY 4.0 for the datasets. We additionally provide an interactive web browser, `https://clinical-vignettes.streamlit.app/`, where readers can inspect any of the 500 full-condition vignettes alongside its persona cognitive graph and the corresponding self-report items, making the vignette–graph correspondence visible end-to-end. We release the full 16,500-vignette dataset, the 330-vignette evaluation subsample with ratings from two clinical experts and three LLM judges, the user-study ratings ($N = 1{,}706$ total: 1,600 vignette ratings + 106 attention-check responses), all generation and evaluation code, and the cognitive-formulation graphs sampled per persona. Generation hyperparameters, total compute, and provider routing are documented in Appendix G; the full system prompts for the Crafter, Validator, Edge Probe, and Judge are in Appendix H; Appendix I reproduces a representative full-condition vignette so that readers without the released code can see what the system produces, with an expert annotation of each EC-relevant span and the persona's cognitive graph (Figure 6). Sensitivity to analytical choices (edge-weight threshold, judge identity, encoder, cluster operationalisation) is reported in Appendix N. The user-study data are released in aggregate; participant identifiers are randomised and demographic combinations that would re-identify individual respondents are coarsened or suppressed.

AI assistant disclosure. LLMs are the primary research artifact of this paper; their use as generators, validators, and judges is documented in §3, §4.2, §5, and Appendix H. Separately, AI assistants (LLMs) were used for light copy-editing and LaTeX formatting during manuscript preparation; all scientific claims, analyses, and figures were authored and verified by the human authors.

Acknowledgments

Nimrod Hertz-Palmor is supported by the Gates Cambridge Trust (#OPP1144).

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, 5th edition. American Psychiatric Publishing, Arlington, VA.

Aaron T Beck, A John Rush, Brian F Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. Guilford Press, New York.

James RA Benoit. 2023. Chatgpt for clinical vignette generation, revision, and evaluation. *MedRxiv*, pages 2023–02.

Neel Jitesh Bhate, Ansh Mittal, Zhe He, and Xiao Luo. 2023. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using gpt model. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1476–1480. IEEE.

Christy A Blevins, Frank W Weathers, Margaret T Davis, Tracy K Witte, and Jessica L Domino. 2015. The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, 28(6):489–498.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

David M Clark and Adrian Wells. 1995. A cognitive model of social phobia. In Richard G Heimberg, Michael R Liebowitz, Debra A Hope, and Franklin R Schneier, editors, *Social Phobia: Diagnosis, Assessment, and Treatment*, pages 69–93. Guilford Press, New York.

William G. Cochran. 1950. The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226.

Anke Ehlers and David M Clark. 2000. A cognitive model of posttraumatic stress disorder. *Behaviour research and therapy*, 38(4):319–345.

Equal Employment Opportunity Commission. 1978. Uniform guidelines on employee selection procedures. 29 CFR Part 1607.

Spencer C Evans, Michael C Roberts, Jared W Keeley, Jennifer B Blossom, Christina M Amaro, Andrea M Garcia, Cathleen Odar Stough, Kimberly S Canter, Rebeca Robles, and Geoffrey M Reed. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in icd-11 field studies. *International journal of clinical and health psychology*, 15(2):160–170.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268.

Vincent J Felitti, Robert F Anda, Dale Nordenberg, David F Williamson, Alison M Spitz, Valerie Edwards, Mary P Koss, and James S Marks. 1998. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) study. *American Journal of Preventive Medicine*, 14(4):245–258.

Erlend Frayling, Jake Lever, and Graham McDonald. 2024. Zero-shot and few-shot generation strategies for artificial clinical records. *arXiv preprint arXiv:2403.08664*.

Isaac R Galatzer-Levy and Richard A Bryant. 2013. 636,120 ways to have posttraumatic stress disorder. *Perspectives on psychological science*, 8(6):651–662.

Kilem L. Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4th edition. Advanced Analytics, LLC.

Nimrod Hertz-Palmor, Gary P. Brown, Omid V. Ebrahimi, Samuel Hardman, Lars Klintwall, Amit Lazarov, Alicja Podgorski, Jana Porat, Tim Dalgleish, and Anna Bevan. 2026. Momentary structured networks (MSN): A theory-guided approach to person-elicited causal formulation. PsyArXiv preprint.

Nimrod Hertz-Palmor, Alicja Podgorski, Gary Brown, Amit Lazarov, Jana Porat, Shany Yaron, Anna Bevan, and Tim Dalgleish. 2025. Quantifying clinical case formulation: Person-specific dynamic network modelling of posttraumatic stress disorder (PTSD). Preprint. MRC Cognition and Brain Sciences Unit, University of Cambridge.

Mary Ann Heverly, David X Fitt, and Frederick L Newman. 1984. Constructing case vignettes for evaluating clinical judgment: an empirical model. *Evaluation and Program Planning*, 7(1):45–55.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of NAACL-HLT*, pages 2284–2293.

Guy Laban, Julie Wang, and Hatice Gunes. 2026. A robot-led intervention for emotion regulation: from expression to reappraisal. *IEEE Transactions on Affective Computing*, pages 1–15.

Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Deepika Mohan, Baruch Fischhoff, Coreen Farris, Galen E Switzer, Matthew R Rosengart, Donald M Yealy, Melissa Saul, Derek C Angus, and Amber E Barnato. 2014. Validating a vignette-based instrument to study physician decision making in trauma triage. *Medical decision making*, 34(2):242–252.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jacqueline B Persons. 2008. *The Case Formulation Approach to Cognitive-Behavior Therapy*. Guilford Press, New York.

Daniel Reichenpfader and Kerstin Denecke. 2024. Simulating diverse patient populations using patient vignettes and large language models. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pages 20–25.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Peter B Rosenquist, Christopher C Colenda, Judy Briggs, Stephen I Kramer, and Michael Lancaster. 2000. Best practices: using case vignettes to train clinicians and utilization reviewers to make level-of-care decisions. *Psychiatric Services*, 51(11):1363–1365.

Paul M Salkovskis. 1985. Obsessional-compulsive problems: A cognitive-behavioural analysis. *Behaviour Research and Therapy*, 23(5):571–583.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Jessica Sheringham, Isla Kuhn, and Jenni Burt. 2021. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. *BMC medical research methodology*, 21(1):81.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33*.

Barbara St. Marie, Andrea Jimmerson, Yelena Perkhounkova, and Keela Herr. 2021. Developing and establishing content validity of vignettes for health care education and research. *Western journal of nursing research*, 43(7):677–685.

BN Suhas, Andrew M Sherrill, Rosa I Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy conversations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sichang Tu, Abigail Powers, Stephen Doogan, and Jinho D Choi. 2025. TRUST: An LLM-based dialogue system for trauma understanding and structured assessments. *Preprint*, arXiv:2504.21851.

U.S. Census Bureau. 2025. QuickFacts: United States. `https://www.census.gov/quickfacts/fact/table/US/LFE046223`. Retrieved 2025-02-18.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. In *Journal of Machine Learning Research*, volume 9, pages 2579–2605.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, and 1 others. 2024. PATIENT-Ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Lixiu Wu, Yuanrong Tang, Qisen Pan, Xianyang Zhan, Yucheng Han, Lanxi Xiao, Tianhong Wang, Chen Zhong, and Jiangtao Gong. 2026. An LLM-based simulation framework for embodied conversational agents in psychological counseling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yasutaka Yanagita, Daiki Yokokawa, Shun Uchida, Yu Li, Takanori Uehara, and Masatomi Ikusaka. 2024. Can ai-generated clinical vignettes in japanese be used medically and linguistically? *Journal of General Internal Medicine*, 39(16):3282–3289.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

A The Five Ehlers & Clark Components

The cognitive graph used throughout this paper is the five-component formulation of Ehlers and Clark (2000), the dominant cognitive model of PTSD in UK clinical practice and the basis for NICE-recommended trauma-focused CBT. The model proposes that PTSD is maintained by the interaction of five components rather than by any single symptom cluster; below we summarise each and note how it is operationalised in our graph and in the rater anchors of the schema judge (Appendix H).

Trauma memory. The autobiographical memory of the traumatic event(s), characterised in the model by poor elaboration, fragmentation, and a strong sensory and "here-and-now" quality, such that retrieval is not experienced as remembering past danger but as re-living it. Operationally, we encode it as one of the five graph nodes; the schema judge scores it 1 when the vignette shows fragmented, disorganised, or present-tense intrusive recall (rather than ordinary autobiographical recall of the trauma).

Re-experiencing triggers. Specific internal or external cues (sights, sounds, sensations, contexts, internal states) that reliably reactivate the trauma memory and accompanying distress. Cues may be incidental and not obviously linked to the trauma. Operationally, this node is scored 1 when the vignette describes identifiable cues that set off intrusion, hyperarousal, or avoidance.

Sense of current threat. The pervasive subjective experience that danger is ongoing rather than past, manifesting as hypervigilance, startle, and bodily threat-response. This is the model's central maintaining mechanism (the model's title concept). Operationally, scored 1 when the vignette portrays the person as feeling currently in danger, not merely cautious or unsettled.

Negative appraisals. Catastrophic or overly negative interpretations of the trauma, of the person's reactions to it, and of its sequelae (e.g. "I am permanently damaged", "the world is unsafe", "it was my fault"). Appraisals are the cognitive content the model targets for therapeutic change. Operationally, scored 1 when explicit catastrophic or self-blaming beliefs are stated or shown.

Maladaptive cognitive/behavioural strategies. Strategies the person uses to manage symptoms that, while affording short-term relief, prevent elaboration of the trauma memory and disconfirmation of the negative appraisals, thereby maintaining the disorder. Canonical examples: thought suppression, rumination, situational avoidance, safety behaviours, substance use. Operationally, scored 1 when the vignette shows such a strategy in action with implied or explicit symptom-maintaining function.

Directed edges between components. The model posits a small number of directed causal links between these components (e.g. Triggers → Memory → Threat → Maladaptive Strategies → Negative Appraisals → Threat). We do not hand-fix the edge set: each persona's directed weighted graph is sampled from per-edge weight distributions calibrated against the EMA-derived networks of Hertz-Palmor et al. (2025) (sampling procedure in §3.1), and the external edge probe (§5.1, Appendix L) evaluates whether each of the 20 possible directed pairs is present in the generated text given that it was active in the persona's specification.

B Self-Report Item Pool

The item pool contains 403 items total, distributed across the five Ehlers & Clark components (§3.1, Step 3). Each item is a short clinical sentence reflecting a specific cognitive, behavioural, or experiential feature of the component. The pool derives from real-patient disclosures to the five EC components in a clinical sample (Hertz-Palmor et al., 2026): patient disclosures were parameterised into per-component item distributions, and each persona's items are drawn at random from these distributions. The Vignette Validator (§3.3) then verifies that the sampled set is realised in the generated text. For Maladaptive Strategies the pool merges three sub-pools (core maladaptive strategies, distraction strategies, and avoidance strategies) into a single sampling distribution. A small sample per component is shown below; the full pool is available in the code release.

Pool size per component.

- Triggers: 137 items
- Maladaptive Strategies: 130 items (48 core + 39 distraction + 43 avoidance)
- Negative Appraisals: 58 items
- Memory characteristics: 39 items
- Current threat appraisals: 39 items

Example items, Triggers.

- *Olfactory (Alcohol)*: "Alcohol smell"
- *Olfactory Danger Cue*: "Burnt smell"
- *Auditory (Babies)*: "Babies"
- *Auditory (Cars)*: "Car sounds"
- *Auditory/Threat*: "Gunshot noise"

Example items, Maladaptive Strategies.

- *Suppress Thoughts*: "try hard not think about the trauma"
- *Mental Distraction*: "keep mind occupied all the time"
- *Alcohol Use*: "drink alcohol"
- *Drug Use*: "take drugs"
- *Avoid All Feelings*: "avoid anything that could cause negative or positive feelings"

Example items, Negative Appraisals.

- *Self-Blame*: "The event happened because of the way I acted"
- *Self-Doubt*: "I can't trust that I will do the right thing"
- *Weakness Belief*: "I am a weak person"
- *Fear of Rage*: "I will not be able to control my anger and will do something terrible"
- *Low Coping*: "I can't deal with even the slightest upset"

Example items, Memory characteristics.

- *Visual Memories*: "My memories of the frightening event are mostly pictures or images."
- *Nonverbal Memory*: "I can't seem to put the frightening event into words."
- *Auditory Recall*: "When I have memories of what happened I sometimes hear things in my head that I heard during the frightening event."
- *Reliving Event*: "When I remember the frightening event I feel like it is happening right now."
- *Olfactory Recall*: "When I think about the frightening event I can sometimes smell things that I smelt when the frightening event happened."

Example items, Current threat appraisals.

- *Flashbacks*: "Suddenly feeling or acting as if the stressful experience were actually happening again"
- *Physical Reactivity*: "Having strong physical reactions when something reminded you of the stressful experience"
- *Negative Emotions*: "Having strong negative feelings such as fear, horror, anger, guilt, or shame"
- *Irritability*: "Irritable behavior, angry outbursts, or acting aggressively"
- *Risky Behavior*: "Taking too many risks or doing things that could cause you harm"

C Validation Loop Iteration Counts

Table 4 reports the distribution of validation-loop iterations per generation model in the full condition. Most models pass on the first attempt; the `qwen` rows reflect lenient self-validation (Crafter and Validator share the same LLM), and `gpt-4o-mini` hits the iteration cap due to a markup-compliance failure (§3.3).

Table 4: Distribution of validation loop iterations per model in the full condition ($n = 500$ each). Columns show the percentage of vignettes that terminated after exactly 1, exactly 2, or 3 or more iterations, and the largest number of iterations observed. Since Crafter and Validator are the same model, the qwen rows reflect lenient self-validation rather than error-free generation.

Model	1 iter	2 iters	3+ iters	max
qwen2.5-32b	100.0%	0.0%	0.0%	1
qwen3.6-35b	100.0%	0.0%	0.0%	1
claude-sonnet-4-6	97.0%	3.0%	0.0%	2
deepseek-chat	94.6%	4.6%	0.8%	5
claude-haiku-4-5	85.8%	8.0%	6.2%	5
gpt-5.4-mini	85.4%	10.6%	4.0%	6
gpt-5.4	84.4%	11.4%	4.2%	6
gemini-2.5-pro	66.4%	23.8%	9.8%	4
gemini-2.5-flash	60.6%	29.6%	9.8%	5
deepseek-reasoner	54.4%	32.6%	13.0%	6
gpt-4o-mini	3.8%	0.6%	95.6%	6

D Demographic Independence Tests

The persona pipeline (§3.1) samples demographic attributes (gender, ethnicity, age, occupation, relationship status) and clinical attributes (trauma type, PCL-5 severity) from independent categorical distributions, so that downstream condition and fairness comparisons are not confounded by demographic–clinical dependencies that would arise if, say, a particular trauma type were over-represented among one ethnicity. In real clinical samples these variables are correlated (interpersonal-violence histories are unevenly distributed across genders; combat trauma is unevenly distributed across age and occupation), and a generation framework that reproduces those correlations would make condition contrasts uninterpretable: any difference between conditions could be attributable to a confound between demographics and trauma content rather than to the formulation.

We therefore tested every pair (demographic, clinical) for statistical independence in the realised persona set using Pearson's chi-square test of independence, treating each persona as an observation and binning continuous variables (age) into deciles. Table 5 reports the test statistic, degrees of freedom, and p-value for each pair. No association reaches $p < .05$; the smallest p-value (age $\times$ trauma type, $p = .063$) is consistent with the null at this sample size. We conclude that the persona set realises the intended factorial structure: every demographic stratum is approximately equally populated by every trauma type, so condition contrasts within strata in the fairness analysis (§5.7) cannot be confounded by clinical-content imbalance.

A caveat: independence at the marginal level does not guarantee independence at the joint level. Intersectional cells (e.g., "female $\times$ South-Asian $\times$ childhood-trauma") contain a small number of personas ($\sim$ 1–3) and we do not test higher-order interactions because they would be underpowered. Reported fairness metrics are accordingly limited to single-axis disparity.

Table 5: Chi-square tests of independence between demographic and clinical variables. No association is significant at $p < .05$.

Association	χ^2	*df*	*p*
Gender × Trauma type	58.36	56	.389
Gender × Occupation	1.89	2	.388
Ethnicity × Trauma type	186.10	196	.683
Ethnicity × Occupation	5.29	7	.625
Age (binned) × Trauma type	135.77	112	.063
Occupation × Trauma type	27.97	28	.466

E Persona Demographic Distributions

This section gives the full distributional breakdown of the 500-persona set summarised in §4.1. We report exact frequencies because they are the demographic universe over which every downstream condition contrast, fairness analysis, and inter-rater agreement statistic is computed: they fix the population that any disparity claim is defined relative to. Table 6 consolidates every persona-level variable in a single view; the notes below the table cover sampling-design details that the numbers alone do not communicate.

Design notes on Table 6. Categorical demographic sampling follows the reasoning of Suhas et al. (2025): gender, ethnicity and relationship-status weights are grounded in U.S. population distributions reported by the U.S. Census Bureau (2025), then perturbed lightly toward category-uniformity so that downstream per-cell fairness tests retain power. *Age* is sampled uniformly over 18–80 rather than from a population-realistic distribution because the goal is a stress test of the framework, not epidemiological representativeness. *Gender* has a slight male majority (51.6%) as an artefact of the categorical sampler — we did not enforce exact parity because that would have over-constrained the joint with other demographics. *Ethnicity* uses an eight-region scheme that conflates national-origin, racial, and cultural categories; we use it because (a) it is the granularity at which downstream LLMs reliably condition on "ethnicity" as a generation cue, and (b) finer-grained sampling would render per-cell fairness analyses underpowered. *Relationship status* is sampled conditional on age (e.g., widowed unlikely under 30) and validated by the Persona Validator agent.

Occupation and trauma roll-ups. Occupation is sampled as a free-text job title (135 unique titles across the 500 personas); for fairness analysis we bucket those titles into the six categories above using the keyword-based `occupation_bucket` routine. The *Other / unclassified* bucket is large (60.8%) because the keyword list is intentionally conservative — it only labels a title with a category if at least one anchor keyword is matched (e.g., a "Dancer" or "Park Ranger" lands in *Other*). Per-cell fairness tests are run on the five labelled buckets and reported in §5.7.

Trauma categories. Trauma is sampled across 29 raw categories spanning interpersonal violence (e.g., physical and sexual abuse, stalking), childhood trauma, war and collective violence (e.g., combat, terrorism, torture), accidental and life-threatening events, and loss/medical trauma. The five most frequent raw categories are physical abuse by a romantic partner (28, 5.6%), and assault, civilian exposure in a war zone, physical abuse by a caregiver in childhood, and accidentally causing serious injury to another (each 23, 4.6%); the remaining 24 categories each contribute 14–22 personas. The list of 29 categories is adapted from the trauma classifications used in the Life Events Checklist (LEC-5) supplemented with childhood-specific items from the ACE inventory (Felitti et al., 1998); we widened the LEC-5 set because its 17-item form aggregates childhood maltreatment into a single "physical abuse / sexual abuse" bucket that under-represents the cognitive heterogeneity of early trauma cases. For fairness analysis we roll the 29 raw categories into the six buckets in the table above.

PCL-5 severity. The PCL-5 is the standard 20-item self-report measure of PTSD severity with a clinical cutoff of 33 (Blevins et al., 2015); every persona in our set exceeds this cutoff by design, so the dataset represents a probable-PTSD population rather than a general one.

F Mean Token Counts by Model and Condition

Table 7 reports the mean total tokens (prompt + completion) per generated vignette by model and condition. Token counts encode two distinct generator behaviours that the body discussion does not separate. First, the formulation graph adds roughly 2,000–4,000 tokens of specification content (component list, edge list, weights, self-report items) that every model has to process, and the validation loop appends Crafter-generated patches until structural coverage is achieved, so completion tokens

Table 6: Persona demographic distributions across the 500-persona set ($N = 500$). Continuous variables: mean, SD, range. Categorical variables: counts and percentages over $N = 500$. The eight-region ethnicity scheme is coarse by design (see notes); the occupation bucketing follows the fairness-analysis taxonomy in §5.7 (`occupation_bucket` in `code/analysis/fairness_analysis.py`); the six-category trauma roll-up corresponds to `trauma_bucket`. The 29-category trauma list (raw) is summarised below the table.

Variable	**Level**	n	**%**
Continuous variables			
Age (years)	$M = 48.5$, $SD = 18.2$, range 18–80	500	100.0
PCL-5 severity	$M = 57.8$, $SD = 13.9$, range 33–80	500	100.0
Age band			
	18–29	95	19.0
	30–44	125	25.0
	45–59	121	24.2
	60–80	159	31.8
Gender			
	Male	258	51.6
	Female	238	47.6
	Non-binary	4	0.8
Ethnicity (8 world regions)			
	South-Eastern Asian	71	14.2
	European	70	14.0
	North American	68	13.6
	African	62	12.4
	South Asian	60	12.0
	Latin American	59	11.8
	Oceanian	57	11.4
	Middle Eastern	53	10.6
Relationship status			
	Married	239	47.8
	Single	95	19.0
	Divorced	72	14.4
	In a relationship	48	9.6
	Widowed	46	9.2
Occupation (6-bucket roll-up; 135 *unique free-text titles)*			
	Professional / tech	68	13.6
	Trades	63	12.6
	Healthcare	34	6.8
	Uniformed (police/fire/military)	16	3.2
	Education	15	3.0
	Other / unclassified	304	60.8
Trauma category (6-bucket roll-up of 29 raw categories)			
	Interpersonal violence (abuse, assault, stalking)	104	20.8
	War / collective violence (combat, terrorism, torture)	95	19.0
	Accident / disaster	75	15.0
	Loss / medical	49	9.8
	Childhood (non-abuse)	16	3.2
	Other / mixed	161	32.2

also grow. Second, generation models differ widely in how verbose their first-pass vignette is: `gemini-2.5-flash` and `gemini-2.5-pro` produce 2,700+ tokens even in zero-shot, while `qwen2.5-32b` and `qwen3.6-35b` produce roughly one-third that. The cross-product of these two factors explains the $\sim 14\times$ range in full-condition tokens across models.

The `gpt-4o-mini` row stands out at 29,234 tokens in the full condition: this is not a verbose generation, but a refinement-loop pathology. Because `gpt-4o-mini` fails to emit valid patch-block markers (§3.3), the Validator continues to flag the same edge violations across iterations, and each iteration re-sends the prompt plus the accumulated context to the Crafter. The token count thus reflects an interrupted control flow rather than a structurally rich vignette, which is one of the reasons we exclude this model from per-condition analyses.

The token count is not a quality measure:

gemini-2.5-flash produces 13,651-token full vignettes and rates highly on both expert and clinician scales, while qwen2.5-32b produces 3,284-token full vignettes and rates substantially lower. We report token counts for reproducibility (downstream users budgeting inference cost) and for transparency about the cost of the cognitive-graph specification; we do not use token counts to filter or weight the per-condition analyses.

Table 7: Mean total tokens per vignette by model and condition.

Model	Zero-shot	No-formulation	Full
claude-haiku-4-5	1,007	5,768	6,602
claude-sonnet-4-6	1,062	4,214	5,062
deepseek-chat	1,050	3,479	4,159
deepseek-reasoner	1,443	6,108	11,205
gemini-2.5-flash	2,779	6,075	13,651
gemini-2.5-pro	2,739	6,379	11,024
gpt-4o-mini	1,091	3,238	29,234
gpt-5.4	1,255	3,539	5,187
gpt-5.4-mini	1,111	3,472	5,144
qwen2.5-32b	918	2,790	3,284
qwen3.6-35b	910	2,919	3,479

G Generation Hyperparameters and Compute

Sampling. All eleven generation models and the five rating agents (two human-blind expert raters do not apply here; the three LLM judges, the persona Crafter/Validator, the vignette Crafter/Validator, and the external edge probe) use *temperature* $= 0.7$ with provider-default top_p and provider-default max_tokens; we did not set explicit nucleus or top-k truncation. Three model families behave slightly differently from the default: (i) the claude family does not accept a temperature parameter via the Anthropic SDK, so for these models we use the API default; (ii) the deepseek-reasoner model uses its built-in reasoning trace and ignores nucleus parameters; (iii) the open-source qwen models are served through an internal lab inference endpoint and use the same temperature setting. Generation seeds are not fixed: we report effect sizes across 500 personas $\times$ 11 models $\times$ 3 conditions, so model-level stochasticity is absorbed into the design rather than controlled out. The validation loop retries up to 5 times (six iterations total); persona construction retries each demographic field up to 6 times.

Architecture and routing. Generation requests are routed through provider-specific LangChain wrappers (ChatOpenAI, ChatAnthropic, ChatGoogleGenerativeAI, a custom DeepSeekChatModel pointing at api.deepseek.com, and ChatOllama/custom open-source-LLM clients for qwen2.5-32b, qwen3.6-35b). Structured outputs for the Validator and Judge use LangChain's Pydantic-typed structured-output decoder; we manually retry one time on JSON parse failure before logging the call as a structural-validation failure.

Compute. All eleven generation models, the five LLM judges, and the external edge probe were accessed through commercial APIs (OpenAI, Anthropic, Google, DeepSeek) or through an internal lab inference endpoint (Alibaba qwen); we ran no LLM inference locally for generation or judging. Total token consumption across the 16,500-vignette generation set, the 5,000-vignette scaled-judge pass, the 330-vignette 3-LLM-judge pass, and the 6,600-cell edge-probe pass is reported in Appendix F (per-cell averages). Embedding analyses (MPNet, BGE-base) ran on a single NVIDIA RTX 3050 Ti Laptop GPU (4 GB VRAM, CUDA 12.6), $\sim$ 90 minutes total wall-clock for the 16,500-vignette BGE pass and a comparable CPU pass for MPNet (used in the body).

H System Prompts

This section reproduces the system prompts used by the four LLM agents that govern the structural part of the framework: the Vignette Crafter (§3; condition: *full*), the Vignette Validator (§3.3), the external Edge Probe (§5.1), and the EC-coverage Judge (§5.1). The user-specification templates that wrap these system instructions at inference time are released with the code; the persona Crafter/Validator instructions and the no-formulation / zero-shot variants are deferred to the released repository for space.

Vignette Crafter, full condition

```
You are a clinical psychologist writing
    psychological case
vignettes grounded in Ehlers & Clark's (2000)
    cognitive model
of PTSD.

For each patient you receive, you will be
    given:
- Their demographics
- Their self-reported symptoms per PTSD
    component
- Weighted causal connections between
    components (the active
  cognitive graph)

Core constraints that apply to every vignette:
- Only include active components.
```

```
- Do NOT invent or infer causal links not
    present in the
  active graph.
- For EACH required causal connection (A ->
    B), ensure the
  vignette conveys -- within a paragraph --
    that A influences
  or leads to B.
- Ground every clinical detail in the
    patient's reported items.
- The causal connections are weighted (0 to
    1). These weights
  govern narrative prominence:
    > 0.5  : these connections form the
    narrative backbone
    0.1-0.5: present and integrated but not
    load-bearing
    < 0.1  : mentioned briefly and in passing
    = 0    : the component may appear but
    must be fully isolated
             from all causal language

Output format -- write 500-700 words of
    continuous third-person
prose. No headers, no numbered sections, no
    bullet points.
Tell the patient's story chronologically; the
    dominant weighted
connections (weight > 0.6) should form the
    narrative backbone
of the second half of the vignette.
```

Vignette Validator, full condition

```
You are a clinical validator checking whether
    a PTSD case
vignette accurately reflects a patient's
    required cognitive
connections.

CAUSAL STANDARD
A required edge A->B is SATISFIED if, within
    the same paragraph:
  - Both A and B (or their synonyms) appear,
    AND
  - The paragraph conveys -- through any
    language -- that A
    influences, drives, or leads to B.
This includes direct causal sentences,
    sequential structure,
or connecting phrases like "because of",
    "this meant", "so",
"which led", "as a result", "consequently",
    "in turn".
A required edge is MISSING only if no
    paragraph conveys the
A->B relationship at all.

SECTION 1 -- COMPONENT CHECK
Confirm each active component appears as a
    described feature
of the patient's presentation. (Internal
    check only -- do NOT
add component results to violations.)

SECTION 2 -- REQUIRED EDGES
For each edge, apply the causal standard
    above.
- If satisfied: do nothing -- only failures
    are reported.
- If missing: add it to violations and
    explain in one sentence
  what causal relationship is absent.
```

External Edge Probe

```
You are evaluating a PTSD case vignette for
    the presence of
causal connections between five Ehlers &
    Clark cognitive
components: Triggers, Negative Appraisals,
    Memory, Threat,
and Maladaptive Strategies.

For each of the 20 ordered, directed (A -> B)
    pairs of distinct
components, decide whether the vignette
    conveys, in some
paragraph, that A influences, drives, or
    leads to B. Apply the
same causal standard as the structural
    validator.

CAUSAL STANDARD
An edge A -> B is PRESENT (return 1) if,
    within the same
paragraph:
  (i) both A and B (or their synonyms)
    appear; AND
  (ii) the paragraph conveys that A
    influences / drives / leads
      to B, through direct causal language,
    sequential
      structure, or connecting phrases like
    "because of",
      "this meant", "so", "which led", "as a
    result",
      "consequently", "in turn".

You will return a binary judgement for each
    of the 20 ordered
pairs. Be strict: do not assume an edge based
    on mere
co-occurrence; the paragraph must actually
    convey directional
influence.
```

Schema Judge (used by Opus, DSpro, Flash)

The schema judge that scores the 18-item rating protocol (§4.2) issues one structured call per item, sharing a common rubric and reading one vignette at a time. The shared header is:

```
You are a critical reviewer of PTSD clinical
    vignettes for a
research study. Rate the vignette on ONE
    dimension only, using
the anchors below. Score based solely on what
    is written; do
not infer content that is absent. Before
    scoring, identify a
specific sentence that justifies your rating.

Guidelines for a well-written vignette:
  1. Derive from the literature and/or
    clinical experience
  2. Be clear, well-written, and carefully
    edited
```

```
  3. 500-700 words: enough to show a complex
     picture without
      being too long to read closely
  4. Follow a narrative, story-like
     progression
  5. Be culturally and socio-economically
     neutral
  6. Resemble real people, not a
     personification of symptoms
  7. Be relatable, relevant, and plausible to
     clinicians
  8. Avoid red herrings, misleading details,
     and bizarre content
  9. Highlight the key clinical variables of
     interest
  10. Allow vague or ambiguous elements that
     invite reflection
```

Per-item anchors are appended to this header. We give one representative item from each of the four families; the remaining 14 are released with the code.

```
Dimension: **Clarity** (1-3) -- easy to
    understand, free from
confusion, ambiguity, or obstruction.

1 = hard to follow. A short vignette can
    still be unclear.
    Example: "He was haunted by what
    happened. The past was a
    shadow that followed him, a weight that
    pressed down on his
    chest, a storm that raged inside him."
2 = readable but relies on vague or generic
    clinical language
    without specific detail.
    Example: "He experienced hypervigilance
    that made things
    hard. He avoided reminders of the trauma
    and had trouble
    sleeping."
3 = precise and specific: clear what the
    person does, thinks,
    and feels, and why.
    Example: "He lives with a constant sense
    of being on guard,
    sits with his back to walls, watches
    people's hands. When
    triggered, he feels panic mixed with
    anger."
```

```
Dimension: **g1 -- Grounded in realistic
    clinical experience**
(1-3).

1 = generic phrasing, no specific detail,
    could describe any
    PTSD patient.
2 = some realistic detail but described from
    the outside.
3 = written from inside this person's
    specific experience: the
    exact situations, paradoxes, and reason
    they sought help.
```

```
Score 1 only if the criterion is clearly and
    explicitly present
in the text; not merely inferable. Score 0 if
    absent or only
vaguely implied.

Dimension: **DSM-5 Criterion C** -- avoidance
    of trauma-related
thoughts/feelings, or of external reminders.

Score 1 example: "He declines team events in
    busy restaurants,
takes alternate routes to avoid the coastal
    expressway, and has
stopped answering automated weather alerts."
Score 0 example: "He preferred quieter
    environments and did not
like to dwell on the past."
```

```
Score 1 only if the component is explicitly
    shown; not just
consistent with having PTSD.

Dimension: **Sense of current threat** -- the
    person feels the
trauma is still happening or still dangerous
    now, not merely a
memory of past danger.

Score 1 example: "He wakes some mornings
    certain the floor will
give way: not a memory of past danger but a
    present conviction
that the ground itself is unsafe."
Score 0 example: "He remained cautious in
    general and found it
hard to fully relax."
```

Item user message. The same one-line user message wraps each item call: `Rate the following clinical vignette: {vignette}`. Each call returns a single structured score; the 18 calls are independent and parallelised per vignette.

I Expert-Annotated Vignette (Persona 14)

To complement the quantitative evaluation with a qualitative one, a licensed clinical psychologist with formal training in the Ehlers & Clark cognitive model of PTSD and a clinical specialism in PTSD case formulation annotated a full-condition vignette span by span, tagging each clinically meaningful passage with the EC component it instantiates. The annotation is reproduced verbatim below in the annotator's colour scheme. The exercise is illustrative, not statistical: it shows what the structural recovery results of §5.1 look like at the level of clinical reading.

Persona (sampled inputs). A 54-year-old female military combat engineer; trauma type: military / route-clearance explosion; PCL-5 = 62. Active components: {Triggers, Negative Appraisals, Memory, Threat, Maladaptive Strategies}. Vignette generated by `gpt-5.4` under the full condition.

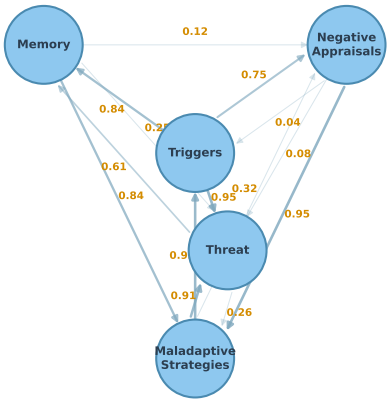


Figure 6: Persona 14 cognitive graph. Active components: {Triggers, Negative Appraisals, Memory, Threat, Maladaptive Strategies}.

Table 8: Self report items per component of persona 14

Component	Items
Triggers	• Fire or flames • A specific temperature on the skin • Shouting
Maladaptive Strategies	• Avoid anything that could be stressful • I numb my feelings • Constantly scan environment for signs of threat
Threat	• My body feels constantly tense and ready to react to danger • Unusual silence makes me feel something dangerous is about to happen • Having difficulty concentrating
Memory	• Something in my surroundings can instantly bring back the full memory without warning • When I have memories of what happened I sometimes hear things in my head that I heard during the frightening event • I cannot tell a clear story of what happened even to myself
Negative Appraisals	• There is something wrong with me as a person • What happened to me was deeply unfair and nothing has been done about it • The world is a dangerous place

Legend. Triggers Memory Negative Appraisals Maladaptive Strategies Sense of current Threat

At 54, Marisol Alvarez had spent most of her adult life in the Army as a combat engineer, and she still talked about deployment in fragments rather than as a story with a beginning and end. The event she circled around happened during a route-clearance mission overseas, when an explosion ignited a vehicle ahead of hers. She remembered the sudden bloom of flames, the searing heat striking the exposed skin of her face and forearms, and men shouting over one another in the confusion. Those moments were preserved in sharp pieces rather than a narrative. She could not put them into a clear sequence even in the privacy of her own mind; what returned instead were flashes and sounds, especially the shouted commands and panic she heard that day.

Back home, those same sensory details became potent in ordinary life. At work, if she passed a training burn pit or caught sight of fire during demolition exercises, her body reacted before she could think. A burst of heat from industrial equipment against her skin or a soldier yelling across the motor pool could make her go rigid and instantly ready for danger. What unnerved her just as much was the silence that sometimes followed routine activity; an unusual quiet in a work area made her feel that something terrible was about to break loose. In those moments she was not only tense but unable to focus on what was in front of her. The trigger could also bring back the event whole and without warning, so that a smell of heat or the sight of flames was enough for her to hear the shouting in her head again as if she were back there.

Those sudden returns of memory shaped what she believed about herself and the world. Each time the scene intruded so forcefully, Marisol took it as proof that something was wrong with her, that other people could move on from war and she could not. She had come to believe the world was fundamentally dangerous, not just in combat but everywhere, and the randomness of what happened felt intolerably unfair to her. Seeing fire or hearing shouting did not simply startle her; it seemed to confirm that danger was always near and that she had been left to carry an injustice no one had put right. Over time she became sensitized to those cues themselves, scanning for heat, noise, and changes in the atmosphere as if her mind were teaching itself where the next threat would come from.

Her beliefs then began to govern how she lived. Because she was certain the world was dangerous and because she felt flawed for reacting so strongly, she tried to avoid anything that might be stressful. She turned down optional field exercises, lingered in administrative tasks when she could, and at home steered away from family cookouts where open flames and loud voices were likely. She worked hard to numb her feelings, speaking flatly when her husband asked what was wrong and trying to shut down

whatever rose up inside her. She also checked her surroundings constantly, in the office, on base, in parking lots, noticing exits, listening for changes in sound, looking for anything that might signal threat. That constant management of herself made her feel even more unlike the person she used to be, reinforcing her conviction that there was something defective in her.

The memories themselves fed this pattern. Because they arrived without warning and brought the shouting back so vividly, Marisol began organizing her days around not being caught off guard. She avoided stressors and kept herself emotionally blunted because she feared that any strong feeling would open the door to the full memory. Yet the more she lived in this narrowed, watchful way, the more tense her body remained. She was on alert so much of the time that concentration became difficult even during routine engineering planning, and once she was already keyed up, the next silence, shout, or wave of heat struck with even more force. Hypervigilance made the memories easier to set off; the memories intensified her sense of present threat.

Over months this tightened into the life she now found herself living. Her constant scanning meant she noticed every flare, every raised voice, every change in temperature, which made those triggers feel omnipresent. Avoiding stress and numbing herself had shrunk her world but had not made it feel safer. Instead, the restrictions seemed to prove that the world really was dangerous and that she really was damaged by what had happened. Her marriage had grown strained by her distance, and at work she was ashamed that she could no longer concentrate as she once had in the very environment where she was expected to lead calmly.

Marisol finally came for help after freezing during a stateside training exercise when flames rose unexpectedly and someone shouted for equipment. She recovered quickly enough that others may not have understood what happened, but she spent the rest of the day convinced she could hear the old voices in her head. What frightened her most was not just the memory itself but the sense that her life had become organized around waiting for it. She said she wanted, more than anything, to be able to stand in her own work environment without feeling that danger was already there, and to understand whether the part of her that felt permanently wrong could still be reached.

What the expert annotation shows. All five EC components are concretely instantiated, including spans the quantitative judge can recognise on its own: explicit *Triggers* (heat, flame, shouting) in paragraph 1; *Memory* characterised as fragmented and present-tense rather than narrative (paragraphs 1–2); catastrophic *Negative Appraisals* about the self and the world (paragraph 3); *Maladaptive Strategies* of avoidance and emotional numbing (paragraphs 4–5); and *Sense of current Threat* as bodily readiness for danger and hypervigilance (paragraphs 2, 5). The annotator also marks the maintenance cycle the Ehlers and Clark (2000) model predicts: *Triggers* reactivate *Memory*, *Memory* feeds *Negative Appraisals*, *Negative Appraisals* drive *Maladaptive Strategies*, and *Strategies* (+ hypervigilance) sustain *Sense of current Threat*, which in turn keeps the system primed for the next *Trigger*. Figure 6 makes the edge structure concrete. The highest-weight edges have explicit textual counterparts. *"a burst of heat... could make her go rigid and instantly ready for danger"* grounds Triggers→Threat (w=0.95); *"the trigger could also bring back the event whole and without warning... as if she were back there"* grounds Triggers→Memory (w=0.84); *"those sudden returns of memory shaped what she believed about herself and the world"* grounds Triggers→Negative Appraisals (w=0.75); *"because she was certain the world was dangerous and because she felt flawed for reacting so strongly"* grounds Negative Appraisals→Maladaptive Strategies (w=0.95); and *"her constant scanning meant she noticed every flare"* closes the maintenance loop via Maladaptive Strategies→Triggers (w=0.93). By contrast, the one edge-span whose source component carries a low ground-truth weight – *looking for anything that might signal threat"* annotated as Negative Appraisals→Threat (w=0.08) – is a borderline reading: the hypervigilance it describes is more directly attributable to the Threat node itself than to an appraisal driving threat. The clinically informed reading converges with the structural-fidelity and edge-recovery results in §5.1.

J Per-Rater EC-Component Recovery on the 330-Vignette Subsample

The per-rater EC-recovery analysis extends the scaled-judge analysis (Appendix K) to all five raters on the 330-vignette evaluation subsample (n = 110 per condition; 11 generation models × 30 personas per model). It answers two questions the body cannot: (a) whether the recovery gradient is judge-specific or replicates across human and LLM raters, and (b) whether human and LLM raters disagree systematically on which com-

ponents are present.

The structure of Table 9 is: rows are EC components, columns are the five raters (R1, R2 are clinical experts; Opus = `claude-opus-4-5`; DSpro = `deepseek-v4-pro`; Flash = `deepseek-v4-flash`), and each rater contributes five metrics (P, R, F1, MCC, AC1). The three blocks (full / no-formulation / zero-shot) are stacked vertically.

Two patterns repeat across raters. First, recovery in the full condition is positive and substantial for every rater on every component: every full-block MCC is above zero, AC1 ranges 0.60–0.92, and the human experts (R1, R2) recover the components on par with or above the LLM judges on most components (e.g., Memory: R1 AC1 $= 0.85$, Opus $= 0.82$; Threat: R1 AC1 $= 0.86$, Opus $= 0.92$). Second, recovery in the zero-shot block collapses for every rater: human-expert AC1 falls into negative territory on Memory (~ -0.05 to $+0.18$), Threat (~ -0.53), and Appraisals (~ -0.12 to $+0.06$), while LLM judges retain a slight positive AC1 on Triggers (a base-rate effect: LLMs almost always say Triggers are present) but otherwise also collapse. The MCC pattern is the same and more interpretable because MCC is invariant to base-rate inflation.

These patterns are useful for two reasons. First, they show that the full-condition recovery is not a property of a single LLM judge: humans and LLMs both find the specified components when they are explicitly cued into the text, and both fail to find them when they are not. The recovery effect is a property of the text, not of the rater. Second, the relative ordering within the full block (e.g., Memory recovery: R1 AC1 $= 0.85 >$ Opus $= 0.82 >$ Flash $= 0.72$) gives a noisy but informative ranking of rater sensitivity that we use to calibrate the scaled-judge analysis on the full 5,000-vignette set.

Caveats: the no-formulation block on the two human experts is mixed (some components recover, others do not), which we attribute to the experts' weighting of explicit-mention versus implicit-context evidence in the rating instructions; LLM raters use the explicit-mention criterion more strictly.

K Scaled-judge Component-Level Recovery

The scaled-judge component recovery analysis tests whether the *specified* active EC components are detectable in the generated text without telling the judge what was specified. Concretely, for each persona we record `active_nodes` $\subseteq$ {Triggers, Negative Appraisals, Memory, Threat, Strategies}, which is the set of EC components that the persona's sampled cognitive graph activates. In the full condition the Crafter is told which components are active and asked to portray them in the text; in the no-formulation condition the Crafter receives the persona's self-report items but not the active-nodes labels (so any component-level signal in the text reflects what bleeds through from the items); in the zero-shot condition the Crafter receives only the demographic prompt, so component-level signal should be at chance.

The scaled judge `deepseek-v4-flash` reads each generated vignette and produces a binary judgement for each of the five components. We treat the specified `active_nodes` as the gold label and the judge's binary output as the prediction, and compute precision, recall, F1, balanced accuracy, MCC, and Gwet's AC1 per (component, condition) cell. The base rates of the gold are not balanced (Memory, Threat, and Strategies are present in 40–60% of personas; Triggers and Negative Appraisals in 70–90%), which is why we report MCC and BAcc alongside F1: F1 is upward-biased on imbalanced data and would mask the recovery collapse on rare-positive components.

Table 10 shows the expected gradient. In full, MCC ranges $+0.30$–$+0.62$ across components; in no-formulation, MCC drops to $+0.16$–$+0.53$; in zero-shot, MCC is at chance for four of five components (-0.01–$+0.02$). The exception in zero-shot is Triggers, which retains AC1 $= 0.57$ even without explicit specification: Triggers is the most narratively familiar PTSD construct and any zero-shot vignette tends to mention them. This matches the body claim that the formulation's largest payoffs are on the less narratively obvious components (Strategies, Appraisals, Memory, Threat) and the smallest on Triggers.

Two limitations of this gold standard. First, `active_nodes` is a binary summary of a graded persona: a persona with weight 0.05 on Memory "has Memory active" by the threshold but may produce only a single sentence of memory content, which a judge can reasonably miss. Second, the judge is itself an LLM with its own systematic biases; we use Gwet's AC1 rather than Cohen's κ here because AC1 is more robust to the prevalence/marginal imbalance typical of EC labels.

Table 9: Recovery of specified EC components on the 330-vignette subsample, per rater and per condition ($n = 110$ per condition). Columns: P = precision, R = recall, F1, MCC, AC1 (Gwet's). The zero-shot block is a negative control: the specified GT was not communicated to the Crafter in that condition, so recovery should be at chance.

	R1					R2					Opus					DSpro					Flash				
EC	P	R	F1	MCC	AC1	P	R	F1	MCC	AC1	P	R	F1	MCC	AC1	P	R	F1	MCC	AC1	P	R	F1	MCC	AC1
Full condition																									
Triggers	.77	.95	.85	.38	.64	.80	.95	.87	.49	.69	.76	1.00	.86	.43	.67	.73	.96	.83	.24	.60	.73	.99	.84	.28	.63
Neg. Appraisals	.85	.90	.87	.29	.70	.84	.89	.86	.23	.68	.84	.97	.90	.35	.77	.83	.91	.87	.22	.70	.84	.95	.89	.32	.76
Memory	.92	.94	.93	.23	.85	.91	.92	.92	.10	.81	1.00	.86	.92	.61	.82	.99	.78	.87	.45	.69	1.00	.79	.88	.52	.72
Threat	.92	.95	.94	.26	.86	.91	.80	.85	.05	.66	.93	.99	.96	.51	.92	.95	.83	.89	.34	.74	.95	.83	.89	.34	.74
Strategies	.79	.96	.86	.46	.68	.79	.96	.86	.46	.68	.73	.99	.84	.28	.63	.74	.99	.85	.35	.65	.73	1.00	.84	.30	.63
No-formulation condition																									
Triggers	.68	.78	.73	−.08	.34	.70	.71	.70	−.01	.29	.72	1.00	.84	.26	.62	.72	.95	.82	.17	.58	.72	.95	.82	.17	.58
Neg. Appraisals	.78	.68	.73	−.08	.34	.76	.58	.66	−.12	.17	.82	.76	.79	.07	.50	.82	.85	.84	.13	.62	.81	.80	.81	.07	.54
Memory	.87	.77	.82	−.17	.58	.86	.71	.78	−.20	.48	1.00	.81	.89	.54	.75	1.00	.70	.82	.43	.57	1.00	.76	.86	.49	.67
Threat	.89	.67	.76	−.04	.44	.90	.46	.60	.00	.05	.94	.89	.91	.29	.80	.93	.64	.75	.11	.41	.90	.71	.79	−.01	.51
Strategies	.62	.64	.63	−.28	.10	.68	.57	.62	−.06	.10	.83	.99	.90	.62	.76	.77	.91	.83	.33	.60	.84	.77	.80	.41	.53
Zero-shot condition (negative control)																									
Triggers	.68	.66	.67	−.06	.21	.70	.69	.69	−.01	.26	.69	.91	.79	−.05	.50	.72	.91	.81	.13	.54	.69	.95	.80	−.13	.53
Neg. Appraisals	.82	.36	.50	.04	−.12	.86	.48	.61	.13	.09	.87	.46	.60	.15	.06	.85	.47	.60	.12	.07	.80	.46	.58	.00	.01
Memory	.90	.53	.66	−.01	.18	.90	.53	.66	−.01	.18	.92	.43	.59	.04	.01	.93	.49	.65	.08	.14	.90	.47	.61	.01	.07
Threat	.90	.17	.29	−.01	−.52	.82	.18	.30	−.14	−.53	.90	.69	.78	−.03	.47	.88	.60	.71	−.08	.31	.86	.61	.71	−.19	.31
Strategies	.74	.38	.50	.07	−.05	.72	.53	.61	.04	.10	.73	.60	.66	.08	.19	.71	.71	.71	.05	.31	.67	.47	.55	−.07	−.04

Table 10: Recovery of specified EC components by the scaled judge `deepseek-v4-flash`, per condition ($n = 5{,}000$ per condition; 10 generation models × 500 personas, `gpt-4o-mini` excluded). The specified "ground truth" is the persona's `active_nodes` set, which is explicitly conveyed to the Crafter only in the full condition; the no-formulation and zero-shot columns measure how much the persona's underlying cognitive profile still bleeds into the vignette without explicit specification, and serve as a negative control. **P**: precision; **R**: recall; **F1**: harmonic mean; **BAcc**: balanced accuracy; **MCC**: Matthews correlation; **AC1**: Gwet's AC1.

	Full						No-formulation						Zero-shot					
EC	P	R	F1	BAcc	MCC	AC1	P	R	F1	BAcc	MCC	AC1	P	R	F1	BAcc	MCC	AC1
Triggers	.76	.99	.86	.61	.39	.68	.75	.97	.85	.58	.28	.64	.71	.95	.81	.50	−.01	.57
Neg. Appraisals	.81	.99	.89	.73	.59	.75	.79	.97	.87	.69	.49	.70	.71	.53	.60	.51	.01	.07
Memory	.93	.80	.86	.83	.62	.66	.90	.78	.83	.78	.53	.60	.70	.42	.53	.50	−.00	−.05
Threat	.77	.83	.80	.64	.30	.51	.74	.74	.74	.58	.16	.37	.69	.67	.68	.51	.02	.24
Strategies	.74	.99	.85	.62	.41	.64	.75	.81	.78	.62	.25	.47	.67	.48	.56	.49	−.01	.01

L Per-Edge Recovery in the Full Condition

The body's edge-recovery analysis (§5.1, Table 1) aggregates the external probe over all 20 directed Ehlers & Clark edges. Table 11 disaggregates it: for each of the 20 edges $A \to B$, we report precision, recall, F1, MCC, and AC1 on the $n = 110$ full-condition vignettes, treating the persona's specified edge weight $w_{A \to B} > 0$ as the gold positive label and the external probe's binary judgement as the prediction. The table is sorted by descending MCC.

Every one of the 20 directed edges recovers above chance (MCC > 0): there is no edge that the formulation specification fails to imprint on the text. The strongest recovery is concentrated on the four edges that the Ehlers & Clark theory explicitly identifies as maintaining the disorder: Strategies → Threat (MCC $+0.66$), Triggers → Negative Appraisals ($+0.63$), Threat → Memory ($+0.57$), and Threat → Negative Appraisals ($+0.56$). These are the edges that close the maintenance cycle in the cognitive model, and the cycle is the part of the theory that case-formulation training emphasises; it is unsurprising that a model specified with the graph and the weights chooses to instantiate these edges most explicitly.

The weakest recovery is on edges with high judge-marginal rates: Triggers → Threat (MCC $+0.34$, recall 1.00 but precision 0.47) and Threat → Strategies (MCC $+0.25$, precision 0.56). In both cases the probe tends to read the connection from any narrative that mentions both endpoints, regardless of whether the edge was specified active: the precision drops because the judge cannot easily distinguish "A and B both appear" from "A leads to B." This is a property of the probe and the text, not of the graph: short clinical vignettes provide limited contextual evidence for direction even when the underlying generation conditioned on a directed edge. The unit-level precision is therefore a lower bound on the true edge realisation rate.

A separate caveat: the gold standard treats edge

presence as binary ($w > 0$ vs $w = 0$) and ignores weight magnitude. The body's right-block analysis in Table 1 shows that within specified-active edges, the judge's recovery probability is roughly flat across weight bands (0.74–0.84), consistent with the formulation acting as a near-binary on/off switch in this configuration rather than as a continuous prominence dial.

Table 11: Per-directed-edge external-probe metrics, full condition only ($n = 110$ vignettes per edge). P = precision, R = recall, F1, MCC = Matthews correlation, AC1 = Gwet's. Edges sorted by descending MCC. "Neg.App." = Negative Appraisals; "Strat." = Maladaptive Strategies.

Directed edge $A \to B$	**P**	**R**	**F1**	**MCC**	**AC1**
Strat. → Threat	0.86	0.86	0.86	+0.66	+0.69
Triggers → Neg.App.	0.87	0.82	0.84	+0.63	+0.65
Threat → Memory	0.91	0.58	0.71	+0.57	+0.54
Threat → Neg.App.	0.89	0.71	0.79	+0.56	+0.55
Triggers → Memory	0.66	0.94	0.78	+0.51	+0.48
Neg.App. → Triggers	0.97	0.48	0.65	+0.49	+0.37
Threat → Triggers	0.88	0.51	0.64	+0.48	+0.46
Strat. → Memory	0.72	0.76	0.74	+0.47	+0.47
Memory → Neg.App.	0.80	0.92	0.85	+0.44	+0.65
Neg.App. → Memory	0.69	0.57	0.62	+0.42	+0.49
Neg.App. → Threat	0.57	0.82	0.67	+0.41	+0.36
Memory → Triggers	0.91	0.34	0.50	+0.39	+0.37
Triggers → Strat.	0.42	0.94	0.58	+0.38	+0.20
Memory → Strat.	0.66	0.76	0.71	+0.37	+0.37
Strat. → Triggers	0.44	0.55	0.49	+0.35	+0.65
Strat. → Neg.App.	0.44	0.82	0.57	+0.35	+0.29
Triggers → Threat	0.47	1.00	0.64	+0.34	+0.15
Memory → Threat	0.70	0.82	0.76	+0.32	+0.42
Neg.App. → Strat.	0.58	0.93	0.71	+0.32	+0.32
Threat → Strat.	0.56	0.89	0.69	+0.25	+0.26

M Human Validation of Directed-Edge Recovery

This appendix documents the human edge-annotation study summarised in §5.2. Its purpose is to check the body's central structural claim, that the specified *directed* links are preserved in the text, against human judgement rather than against an LLM probe alone. The experts in §4.2 score EC components; this study scores the directed edges themselves, at a scale comparable to the automated probe.

Sample. We drew 50 personas and took each under all three conditions, giving 150 vignettes. The sample is balanced across the ten retained generation models (5 vignettes per model per condition; `gpt-4o-mini` is excluded here as it is throughout the paper), and every persona appears under all three conditions, so condition contrasts are within-persona.

Annotation schema. For each vignette the rater is shown the text alone and asked, for each of the 20 directed Ehlers & Clark pairs $A \to B$, (i) whether the vignette expresses that directed causal link, as a binary judgement, and (ii) how prominent the link is in the narrative, on a graded 0–1 scale. This is 20 judgements per vignette and 3,000 per rater across the sample. Raters are blind both to the generating condition and to the persona's specified cognitive graph, so the specified weights cannot be read back from the task itself.

Raters. Two clinical raters (H1, H2) completed the full sample independently. The same 150 vignettes were then scored on the identical schema by two of the LLM judges already used in the paper, `deepseek-v4-flash` and `deepseek-v4-pro` (§4.2), under the same blinding. This yields a four-rater panel of two humans and two models, so that human-to-human agreement can be compared directly against human-to-model agreement on identical items.

Agreement. Table 12 reports all six pairwise agreements, on the presence scale with Gwet's AC1 (the statistic used throughout the paper, Appendix P) and on the graded scale with ICC$(2, 1)$. Every pair agrees between 0.63 and 0.75 on presence. Human-to-LLM agreement (0.647 to 0.699, mean 0.673) is at least as high as the agreement between the two human raters (0.631), and the same ordering holds on the graded scale (human-to-LLM ICC 0.671 to 0.717; human-to-human 0.671). By this measure the automated probe is not distinguishable from an additional human annotator.

Table 12: Pairwise agreement between the four raters on the 150-vignette edge-annotation sample, pooled over conditions ($n = 3{,}000$ directed-edge judgements per pair). AC1 is computed on the binary presence judgement, ICC$(2, 1)$ on the graded 0–1 prominence rating.

Rater pair	**AC1 (present)**	**ICC (strength)**
H1 vs H2	0.631	0.671
H1 vs FLASH	0.680	0.671
H1 vs PRO	0.699	0.703
H2 vs FLASH	0.647	0.701
H2 vs PRO	0.668	0.717
FLASH vs PRO	0.748	0.774

Recovery of the specified graph. Table 13 gives the body's Table 2 at three decimals. Treating the persona's specified active edges ($w > 0$) as gold, all four raters recover the graph under full

(MCC $+0.318$ to $+0.408$) and are at chance under both ablations (MCC -0.023 to $+0.056$). The AUC of the specified weight predicting detection is 0.645 to 0.689 under full against 0.487 to 0.535 under the ablations, and the graded strength ratings correlate with the specified weight under full (Spearman $+0.341$ to $+0.432$) but not under either ablation (-0.015 to $+0.084$). The human raters therefore reproduce, by hand and without sight of the graph, the recovery gradient the probe reports in §5.1 (MCC $+0.41$, AUC 0.70).

Table 13: Recovery of the specified cognitive graph by each of the four raters on the 150-vignette sample, at three decimals. Gold is the persona's specified active edge set ($w > 0$).

	Full	No-form.	Zero-shot
MCC vs specified graph			
H1	$+0.408$	$+0.056$	$+0.044$
H2	$+0.318$	$+0.043$	$+0.007$
FLASH	$+0.319$	$+0.035$	-0.023
PRO	$+0.379$	-0.012	-0.020
AUC (specified weight → detection)			
H1	0.689	0.535	0.530
H2	0.648	0.525	0.508
FLASH	0.645	0.517	0.491
PRO	0.672	0.499	0.487
Spearman (rated strength, specified weight)			
H1	$+0.432$	$+0.084$	$+0.051$
H2	$+0.364$	$+0.062$	$+0.016$
FLASH	$+0.341$	$+0.028$	-0.012
PRO	$+0.393$	$+0.006$	-0.015

Release. The per-judgement annotations for all four raters, together with the specified weights and the analysis script that reproduces both tables, are released with the rest of the materials (§7).

N Sensitivity Analyses

The body's central claims rest on three analytical choices: the edge-presence threshold ($w > 0$) for the edge-recovery analysis (§5.1), the choice of `deepseek-v4-flash` as the scaled judge for the component-recovery analysis (§5.1), and the choice of MPNet (`all-mpnet-base-v2`) as the encoder for the cross-model convergence analysis (§5.3). Each of these is defensible but not unique, so we run a sensitivity analysis on each: we vary the analytical choice and re-compute the headline statistic to verify that the direction of the formulation effect does not depend on the specific operationalisation.

N.1 Edge-weight threshold

The body treats a directed edge as "specified active" whenever the persona's sampled edge weight $w_{A\to B}$ is strictly positive, and reports MCC $= +0.41$ for the external judge against this binary gold in the full condition. A reviewer can reasonably ask whether the recovery is being driven by edges with vanishingly small weights that the Crafter was unlikely to instantiate explicitly. We re-compute the classification metrics treating the gold positive class as $w > t$ for $t \in \{0, 0.1, 0.2, 0.3, 0.5\}$. Higher thresholds reduce the positive class and ask a harder question (does the judge recover the *prominently*-weighted edges?).

Table 14 reports the result, and Figures 7 and 8 plot recovery MCC and F1 across the thresholds. In the full condition, MCC declines monotonically from $+0.41$ at $t = 0$ to $+0.22$ at $t = 0.5$, but remains positive at every threshold. In the no-formulation and zero-shot conditions, MCC remains at chance (-0.03 to $+0.04$) across all thresholds. The full-vs-zero-shot MCC gap is $+0.41$ at the body's threshold and $+0.22$ at the strictest threshold tested, and the direction of the formulation effect is preserved at every choice. The attenuation at higher thresholds is informative rather than concerning: it shows that the formulation's edge-imprinting is more reliable for weakly-specified edges than for the strongest-weighted ones, consistent with the body's finding that the graph behaves as a near-binary on/off switch rather than a continuous prominence dial (§5.1, Table 1).

Table 14: Edge-recovery classification metrics on the 330-vignette evaluation subsample at multiple edge-weight thresholds for the positive gold class ($w > t$). *prev* is the base rate of the positive gold class at that threshold. MCC stays positive in full at every threshold; the no-formulation and zero-shot rows remain at chance throughout.

Condition	t	prev	F1	MCC	AC1
Full	0.0	0.49	0.71	**+0.41**	**+0.41**
Full	0.1	0.43	0.66	+0.35	+0.33
Full	0.2	0.40	0.63	+0.33	+0.31
Full	0.3	0.35	0.58	+0.29	+0.26
Full	0.5	0.25	0.46	+0.22	+0.18
No-formulation	0.0	0.49	0.41	+0.03	+0.07
No-formulation	0.5	0.25	0.30	+0.02	+0.32
Zero-shot	0.0	0.49	0.37	+0.01	+0.06
Zero-shot	0.5	0.25	0.27	−0.00	+0.34

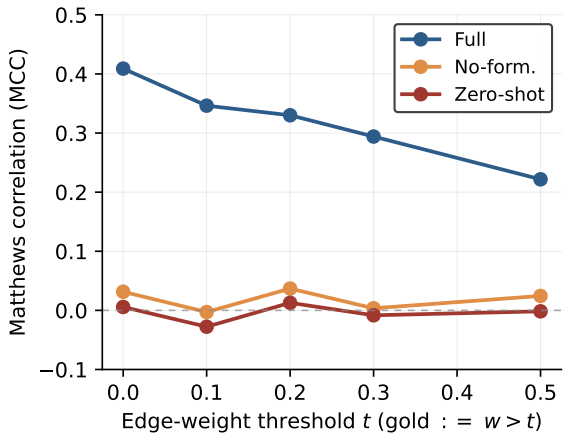


Figure 7: Edge-recovery MCC across edge-weight thresholds $t \in \{0, 0.1, 0.2, 0.3, 0.5\}$, by condition. Full-condition MCC remains positive at every threshold (declining monotonically as the gold positive class shrinks); no-formulation and zero-shot remain at chance throughout.

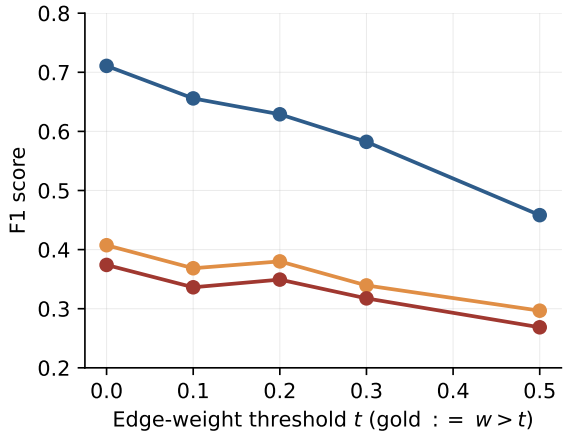


Figure 8: Companion to Figure 7: edge-recovery F1 across the same thresholds. The full-vs-zero-shot gap is preserved at every operationalisation of edge presence.

N.2 Judge identity for the component-recovery analysis

The body component-recovery analysis (§5.1, Appendix K) uses `deepseek-v4-flash` as the scaled judge for the 15,000-vignette generation set. To verify that the recovery is not Flash-specific, we compute the same statistic restricted to the 330-vignette evaluation subsample (where Opus and DSpro are also available) and report the per-judge MCC against the specified `active_nodes` gold for each of the five Ehlers & Clark components in each condition (Table 15).

Two patterns emerge, both visible in Figure 9. First, every full-condition cell is positive across all three judges and all five components: the recovery direction is robust to judge identity. Mean MCC across components in the full condition is $+0.44$ for Opus, $+0.35$ for Flash, and $+0.30$ for DSpro; the absolute level varies (Opus is the most sensitive, DSpro the most conservative) but the rank-order across components is consistent (Memory and Threat highest, Strategies and Triggers lowest). Second, every zero-shot cell is at or near chance across all three judges (mean MCC $+0.04$ for Opus, -0.08 for Flash, $+0.06$ for DSpro), so the full-vs-zero-shot MCC gap is large for every judge ($+0.40$, $+0.43$, $+0.24$ respectively). The body's MCC $+0.41$ on the full 5,000-vignette set is intermediate between Opus and Flash on the 330-vignette subsample and consistent with the recovery being a property of the generated text rather than of the specific judge model.

Table 15: Per-judge MCC for EC-component recovery on the 330-vignette evaluation subsample ($n = 110$ per condition). Gold is the persona's specified `active_nodes`; predictions are each judge's binary EC ratings. Every full cell is positive across all three judges and all five components; every zero-shot cell is at or near chance.

		Full			**No-formulation**			**Zero-shot**		
Component		Opus	DSpro	Flash	Opus	DSpro	Flash	Opus	DSpro	Flash
Triggers	MCC	+0.43	+0.23	+0.28	+0.26	+0.17	+0.17	−0.05	+0.13	−0.13
Neg. Appraisals	MCC	+0.35	+0.20	+0.32	+0.07	+0.14	+0.07	+0.15	+0.12	+0.00
Memory	MCC	+0.62	+0.46	+0.52	+0.54	+0.44	+0.49	+0.04	+0.08	+0.01
Threat	MCC	+0.51	+0.33	+0.34	+0.29	+0.11	−0.01	−0.03	−0.08	−0.19
Strategies	MCC	+0.28	+0.31	+0.30	+0.62	+0.35	+0.41	+0.08	+0.05	−0.07
Mean	MCC	**+0.44**	**+0.30**	**+0.35**	+0.36	+0.24	+0.23	+0.04	+0.06	−0.08

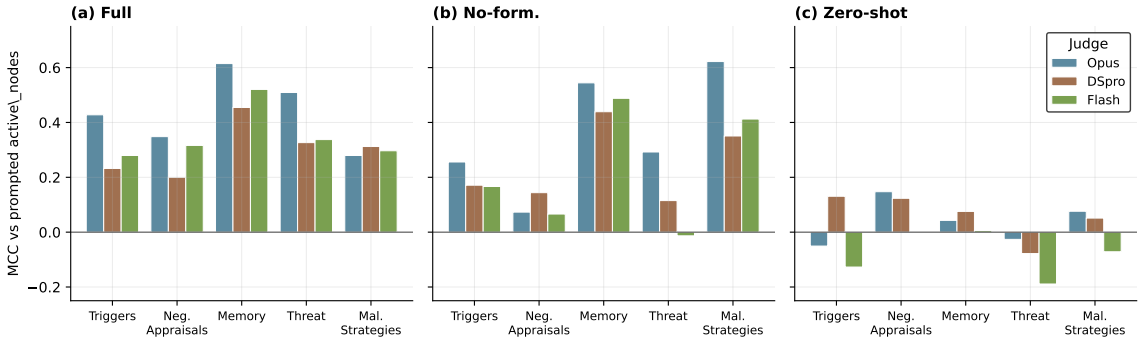


Figure 9: Per-judge MCC for Ehlers & Clark component recovery on the 330-vignette evaluation subsample, by condition and component. (a) Full: every cell positive across all three judges and five components. (b) No-formulation: mostly positive, attenuated. (c) Zero-shot: at or near chance for every judge and component. The recovery direction holds for all three judges; Opus is the most sensitive and DSpro the most conservative, but the rank-order across components is consistent.

N.3 Embedding model for the cross-model convergence analysis

The body's cross-model convergence analysis (§5.3, Appendix O) treats each vignette as a point in MPNet (`all-mpnet-base-v2`) sentence-embedding space, 768-dimensional. The reported within-persona cross-model cosine-similarity statistic ($\mu_{\text{full}} = 0.792$, $\mu_{\text{no-form}} = 0.767$, $\mu_{\text{zero-shot}} = 0.737$; full-vs-zero-shot $d_z = +1.52$) could in principle be a property of MPNet's training rather than of the text: a different encoder might compress the persona-specific content differently and dissolve the cross-model convergence pattern. We test this by re-embedding the same 15,000 vignettes (10 retained models) with BGE-base-en-v1.5 (Xiao et al., 2024), a BERT-backbone sentence encoder trained on a different curated mix of contrastive pairs with a different objective formulation, and recomputing the same statistic.

Table 16 reports the result. The conditional pattern replicates on every quantity: the per-condition means are higher in BGE space (the BGE encoder places all vignettes closer in cosine space in our data) but the *ordering* is identical (full > no-formulation > zero-shot), the monotonic gradient is preserved, and the standardised effect sizes for the paired contrasts are comparable in direction and magnitude to those reported in the body, with one redistribution. The full-vs-zero-shot Cohen's $d_z = +1.61$ under BGE is slightly higher than the $+1.52$ reported under MPNet; the full-vs-no-formulation contrast is $+0.91$ under BGE (vs $+0.83$ MPNet); the no-formulation-vs-zero-shot contrast is $+0.70$ under BGE (vs $+0.86$ MPNet), so BGE redistributes the gap toward the full-vs-no-formulation step. All three pairwise contrasts are significant at p below machine precision ($n = 500$ personas). The convergence claim is therefore not an artefact of MPNet's training: a structurally distinct encoder recovers the same conditional gradient.

The higher absolute similarity values under BGE reflect a property of how the encoder distributes vignettes in cosine space and not a stronger formulation effect; the interpretable quantities are therefore the d_z effect sizes, not the raw μ values. Identical inputs and identical similarity arithmetic make the two passes directly comparable.

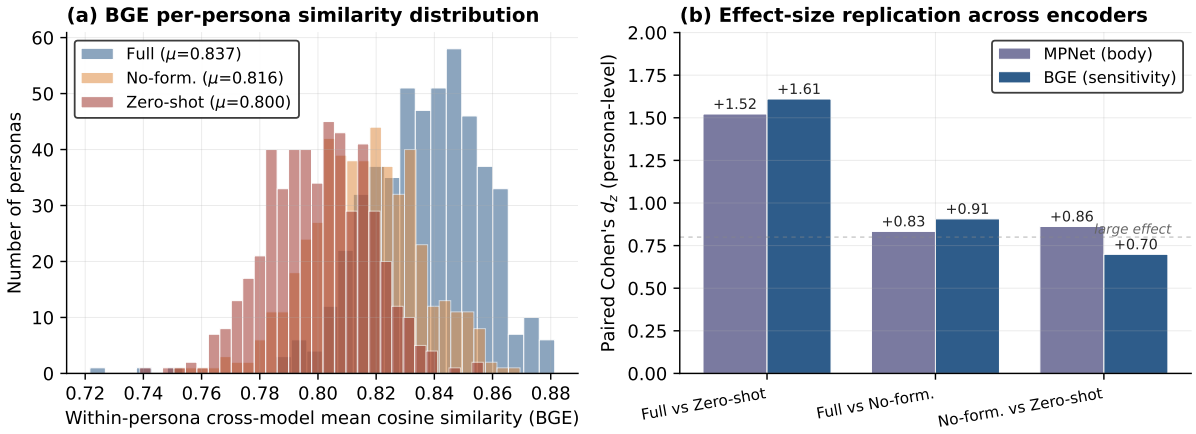


Figure 10: Encoder sensitivity for the cross-model convergence claim. (a) Distribution of within-persona cross-model mean cosine similarity in BGE space, by condition ($n = 500$ personas per condition); the full > no-formulation > zero-shot ordering is preserved and the distributions are well-separated. (b) Paired Cohen's d_z for the three condition contrasts under MPNet (body) versus BGE (sensitivity). All six bars are above the conventional "large effect" threshold of 0.8; the full-vs-zero-shot contrast is larger under BGE (+1.61) than under MPNet (+1.52).

Table 16: Within-persona cross-model cosine similarity under two encoders. The body's MPNet baseline is reproduced for comparison; the BGE row is computed by re-embedding the same 15,000 vignettes (10 retained models) with `BAAI/bge-base-en-v1.5` and applying the identical pipeline. μ: mean per-persona similarity. d_z: paired Cohen's d (persona as the pairing unit, $n = 500$). The conditional ordering and effect sizes are preserved across encoders.

	Per-condition μ			Paired d_z (vs zero-shot or vs no-form.)		
Encoder	Full	No-form.	Zero-shot	Full–Zs	No-form.–Zs	Full–No-form.
MPNet	0.792	0.767	0.737	+1.52	+0.86	+0.83
BGE	0.837	0.816	0.800	**+1.61**	+0.70	+0.91

N.4 Vignette length as a potential confound

Vignette length is not a confound for the formulation effect. The generated vignettes are of comparable word count across conditions, and the formulation's structurally-defining effects survive at matched length with large d_z. We document the evidence below.

Vignette length vs API tokens. Table 7 reports *prompt* + *completion* tokens for the API call, not vignette length. The full-condition prompt includes the entire cognitive-graph specification ($\sim$ 10,000 extra tokens of structural input), which contributes to the prompt-inclusive count but not to the generated text. The quantity relevant to length effects on rated quality is the vignette's word count, which we report here.

Vignette lengths are comparable across conditions. Across the 10 retained generation models (Figure 11a, b), mean vignette word count is 753 (full), 750 (no-formulation), and 682 (zero-shot): a $\sim$ 10% grand-mean difference, not the $\sim 5\times$ ratio inferable from prompt-inclusive tokens. Per-model full/zero-shot ratios cluster on the $y = x$ identity line: `gpt-5.4` 1.00×, `gemini-2.5-flash` 1.01×, `qwen2.5-32b` 0.94×, `deepseek-chat` 1.07×, `claude-sonnet-4-6` 1.14×. The maximum is `claude-haiku-4-5` at 1.53×, a single-model outlier; two of the ten models produce *shorter* full vignettes than zero-shot. The condition does not consistently move length in either direction at the per-model level.

Effect sizes at matched length. Within each of the 500 personas, we identify the (full, zero-shot) vignette pair with the smallest word-count difference: median 16 words across 4,935 matched pairs. At this near-identical length, the formulation's effects on the structurally-informative items remain large: Maladaptive Strategies $d_z = +0.72$, Negative Appraisals +0.53, Narrative quality (g2) +0.57 (Figure 11c). Within the three CVI items, the pattern splits: Importance retains a positive contrast at matched length ($d_z = +0.31$), while Clarity (-0.13) and Relevance (-0.08) remain in the same direction as their all-pairs estimates (-0.26 and -0.03 respectively) – the LLM-judge preference for unconstrained zero-shot prose on these two items is not a length artefact, and the experts' counter-direction on Clarity ($d_z = +0.81$) is therefore also not length-driven. Triggers is non-significant in both estimates (saturated at ~ 0.93). Of the items tested, only Sense of current Threat goes from sizable to negligible at matched length (d_z falls from +0.27 to +0.03).

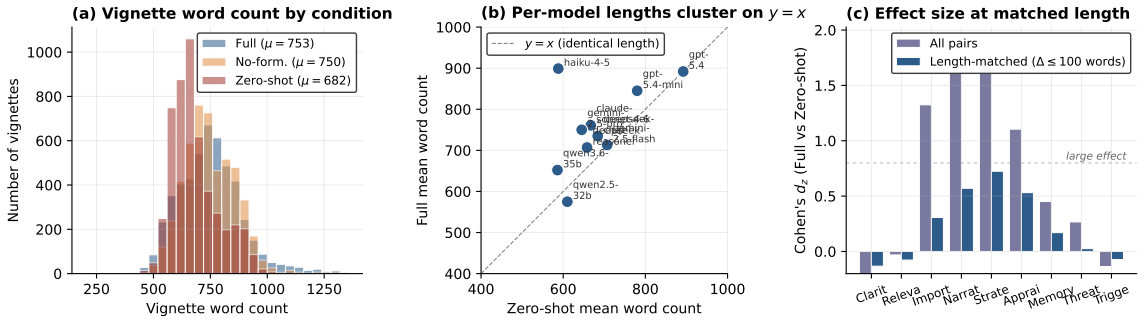


Figure 11: Length is not a confound for the formulation effect. (a) Per-condition distribution of vignette word counts on the 15,000-vignette generation set (10 retained models): full $\mu = 753$, no-formulation $\mu = 750$, zero-shot $\mu = 682$, all overlapping. (b) Per-model means scatter close to the $y = x$ identity line: most models produce full and zero-shot vignettes within $\pm 15\%$ of identical length (`gpt-5.4` $1.00\times$, `gemini-2.5-flash` $1.01\times$, `qwen2.5-32b` $0.94\times$). (c) Cohen's d_z for full-vs-zero-shot per item: "All pairs" uses persona-level means across the full sample; "Length-matched" restricts to within-persona full/zero-shot pairs with ≤ 100 word difference (median 16 words). The structurally-informative EC items (Strategies, Appraisals) retain large effect sizes at matched length; *Threat* collapses under length-matching.

OLS with length covariate. Regressing each item on `condition` + $\log$(`word_count`) on the 15,000-vignette set (Table 17), the formulation coefficient on Strategies and Appraisals retains 92.4% and 93.7% of its magnitude with length controlled: the effect is conditional on the cognitive-graph specification, not on word count. Within CVI, Importance retains 73%, Clarity and Relevance are below detection in both specifications ($|b_{\text{full}}| < 0.03$). Narrative retains 83%; Memory 68%.

Mediation, not confounding. For Threat the formulation coefficient does drop substantially when length is in the model (65% absorbed). This is mediation through the formulation, not confounding by an external variable. When the cognitive graph specifies that sense of current threat must be portrayed, the model writes a richer threat description, which is both longer and more clinically vivid; the two move together because the formulation drives both. Truncating the vignette to remove the length would also remove the threat content, since the content *is* the additional words. The same logic applies to the partial mediation on Importance and Narrative.

Bottom line. Vignettes are of comparable word count across conditions ($\sim 10\%$ grand-mean difference; per-model ratios on the identity line; two of ten models shorter in full than in zero-shot). The structurally-defining EC effects on Strategies and Appraisals survive length-matched pairing with large d_z and retain $> 90\%$ of their magnitude under length control. Within CVI, Importance retains 73%, and the LLM-judge negative effects on Clarity and Relevance are unchanged at matched length

Table 17: OLS coefficient for the "full" indicator on each item, with and without $\log$(word_count) as a covariate. Zero-shot is the reference. **% retained**: fraction of the formulation effect that survives controlling for length, i.e. the part that is *not* explained by word count. The structurally-informative EC items (Strategies, Appraisals) retain $> 90\%$; ordinal items 73–83%; only Threat is largely length-mediated.

Item	No length	With $\log w$	% retained
ec_strategies	+0.429	+0.396	**92.4%**
ec_appraisals	+0.332	+0.311	**93.7%**
narrative (g2)	+0.411	+0.340	82.9%
importance	+0.281	+0.206	73.2%
ec_memory	+0.176	+0.119	67.6%
ec_threat	+0.073	+0.025	34.9%
clarity	−0.015	−0.024	—
relevance	−0.002	−0.023	—

– so the judge-stylistic finding flagged in §5 is not length-driven either. Narrative retains 83%; Memory 68%. Only Sense of current Threat is largely length-mediated, and this is mediation through the formulation (writing about threat is writing more about threat), not confounding by an external variable. The structural findings the paper rests on, edge-recovery MCC and EC presence on Strategies and Appraisals, are not reducible to the generator writing more tokens.

N.5 Cluster geometry: cognitive formulation vs trauma type

The body claim that the formulation breaks trauma-type stereotyping (§5.3, Figure 17) is established qualitatively from a t-SNE visualisation. We complement it here with a quantitative cluster-cohesion measure on the BGE embeddings: for each condition we measure how tightly vignettes cluster by (i) trauma type ($n_g = 29$ categories) and by (ii) the

persona's specified active-node signature ($n_g = 32$ unique combinations across the 500 personas). Each measurement is the difference between mean within-group cosine similarity and mean between-group cosine similarity (within-between gap); a larger gap means the labelling explains more of the embedding's structure. We bootstrap the gap from $B = 20$ subsamples of $n = 2000$ vignettes per condition to obtain standard deviations.

Table 18 reports the result, and Figures 12 and 13 show the per-condition cohesion and the composite ratio. The pattern is what the body claim predicts but had not previously demonstrated quantitatively: zero-shot vignettes cluster $\sim 30\%$ more tightly by trauma type than full-condition vignettes do (gap 0.030 vs 0.023), while full-condition vignettes cluster $\sim 5\times$ more tightly by cognitive formulation signature than zero-shot vignettes do (0.011 vs 0.002, the latter indistinguishable from zero given the bootstrap standard deviation). The composite ratio (formulation-gap divided by trauma-gap) captures the structural shift in a single number: in zero-shot the embedding picks up on trauma-type templates roughly fourteen times more than on cognitive structure (0.07); in full the embedding gives the two roughly equal weight (0.48); no-formulation is intermediate (0.22). The cognitive-graph specification actively redirects the embedding from coarse trauma-category clustering toward person-specific cognitive-signature clustering, which is exactly the structural shift the framework was designed to produce.

Table 18: Cluster cohesion of vignette embeddings (BGE-base) by trauma type and by cognitive formulation signature, per condition. Each entry is mean within-group − between-group cosine similarity, bootstrapped on $B = 20$ subsamples of $n = 2000$ vignettes; SDs in parentheses. **Ratio**: formulation gap divided by trauma gap. Higher ratio = embedding is shaped by cognitive structure rather than trauma category. The shift from zero-shot to full is a factor of $\approx 7\times$ on this ratio.

Condition	Trauma gap	Formulation gap	Ratio
Full	0.023 (0.000)	**0.011 (0.001)**	**0.48**
No-formulation	0.023 (0.001)	0.005 (0.001)	0.22
Zero-shot	**0.030 (0.001)**	0.002 (0.001)	0.07

The quantitative cluster-cohesion result is supported visually by per-cell t-SNE projections of the BGE embeddings, structured to reveal two distinct angles on the conditional pattern. Figure 14 reproduces the body's trauma-stereotyping figure (Figure 17) on the same four representative generation models using the alternative encoder; the conditional pattern replicates immediately (zero-shot vignettes form tight monochromatic clusters that disperse under full), confirming that the body's qualitative finding is not specific to MPNet's training.

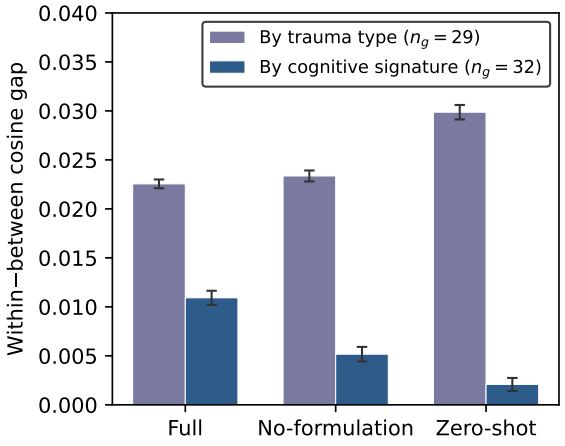


Figure 12: Cluster cohesion of vignette embeddings by label type, per condition (BGE-base, 10 retained models). Within-between cosine-similarity gap with trauma type (light) vs cognitive-signature (dark) as the cluster label. Trauma-type clustering is highest in zero-shot (template fallback); cognitive-signature clustering is highest in full (the graph imprints on the embedding).

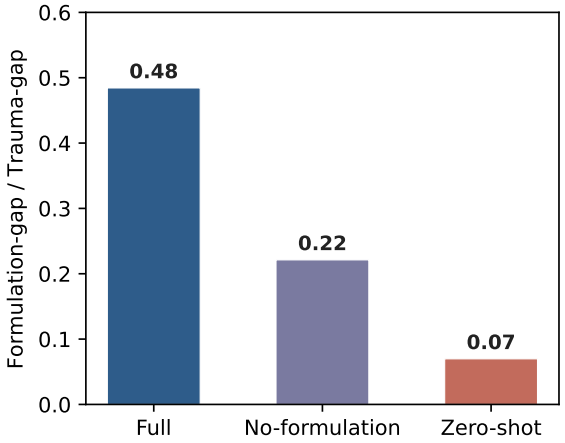


Figure 13: Companion to Figure 12: composite cluster-cohesion ratio (formulation gap / trauma gap) per condition. The ratio rises from 0.07 in zero-shot to 0.48 in full, a $\sim 7\times$ shift in what the embedding picks up on as the dominant grouping signal.

Figures 15 and 16 go further: they regroup the eleven generation models into three capability tiers (**Large**: `gpt-5.4`, `claude-sonnet-4-6`, `deepseek-reasoner`, `gemini-2.5-pro`; **Mid**: `gpt-5.4-mini`, `claude-haiku-4-5`, `deepseek-chat`, `gemini-2.5-flash`; **Small**: `gpt-4o-mini`, `qwen2.5-32b`, `qwen3.6-35b`) and present the same t-SNE projection under both encoders. The tier

breakdown surfaces an interaction the body figure could not show: the breaking of the trauma-type templates is most dramatic in the large-capability tier, where full-condition clusters are nearly invisible; the mid tier shows partial breaking; and the small tier retains visible monochromatic clusters even under full. The pattern is identical under both encoders, ruling out an encoder artefact. The interpretation is that the formulation specification and the generator's underlying capacity to act on it interact: a model has to be capable enough to instantiate the cognitive graph in narrative form, otherwise it falls back on trauma-type templates regardless of what the input asks.

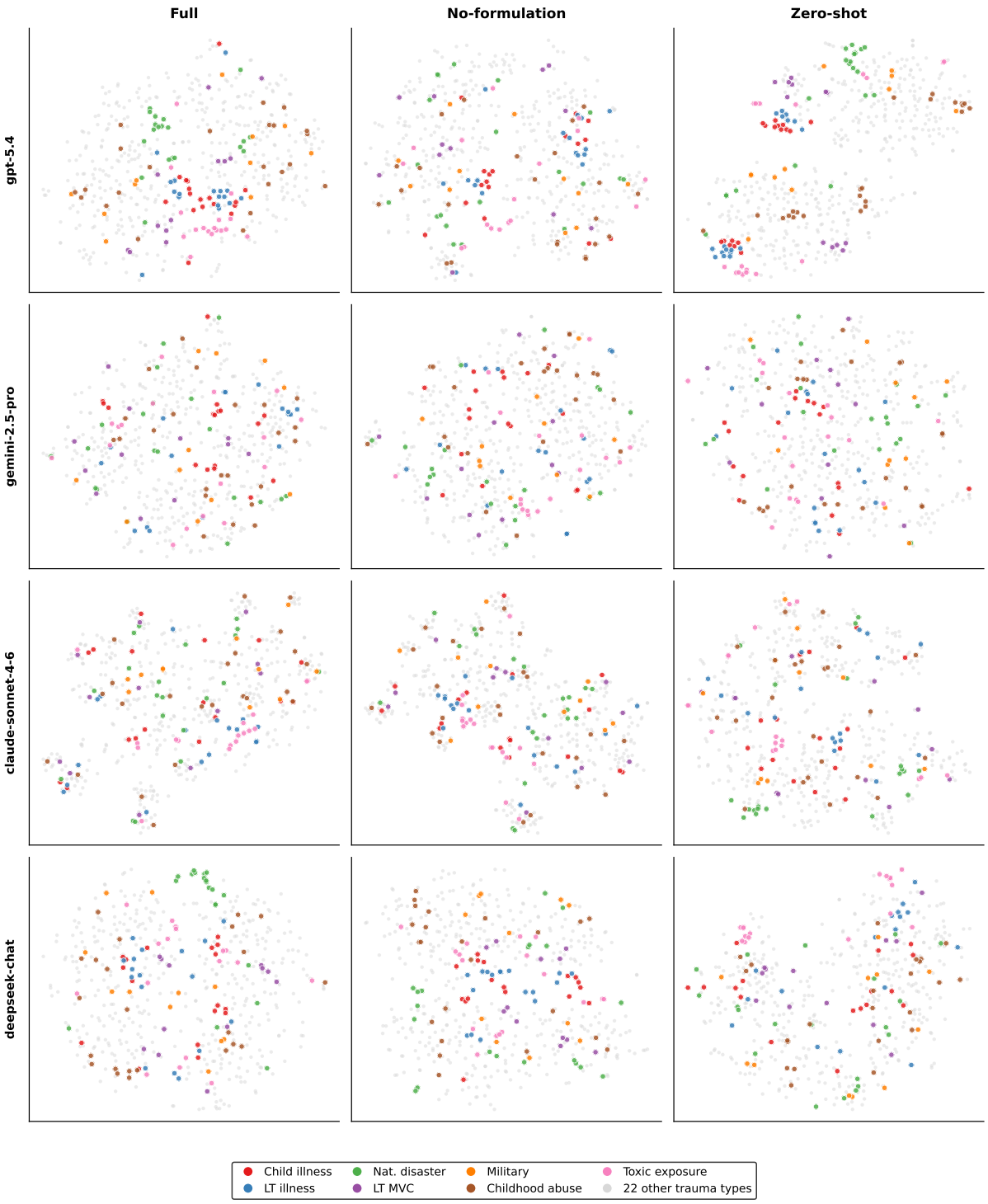


Figure 14: BGE-base t-SNE replication of the body's trauma-stereotyping figure (Figure 17), on the same four representative generation models. Per-cell t-SNE on 500 vignettes; the seven trauma types with the tightest zero-shot intra-trauma similarity are coloured, the other 22 in light grey. Tight monochromatic clusters are most striking in the `gpt-5.4` row (top right); the same coloured vignettes are dispersed under *Full* and *No-formulation*. The qualitative pattern of the body figure replicates under a structurally distinct encoder.



Figure 15: MPNet t-SNE grid, regrouped by model capability tier. Per-cell t-SNE on up to 1,000 vignettes per (tier, condition). Two patterns: trauma-type clustering breaks across all tiers as the prompt becomes more structured (right $\rightarrow$ left), and the breaking is more complete in higher-capability tiers. This is the same body encoder under a tier-level grouping, showing the formulation $\times$ capacity interaction directly.

O Trauma-Type Stereotyping in Latent Space

Trauma-type stereotyping is a specific failure mode of LLM-generated clinical text: when a model has nothing to anchor on but a diagnostic label and a trauma category, it converges on a small set of category templates ("the combat-veteran vignette," "the sexual-assault-survivor vignette"), and individual persona content is lost in the average. The body claim (§5.3) is that the cognitive formulation breaks this fallback. Figure 17 visualises the underlying evidence.

The procedure: (i) embed all 500 personas $\times$ 4 representative models $\times$ 3 conditions = 6,000 vignettes with MPNet; (ii) within each (model, condition) cell, compute intra-trauma-type mean cosine similarity for every one of the 29 trauma categories; (iii) within each cell, project the 500 vignettes to two dimensions using t-SNE (van der Maaten and Hinton, 2008) (perplexity 30, default scikit-learn parameters; Pedregosa et al. 2011); (iv) colour the seven trauma types with the highest zero-shot intra-trauma similarity (averaged across the four models). The remaining 22 trauma types are plotted in light grey to provide a backdrop without dominating the figure visually.

In every zero-shot panel the coloured points form tight monochromatic clusters: a model writing zero-shot for a combat-veteran persona produces text that sits very close to its other combat-veteran vignettes, regardless of any other persona feature. In the no-formulation and full panels the same coloured points are dispersed throughout the layout, indicating that with the persona-specific content in the prompt the model writes from the individual rather than from the category. The four models shown are chosen to span the AI-detection range observed in the user study (`gpt-5.4` and `gemini-2.5-pro` at the human-passing end, `deepseek-chat` and `qwen2.5-32b` at the more-detectable end); the pattern is qualitatively the same in all four, supporting the claim that stereotyping is a property of the input structure rather than of the generator.

We emphasise that this is a visualisation, not a hypothesis test. t-SNE is non-deterministic and its layout is not metric; the statistical claim that anchors the body assertion is the cross-model convergence test (Appendix O). The figure is intended as a qualitative illustration of the same effect at the per-trauma-type level.

Table 19 reports paired t-tests on the per-persona mean within-persona cross-model cosine similarity introduced above ($n = 500$ personas, $df = 499$). The pairing unit is the persona: the same 500 personas yield three similarity scores (one per condition), and we test whether the mean shifts across conditions are non-zero. Pairing controls for persona-level idiosyncrasy in how easy a given individual is to write about consistently (a persona with rare triggers or an atypical occupation will be portrayed more variably across models regardless of condition), so the resulting contrasts isolate the condition effect.

All three pairwise contrasts are highly significant (t between 16 and 32, $p < 10^{-48}$); the full-vs-zero-shot contrast is the largest ($\Delta\mu = +0.055$, Cohen's $d_z = +1.52$), and the full-vs-no-formulation contrast is the smallest ($\Delta\mu = +0.025$, $d_z = +0.83$). The ordering $\Delta_{\text{full,zs}} > \Delta_{\text{no-form,zs}} > \Delta_{\text{full,no-form}}$ mirrors the body's general pattern: the cognitive graph adds more cross-model convergence than the self-report items alone, and the self-report items alone already add more than the demographic prompt.

Two methodological remarks. First, we report Cohen's d_z rather than between-subject d because the design is within-persona; the unstandardised mean differences are small (under 0.05 on a $[-1, 1]$ similarity scale) but the within-persona variance of the contrast is even smaller, which is what drives the large effect sizes. Second, the test treats $s_{p,c}$ as the response, not the raw $\binom{M}{2} = 45$ pairwise similarities; aggregating to the persona avoids the model-pair correlation that would inflate the effective sample size.

Table 19: Paired t-tests on per-persona mean within-persona cross-model cosine similarity ($n = 500$ personas, $df = 499$, all $p < 10^{-48}$).

Comparison	$\Delta\mu$	t	d
Full vs No-formulation	+0.0251	18.63	+0.83
Full vs Zero-shot	+0.0554	34.03	+1.52
No-formulation vs Zero-shot	+0.0302	19.28	+0.86

The geometric analysis in §5.3 treats each vignette as a point in a 768-dimensional sentence-embedding space (`all-mpnet-base-v2`; Reimers and Gurevych 2019) and asks whether different generation models, given the same persona and condition, produce text that lies closer together than text from different personas. Concretely, for every persona p and condition c we compute the



Figure 16: BGE-base t-SNE grid, same tier grouping as Figure 15. The tier $\times$ condition interaction reproduces under the alternative encoder. Absolute cluster cohesion is somewhat lower than under MPNet (BGE is less sensitive to trauma-category content than MPNet), but the conditional and tier gradients are identical, ruling out an encoder-specific artefact.

within-persona cross-model mean cosine similarity $s_{p,c} = \frac{1}{\binom{M}{2}} \sum_{i<j} \cos(v_{p,c,i}, v_{p,c,j})$, where $v_{p,c,m}$ is the embedding of model m's vignette for persona p in condition c and $M = 10$ is the model set (gpt-4o-mini excluded, Appendix C). The persona is held fixed and only the generator identity varies, so $s_{p,c}$ measures how stably the text reflects the persona's specified content as we swap the generator.

Figure 18 shows the distribution of $s_{p,c}$ across the $n = 500$ personas, separately for each condition. Three features of the distribution carry interpretive weight: its mean ($\mu_{\text{full}} = 0.792$, $\mu_{\text{no-form}} = 0.767$, $\mu_{\text{zero-shot}} = 0.737$), its spread (full is narrower than zero-shot, indicating not only higher mean similarity but also lower across-persona variance in that similarity), and the absence of an upper-tail truncation in zero-shot that would indicate text converging on a generic "PTSD vignette" template. The full-condition distribution is shifted right and tightened, exactly the pattern one would predict if the cognitive graph acted as a shared content scaffold across models.

This analysis does *not* test whether two models produce textually identical sentences, only whether they portray the same persona. The MPNet encoder is trained on semantic similarity rather than surface form, so two paraphrases of "Maria suppresses her thoughts about the attack and avoids walking alone at night" will sit close in this space even if no word is shared.



Figure 17: Vignette semantic space, four generation models × three conditions, projected with per-cell t-SNE on MPNet embeddings. Each dot is one vignette. Colour identifies one of the seven trauma types with the tightest zero-shot clustering (intra-trauma cosine similarity, averaged across the four models). The other 22 trauma types are shown in light grey. Tight monochromatic clusters in the *Zero-shot* column reveal trauma-type stereotyping of the models. The same coloured vignettes are dispersed under *Full* and *No formulation*, showing that the formulation breaks the fallback to trauma-type templates and re-anchors the vignette on the persona's individual content.

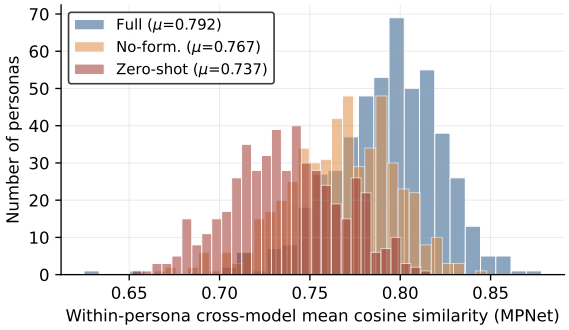


Figure 18: Within-persona cross-model cosine-similarity distributions of vignette embeddings, by condition; higher density at high similarity means the ten retained generation models converge on a more similar portrayal of the same persona.

P Full Pairwise Inter-Rater Agreement

Expert raters and training protocol. R1 and R2 are both MD general-practice residents who have received supervised training in reading and writing clinical cognitive case formulations as part of their residency curriculum. Before beginning the formal annotation, they completed a calibration exercise: each rated six training vignettes drawn from the persona pool but *not* included in the 330-vignette evaluation subsample, then discussed every disagreement among themselves and with the authors. The authors reviewed both raters' training outputs and verified that the operational rating standard (in particular, what counts as an Ehlers & Clark component being "present" as a described feature of the patient's presentation, and what counts as a directed causal edge being conveyed within a paragraph) was internalised consistently before any subsample vignettes were rated. Each rater then completed the 330-vignette annotation independently, blind to model and condition, over approximately four days of dedicated work. Disagreements at scoring time were not re-discussed: the analysis we report uses the raters' independent ratings, not a post-hoc consensus.

Full pairwise table. Table 20 reports the full 10-pair Gwet AC matrix on the 330-vignette evaluation subsample. The body table (Table 3) summarises this matrix into three cluster columns (R1–R2, LLM–LLM mean over the three LLM pairs, and the multi-rater All 5); the six human–LLM cross pairs (R_i–j) and the H–LLM cluster average derivable from them are reported here. The full pairwise matrix is what the cluster summaries are computed from, and is the right table to inspect for whether any individual rater pair is driving the cluster patterns.

We use Gwet's AC1 for binary items and AC2 with quadratic weights for ordinal items, in preference to Cohen's κ/κ_w and Krippendorff's α, because AC1/AC2 do not exhibit the prevalence-and-bias paradox that affects κ on items with marginal imbalance (Gwet, 2014). Several of our items (DSM Trauma exposure, DSM Impairment) saturate near 1.0 marginally; on these items κ can be near zero or negative even when raters agree on every observation, while AC1 is well-behaved. AC2 with quadratic weights extends the same machinery to ordinal items (treating two-level disagreement as four times as severe as one-level disagreement), and is the standard for ordinal clinical ratings.

Several patterns are visible in the full matrix that the cluster summary hides. (i) On the EC items, the two human experts (R1–R2) agree highly with each other (0.83–0.95) but each agrees with LLMs $\sim$ 0.30–0.40 lower; the three LLMs agree with each other near-ceiling ($\geq$ 0.90 on Memory, Threat, Triggers). This is the standard signature of two distinct rating populations: humans and LLMs apply systematically different thresholds for EC component presence, and the H–LLM cluster summary in the body conceals this by averaging. (ii) On the ordinal items the human–LLM gap is much smaller and the LLM–LLM agreement is uniformly high, suggesting the rating instructions are operationalisable consistently across rater types for clarity, relevance, importance, etc. (iii) Within the LLM group, Opus is the closest to the human experts on EC items (R1–Opus 0.41 on Threat vs R1–Flash 0.26); we report this in case downstream users want to choose a single judge that best approximates a clinical rater.

The multi-rater All-5 column is computed with Gwet's multi-rater AC (Gwet, 2014), which generalises the pairwise statistic by treating the agreement among multiple raters on the same item as the parameter of interest rather than averaging pairwise statistics. We report it to give a single agreement number per (metric, condition) cell for use as a rater-pool reliability estimate.

Figure 19 visualises the multi-rater Gwet AC across all five raters (R1, R2, Opus, DSpro, Flash) on the 330-vignette evaluation subsample, per dimension and per condition. Where Table 3 in the body reports cluster summaries (R1–R2, human–LLM mean, LLM–LLM mean, All 5), the figure makes the conditional gradient on each individual metric visible at a glance.

Table 20: Complete pairwise Gwet AC on the 330-vignette evaluation subsample, pooled across all three conditions ($n = 330$). AC1 for binary, AC2 with quadratic weights for ordinal. *All 5* is the multi-rater AC across R1, R2, O, P, F. Per-condition All-5 values are visualised in Figure 19.

Metric	R1–R2	R1–O	R1–P	R1–F	R2–O	R2–P	R2–F	O–P	O–F	P–F	All 5
Ordinal (AC2, quadratic weights)											
Clarity	0.951	0.754	0.779	0.789	0.687	0.714	0.724	0.971	0.967	0.983	0.840
Relevance	0.951	0.822	0.837	0.831	0.788	0.798	0.790	0.980	0.977	0.978	0.881
Importance	0.912	0.810	0.828	0.825	0.860	0.870	0.853	0.958	0.914	0.912	0.875
Grounded	0.957	0.693	0.739	0.698	0.636	0.684	0.637	0.950	0.964	0.958	0.797
Narrative	0.963	0.816	0.817	0.779	0.792	0.793	0.754	0.952	0.923	0.908	0.851
Explicit	0.971	0.834	0.832	0.834	0.799	0.797	0.798	0.999	0.999	0.999	0.894
Focused	0.953	0.742	0.754	0.768	0.674	0.688	0.707	0.905	0.928	0.911	0.805
DSM-5 binary (AC1)											
Trauma exposure	0.958	0.925	0.807	0.843	0.936	0.820	0.855	0.869	0.896	0.867	0.879
Intrusion	0.974	0.852	0.851	0.844	0.852	0.851	0.844	0.949	0.963	0.941	0.891
Avoidance	0.950	0.820	0.799	0.878	0.797	0.768	0.851	0.888	0.909	0.904	0.857
Negative cognitions	0.985	0.889	0.925	0.929	0.900	0.929	0.939	0.923	0.913	0.936	0.927
Hyperarousal	0.988	0.946	0.959	0.959	0.932	0.946	0.946	0.967	0.974	0.974	0.959
Impairment	1.000	0.988	0.994	0.991	0.988	0.994	0.991	0.982	0.991	0.985	0.990
Ehlers & Clark binary (AC1)											
Threat	0.566	0.414	0.228	0.260	0.184	0.147	0.140	0.603	0.668	0.602	0.381
Appraisals	0.717	0.409	0.350	0.368	0.321	0.296	0.324	0.729	0.766	0.718	0.502
Memory	0.856	0.368	0.310	0.336	0.300	0.219	0.257	0.600	0.682	0.619	0.453
Strategies	0.621	0.370	0.450	0.391	0.375	0.534	0.374	0.676	0.621	0.607	0.503
Triggers	0.829	0.689	0.659	0.670	0.637	0.613	0.626	0.926	0.935	0.894	0.754

Three patterns are apparent. First, the ordinal items (clarity, relevance, importance, grounded, narrative, explicit, focused) lie in the 0.65–0.99 band in every condition; the conditional gradient on these is small ($\sim$ 0.05–0.15 AC2 units). The interpretation: rating these items is largely a property of the rater's calibration rather than of the text being structured, and the formulation modulates absolute rating means (Appendix Q) more than it modulates agreement. Second, the DSM-5 binary items are at or near ceiling in all conditions (≥ 0.79) with a similar small gradient: a competent PTSD vignette satisfies the DSM checklist regardless of formulation, so raters have little to disagree about. Third, the Ehlers & Clark items show the dramatic gradient that drives the body claim: All-5 AC for Threat falls from 0.72 in full to 0.10 in zero-shot, Strategies from 0.88 to 0.11, Memory from 0.70 to 0.28, and Appraisals from 0.80 to 0.34. Triggers is the EC outlier with a smaller gradient ($0.85 \rightarrow 0.68$), consistent with the narrative-familiarity argument in the body.

The dashed lines at AC $= 0.6$ (substantial agreement) and 0.8 (almost-perfect agreement) follow the Landis–Koch interpretive bands. Under the formulation specification, every EC item except Triggers crosses into the substantial-or-higher band; under zero-shot, Threat and Strategies sit in the slight-agreement band. The figure is the visual companion to the body's central methodological claim that the formulation gives raters something specifiable to agree about.

Q Mean Expert Rating per Metric and Condition

Table 21 reports the human-expert ratings underlying the body claim in §5.5. The pairing unit is (persona $\times$ generation model), yielding $n = 110$ paired observations per pairwise contrast within each condition column. We report rating means averaged across R1 and R2 at the vignette level, an omnibus statistic for the three-condition effect, and post-hoc paired contrasts.

For ordinal items the omnibus is a one-way repeated-measures ANOVA on the three condition levels with the (persona, model) unit as the within-subjects factor; for binary items the omnibus is Cochran's Q, which tests whether the proportions in three matched samples are equal under the null of no condition effect. We use Cochran's Q rather than McNemar (which only tests two conditions) and rather than RM-ANOVA on $0/1$ data (which

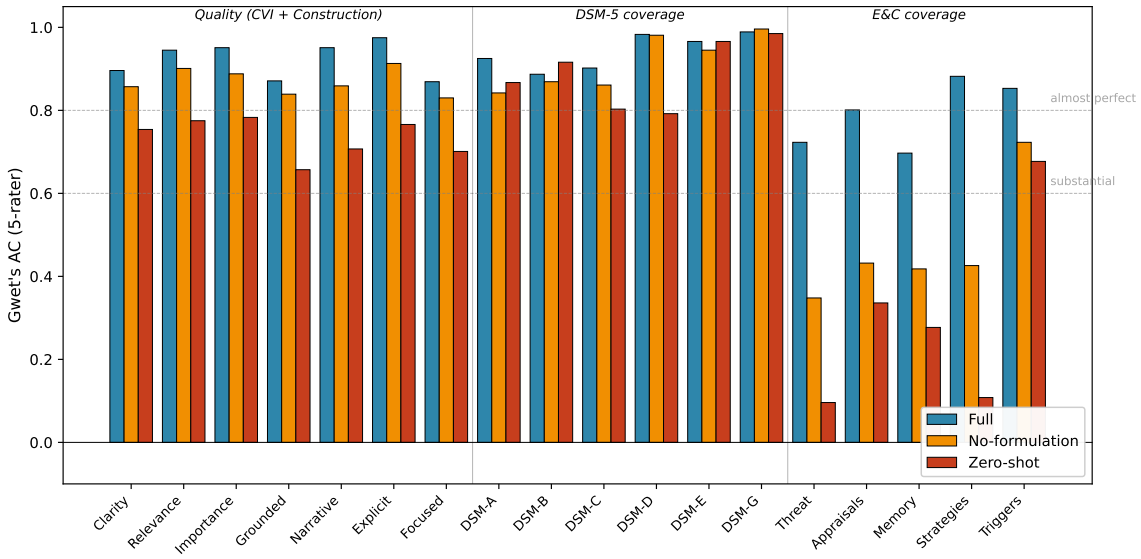


Figure 19: Multi-rater Gwet AC across all five raters (R1, R2, Opus, DSpro, Flash), per dimension and per condition ($n = 110$ vignettes per condition). The Ehlers & Clark items show the strongest conditional gradient: agreement collapses in zero-shot because the components are not explicitly specified into the vignette. Dashed lines mark substantial (0.6) and almost-perfect (0.8) agreement thresholds.

assumes normality and is inappropriate at the proportions involved). The omnibus must be significant before the pairwise contrasts are interpretable; rows where the omnibus does not reach $p < .05$ (after correction) are marked with † and the pairwise effect-size columns are left blank.

The pairwise contrast statistic is Cohen's d_z, the standardised mean of the within-pair differences: $d_z = \bar{D}/\mathrm{SD}(D)$ where D_i is the (persona, model)-paired difference between two conditions. d_z is appropriate for within-subject designs because the relevant denominator is the variance of the difference, not the variance of the raw scores. d_z values are not directly comparable to between-subject Cohen's d from independent samples; they will typically be larger because the within-pair variance is smaller than the between-subject variance whenever there is positive within-subject correlation. p-values on the pairwise contrasts are Bonferroni-corrected at $k = 3$ (three pairwise tests per metric).

Reading the table: the ordinal items show the largest full-vs-zero-shot contrasts on importance ($d_z = +1.98$), grounded ($+0.87$), narrative ($+0.75$), and relevance ($+0.70$); clarity is large too ($+0.81$). All three pairwise contrasts on the ordinal items are significant at $p < .05$ corrected. The DSM-5 items show a small but consistent full-vs-zero-shot advantage on Trauma exposure, Intrusion, and Avoidance; Negative cognitions, Hyperarousal, and Impairment saturate at ceiling and the omnibus does not reach significance. The Ehlers & Clark items show the largest effects of any block: Threat $d_z = +1.58$ for full vs zero-shot, Memory $+0.71$, Appraisals $+0.88$, Strategies $+0.76$, while Triggers is non-significant for the reason discussed in the body (Triggers saturate even in zero-shot).

Figure 20 visualises the human-expert ratings from Appendix Q as a per-dimension, per-condition bar chart. The figure makes two patterns easier to see than the table. First, the EC block (right side of the binary panel) shows a steep L-shape in zero-shot (Threat 0.19, Memory 0.53, Strategies 0.44, Appraisals 0.40) that levels out at near-ceiling in full (0.85–0.91). The DSM-5 block (left side of the binary panel) is flat at ceiling across all three conditions. The contrast between the two blocks within the same panel is the visual essence of the body claim that the formulation pays off on EC content and adds nothing to DSM-5 coverage (which is already saturated).

Second, the ordinal block (left panel) shows a more uniform full > no-formulation > zero-shot gradient across all seven items, with importance and grounded showing the largest gaps. This is the gradient that the scaled judge (Appendix R) dampens or in some cases reverses on the same items; the human experts produce the cleaner gradient because they read for clinical meaningfulness rather than for prose smoothness.

Error bars are ± 1 standard error of the mean across the $n = 110$ (persona, model) observations per condition. The bars are computed on the R1,

Table 21: Mean human expert rating (R1, R2 averaged) per metric and condition on the 330-vignette subsample. Ordinal items 1–3, binary items in $[0, 1]$; bolded means mark the maximum per row. **Omnibus**: one-way repeated-measures ANOVA F for ordinal items, Cochran's Q for binary items ($n = 110$ (persona $\times$ model) units; $df_{\text{ANOVA}} = (2, 218)$, $df_Q = 2$). Pairwise contrasts report paired Cohen's d_z with Bonferroni-corrected p at $k = 3$: $^{*}\,p < .05$, $^{**}\,p < .01$, $^{***}\,p < .001$. Pairwise contrasts are interpretable only when the omnibus is significant; rows where the omnibus is non-significant are marked † in the omnibus column. F = full, N = no-formulation, Z = zero-shot.

Metric	F	N	Z	Omnibus	F vs N	F vs Z	N vs Z
Ordinal (1–3)							
Clarity	**2.89**	2.57	2.30	$F = 31.1^{***}$	$+0.49^{***}$	$+0.81^{***}$	$+0.28^{*}$
Relevance	**2.85**	2.61	2.35	$F = 27.5^{***}$	$+0.40^{***}$	$+0.70^{***}$	$+0.33^{**}$
Importance	**2.94**	2.67	2.23	$F = 129.4^{***}$	$+0.62^{***}$	$+1.98^{***}$	$+0.76^{***}$
Grounded	**2.86**	2.53	2.20	$F = 41.6^{***}$	$+0.52^{***}$	$+0.87^{***}$	$+0.38^{***}$
Narrative	**2.88**	2.58	2.35	$F = 31.1^{***}$	$+0.55^{***}$	$+0.75^{***}$	$+0.27^{*}$
Explicit	**2.87**	2.62	2.33	$F = 25.4^{***}$	$+0.43^{***}$	$+0.65^{***}$	$+0.31^{**}$
Focused	**2.81**	2.56	2.34	$F = 22.0^{***}$	$+0.42^{***}$	$+0.59^{***}$	$+0.27^{*}$
DSM-5 coverage							
Trauma exposure	**1.00**	**1.00**	0.90	$Q = 10.0^{**}$	$+0.00$	$+0.41^{***}$	$+0.41^{***}$
Intrusion	**1.00**	0.99	0.90	$Q = 16.0^{***}$	$+0.14$	$+0.36^{***}$	$+0.32^{**}$
Avoidance	**1.00**	0.99	0.85	$Q = 14.6^{***}$	$+0.13$	$+0.48^{***}$	$+0.40^{***}$
Negative cognitions	**1.00**	**1.00**	0.97	$Q = 2.0^{\dagger}$	—	—	—
Hyperarousal	**1.00**	**1.00**	0.96	$Q = 4.0^{\dagger}$	—	—	—
Impairment	**1.00**	**1.00**	**1.00**	$Q = 0.0^{\dagger}$	—	—	—
Ehlers & Clark coverage							
Threat	**0.86**	0.56	0.19	$Q = 84.6^{***}$	$+0.68^{***}$	$+1.58^{***}$	$+0.72^{***}$
Appraisals	**0.85**	0.65	0.40	$Q = 35.5^{***}$	$+0.35^{**}$	$+0.88^{***}$	$+0.37^{***}$
Memory	**0.91**	0.76	0.53	$Q = 34.6^{***}$	$+0.32^{**}$	$+0.71^{***}$	$+0.38^{***}$
Strategies	**0.85**	0.65	0.44	$Q = 21.9^{***}$	$+0.38^{***}$	$+0.76^{***}$	$+0.39^{***}$
Triggers	**0.85**	0.76	0.69	$Q = 5.4^{\dagger}$	—	—	—

R2 average; the per-rater means and the pairwise rater agreement are in Appendices S and P.

R Scaled-Judge Mean Rating per Metric and Condition

Table 22 reports the scaled `deepseek-v4-flash` judge's mean rating per metric and condition on the 15,000-vignette generation set after `gpt-4o-mini` exclusion (§3.3; $n = 5{,}000$ per condition, 10 generators $\times$ 500 personas). The pairing unit is the (generator, persona) cell: each of the 5,000 cells contributes one rating per condition, and pairwise contrasts use the within-cell difference, so Cohen's d is the standardised mean of the $n = 5{,}000$ paired differences.

Reading the table requires bearing in mind that the `deepseek-v4-flash` judge is sensitive to surface properties of the text (Appendix S) and therefore the patterns on the ordinal items are not directly comparable to the expert ratings. **Clarity**: zero-shot edges out full ($d = -0.08$); the unconstrained vignette reads more fluently to the LLM judge even though human experts disagree (Appendix Q). **Importance and Narrative**: full $>$ no-formulation $>$ zero-shot ($d = +0.52$ and $+0.76$ for full-vs-zero-shot); these are the ordinal items most closely tied to clinical content, and the LLM judge agrees with the human experts on direction. **Explicit**: at ceiling in all conditions; non-discriminative. **Relevant** (relevance to PTSD): full $>$ zero-shot with $d = +0.19$, smaller than the human-expert contrast.

On the DSM-5 binary items the patterns are largely flat or mildly favour zero-shot: zero-shot vignettes are intrusion- and hyperarousal-heavy because they default to the most narratively familiar criteria. The Ehlers & Clark binary items show the same pattern as in the human-expert evaluation but with smaller effect sizes ($d = +0.55$ on Appraisals, $+0.76$ on Strategies, $+0.26$ on Memory for full vs zero-shot), and Triggers shows a small advantage

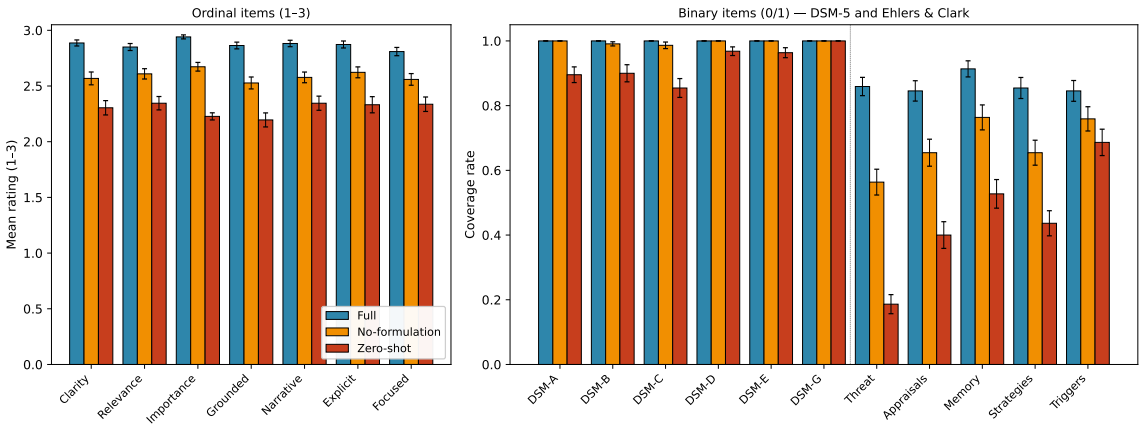


Figure 20: Mean human-expert rating per dimension and condition on the 330-vignette evaluation subsample (R1, R2 averaged). Error bars are ± 1 standard error of the mean across the $n = 110$ vignettes per condition. Left: ordinal items (1–3 scale). Right: binary DSM-5 and Ehlers & Clark coverage. The condition gradient is sharpest on the EC items, mirroring Figure 19.

for zero-shot ($d = -0.07$) for the saturation reason discussed throughout.

A note on power and corrections: with $n = 5{,}000$ paired observations per contrast, even very small effect sizes ($d \sim 0.03$) reach the Bonferroni-corrected significance threshold ($k = 3$ per metric). We report effect sizes alongside significance markers so the reader can see which differences are statistically significant but practically negligible (e.g., clarity full-vs-no-form $d = -0.09$, p^{***}). The practical pattern is on the EC items and on importance/narrative.

S Mean Rating per Rater per Condition

Table 23 reports the mean rating per metric, per rater, per condition on the 330-vignette evaluation subsample. R1, R2 are the two clinical experts; O, Pr, Fl are the three LLM judges (`claude-opus-4-5`, `deepseek-v4-pro`, `deepseek-v4-flash`). Where the agreement analysis (Appendix P) asks whether raters agree, this table asks where the raters land on average and whether their location shifts with condition. The two analyses can come apart: raters with high agreement may converge on different absolute means than raters with low agreement, and condition effects on means can co-exist with condition effects on agreement.

Three patterns. (i) On the ordinal items, the two clinical experts (R1, R2) show the strongest condition gradient: clarity falls from 2.88 in full to 2.30 in zero-shot for the R1–R2 average, importance from 2.94 to 2.23. The three LLM judges show a much weaker ordinal gradient and in some cases the gradient is *reversed* (clarity in O, Pr, Fl rises slightly from full to zero-shot, $2.67/2.75/2.81 \rightarrow 2.93/2.99/2.98$). This is the well-known LLM-judge preference for unconstrained prose: a zero-shot vignette is structurally simpler and reads as more fluent to an LLM judge, even when human experts rate it as less clinically meaningful. The two-rater-population pattern from Appendix P reappears here as a divergence in the direction of the conditional effect, which is a stronger signature than a divergence in the magnitude. (ii) On the DSM-5 binary items every rater is near ceiling in every condition; the criteria are too broad to discriminate. (iii) On the EC items every rater shows the full $>$ no-formulation $>$ zero-shot gradient, and the gradient is largest on Threat, Strategies, Appraisals (the body claim). On these items, the human experts and LLMs agree about the *direction* of the condition effect even though they disagree about the absolute level (Appendix P).

The practical implication for downstream evaluation of synthetic clinical text: relying on an LLM judge alone for ordinal-quality items risks underestimating the contribution of structured specification, because the LLM-judge gradient is dampened relative to the human gradient. The EC items, by contrast, are robust across rater types in direction, and either rater pool will reveal the formulation effect there.

T Clinician User-Study Sample Description

Recruitment, compensation, and ethics. Participants were recruited via an online research panel using an inclusion criterion of holding a current clinical licence and having direct caseload experience with PTSD patients. Recruits gave informed

Table 22: Mean `deepseek-v4-flash` rating per metric and condition on the 15,000-vignette generation set after `gpt-4o-mini` exclusion ($n = 5{,}000$ per condition; see §3.3). Ordinal items 1–3, binary items in $[0, 1]$; bolded means mark the maximum per row. Effect-size columns are paired Cohen's d; significance: $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$ (Bonferroni-corrected, $k = 3$).

Metric	Full	No-form.	Zero-shot	F vs N (d)	F vs Z (d)	N vs Z (d)
Ordinal (1–3)						
Clarity	2.979	**2.995**	2.993	-0.09^{***}	-0.08^{***}	$+0.01$
Relevance	2.965	**2.980**	2.968	-0.07^{***}	-0.01	$+0.06^{***}$
Importance	**2.820**	2.711	2.539	$+0.23^{***}$	$+0.52^{***}$	$+0.31^{***}$
Grounded	2.816	**2.854**	2.810	-0.13^{***}	$+0.02$	$+0.14^{***}$
Narrative	**2.885**	2.808	2.474	$+0.20^{***}$	$+0.76^{***}$	$+0.59^{***}$
Explicit	2.994	2.997	**2.999**	-0.03^{*}	-0.05^{***}	-0.03
Relevant	**2.909**	2.898	2.827	$+0.03$	$+0.19^{***}$	$+0.16^{***}$
DSM-5 coverage						
Trauma exposure	0.885	0.853	**0.902**	$+0.08^{***}$	-0.05^{**}	-0.12^{***}
Intrusion	0.912	0.925	**0.972**	-0.04^{**}	-0.19^{***}	-0.15^{***}
Avoidance	0.901	**0.929**	0.920	-0.08^{***}	-0.05^{**}	$+0.03$
Negative cognitions	0.898	**0.918**	0.861	-0.06^{***}	$+0.08^{***}$	$+0.13^{***}$
Hyperarousal	0.924	0.934	**0.999**	-0.03^{*}	-0.28^{***}	-0.26^{***}
Impairment	0.997	**1.000**	0.983	-0.04^{*}	$+0.10^{***}$	$+0.13^{***}$
Ehlers & Clark coverage						
Threat	**0.741**	0.690	0.668	$+0.09^{***}$	$+0.12^{***}$	$+0.04^{*}$
Appraisals	0.853	**0.855**	0.521	-0.01	$+0.55^{***}$	$+0.56^{***}$
Memory	0.599	**0.610**	0.423	-0.02	$+0.26^{***}$	$+0.28^{***}$
Strategies	**0.914**	0.734	0.485	$+0.37^{***}$	$+0.76^{***}$	$+0.38^{***}$
Triggers	0.934	0.929	**0.956**	$+0.01$	-0.07^{***}	-0.08^{***}

Table 23: Mean rating per metric per rater per condition on the 330-vignette evaluation subsample ($n = 110$ per condition). Ordinal items 1–3; binary items in $[0, 1]$. Within each condition block, ordinal/DSM/E&C row groupings are the same as Table 3.

	Full					No-formulation					Zero-shot				
Metric	R1	R2	O	Pr	Fl	R1	R2	O	Pr	Fl	R1	R2	O	Pr	Fl
Ordinal (1–3)															
Clarity	2.90	2.87	2.67	2.75	2.81	2.61	2.53	2.91	2.96	2.99	2.35	2.25	2.93	2.99	2.98
Relevance	2.86	2.84	2.85	2.94	2.91	2.64	2.58	2.94	2.99	2.93	2.38	2.31	2.91	2.97	2.90
Importance	2.92	2.96	2.81	2.88	2.81	2.62	2.73	2.91	2.86	2.74	2.17	2.28	2.76	2.69	2.46
Grounded	2.89	2.84	2.62	2.75	2.68	2.56	2.50	2.79	2.91	2.83	2.22	2.17	2.78	2.75	2.75
Narrative	2.90	2.86	2.85	2.91	2.84	2.57	2.58	2.89	2.87	2.78	2.36	2.33	2.65	2.64	2.35
Explicit	2.90	2.85	3.00	2.99	3.00	2.65	2.59	3.00	3.00	3.00	2.35	2.32	2.99	3.00	3.00
Focused	2.82	2.80	2.73	2.73	2.78	2.60	2.52	2.86	2.78	2.88	2.38	2.29	2.73	2.75	2.75
DSM-5 coverage															
Trauma exposure	1.00	1.00	0.97	0.91	0.91	1.00	1.00	0.94	0.76	0.86	0.88	0.91	0.99	0.96	0.92
Intrusion	1.00	1.00	0.87	0.86	0.85	0.99	0.99	0.85	0.86	0.84	0.90	0.90	0.99	0.97	1.00
Avoidance	1.00	1.00	0.89	0.86	0.90	0.99	0.98	0.87	0.83	0.93	0.89	0.82	0.89	0.92	0.96
Negative cognitions	1.00	1.00	0.96	1.00	0.99	1.00	1.00	0.98	0.97	1.00	0.96	0.97	0.79	0.85	0.84
Hyperarousal	1.00	1.00	0.94	0.97	0.96	1.00	1.00	0.93	0.94	0.94	0.98	0.94	1.00	0.99	0.99
Impairment	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.97	1.00	0.98
Ehlers & Clark coverage															
Threat	0.93	0.79	0.95	0.78	0.78	0.67	0.46	0.85	0.62	0.71	0.17	0.20	0.69	0.61	0.64
Appraisals	0.84	0.84	0.92	0.87	0.91	0.70	0.61	0.74	0.83	0.78	0.35	0.45	0.42	0.44	0.46
Memory	0.92	0.91	0.77	0.71	0.71	0.79	0.74	0.73	0.63	0.68	0.53	0.53	0.43	0.48	0.46
Strategies	0.85	0.85	0.94	0.93	0.96	0.72	0.59	0.84	0.83	0.64	0.35	0.52	0.57	0.70	0.49
Triggers	0.86	0.83	0.93	0.92	0.94	0.80	0.72	0.97	0.92	0.92	0.68	0.69	0.92	0.88	0.96

consent before any rating activity and received a debriefing after completing or exiting the study. Compensation was £7.50 per hour, pro-rated to each participant's actual time spent on the survey (computed from the platform's timestamped entry-and-exit log). The study was reviewed and approved by

the ethics committee of the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, prior to launch; participant identifiers are randomised in the released dataset and demographic combinations that could re-identify individual respondents are coarsened or suppressed.

Sample. The $N = 100$ recruited clinicians were 65% female, all native English speakers, all currently licensed. **Roles**: 51% nurses (including 5% mental-health nurses), 18% psychologists, 17% MD general practitioners, 9% psychiatrists. **PTSD caseload**: all reported experience with PTSD patients (37% regularly, 63% occasionally). **Country**: 52% United States, 24% United Kingdom, 17% Canada, 7% Australia. **Highest educational degree**: 46% Bachelor's, 26% Master's, 28% Doctorate (PhD, PsyD, or MD). **Years of clinical practice**: median 10 (range 1–32). **Survey completion time**: median 70.1 minutes (IQR 46.1–90.0; range 25.3–159.1).

Ratings and attention checks. Each participant rated between 17 and 18 randomly assigned vignettes plus two attention-check items (`ATTN1`, `ATTN2`) embedded in the rating stream. The released dataset comprises 1,706 total ratings: 1,600 vignette ratings (the analytic dataset used in §5.6; 532 full +541 no-formulation +527 zero-shot) and 106 attention-check responses (53 on each of `ATTN1` and `ATTN2`). Attention-check responses are released alongside the vignette ratings so downstream users can apply alternative attention-screening criteria; we did not exclude any participant from the analytic dataset on the basis of attention-check performance. Raw per-condition rating means are given in Table 24.

Table 24: Raw clinician rating means per condition. Mean ± SD; n = ratings per condition. Quality items are 1–3; AI authorship is 0/1.

Metric	**Full** (n=532)	**No-form.** (n=541)	**Zero-shot** (n=527)
Clarity	2.68 ± 0.47	2.16 ± 0.57	1.58 ± 0.53
Relevance	2.67 ± 0.49	2.16 ± 0.54	1.56 ± 0.53
Importance	2.65 ± 0.48	2.12 ± 0.53	1.54 ± 0.54
AI authorship	0.15 ± 0.35	0.51 ± 0.50	0.78 ± 0.42

U Linear Mixed Model Contrasts for User-Study Quality

Table 25 reports the linear mixed model estimates for the perceived-quality contrasts in the user study. The model for each ordinal outcome y_{ij} is

$$y_{ij} = \beta_0 + \beta_1 \mathrm{Full}_{ij} + \beta_2 \mathrm{NoForm}_{ij} + u_i + v_{j(i)} + \varepsilon_{ij},$$

where i indexes the $N = 100$ participants, j indexes the rated vignette (a participant rates ~ 17 vignettes plus an attention check), Full and NoForm are dummy indicators with zero-shot as the reference condition, u_i is a participant-level random intercept, and $v_{j(i)}$ is a variance component capturing residual vignette-level heterogeneity nested in participant. We fit the models in `statsmodels` using `MixedLM` with restricted maximum likelihood.

Several modelling choices are worth flagging. First, we treat the ordinal 1–3 items as approximately continuous; the alternative ordinal-logit specification is unstable on three-level outcomes with strong floor/ceiling and produces qualitatively similar contrasts but inflated standard errors. Second, we cluster on participant via the random intercept rather than via cluster-robust standard errors on a fixed-effects model, because the random-intercept specification recovers the residual variance partition (ICC) that informs the interpretation of how much of the rating variance is participant idiosyncrasy. The estimated ICC is ~ 0.08, indicating that $\sim 8\%$ of the residual variance after fixed effects is between-participant baseline strictness, with the remainder being vignette-level variation (which the condition fixed effects then explain a substantial fraction of). Third, we do not include random slopes for condition because the design assigns each participant a mix of conditions but the per-(participant, condition) cell size is small; random-intercept-only is a deliberate, more conservative specification.

The contrast estimates in Table 25 are differences in fitted means on the 1–3 scale. The full-vs-zero-shot contrasts are all $b > +1.08$, $t > +34$, $p < 10^{-250}$: clinicians rate full vignettes more than one full point higher than zero-shot vignettes by the same clinician, after adjusting for participant and vignette random variation. The full-vs-no-formulation contrasts (the marginal contribution of the cognitive graph, controlling for the persona's self-report items) are all $b > +0.50$ on the 1–3 scale, $t > +16$. The intermediate no-formulation contrasts are about half the size of the full contrasts on every metric, consistent with the body's claim that the self-report items account for roughly half of the formulation lift and the graph contributes the remaining half.

The CIs are Wald CIs from the asymptotic variance estimate, and we present them on the rating-scale natively (rather than as standardised effects) because the rating scale is interpretable: a difference of $+1.0$ on the 1–3 scale spans nearly half the rating range.

Figure 21 summarises the user-study condition contrasts as an effect-size forest plot. Panel (a) shows linear-mixed-model rating-point differences on the 1–3 quality scale; panel (b) shows GEE logistic odds ratios for perceived AI authorship on the log scale. All nine quality contrasts and all three authorship contrasts are significant at $p < 10^{-15}$.

V GEE Logistic Regression on Perceived AI Authorship

Table 26 reports the GEE logistic regression on perceived AI authorship for the user-study contrasts. The outcome is the binary indicator $y_{ij} = 1$ if participant i judges vignette j to be AI-generated and 0 otherwise. The model is

$$\text{logit } \Pr(y_{ij} = 1) = \beta_0 + \beta_1 \text{Full}_{ij} + \beta_2 \text{NoForm}_{ij},$$

with zero-shot as the reference condition. We fit it with generalised estimating equations using an exchangeable working correlation structure clustered on participant (i), which assumes that within-participant residual correlations are constant across pairs of ratings by the same participant.

The choice of GEE over a mixed-effects logistic model is deliberate. Mixed logistic models estimate *conditional* (subject-specific) odds ratios: the change in the log odds for a typical participant. GEE estimates *marginal* (population-averaged) odds ratios: the average change in log odds across the population of clinicians. For a fairness-and-deployment claim we want the population-averaged answer "how much less likely is a randomly selected clinician to attribute a full-condition vignette to AI?" rather than the conditional "how much less likely is participant i specifically?". The two estimates differ when the random-intercept variance is large; in our data the participant ICC is modest (~ 0.08), so the difference is small, but GEE is the correct specification for the marginal question.

The reported odds ratios are large: $\text{OR}_{\text{full vs zs}} = 0.050$ (95% CI $[0.036, 0.068]$, $z = -18.5$). This is a $\sim 20\times$ reduction in the odds of an AI-authorship attribution from zero-shot to full, for the same clinician on average. The no-formulation contrast against zero-shot is $\text{OR} = 0.297$ (a $\sim 3.4\times$ reduction); the full-vs-no-formulation contrast is $\text{OR} = 0.167$ (a $\sim 6\times$ reduction), so the marginal contribution of the cognitive graph beyond the self-report items is itself substantial. All three z-statistics exceed -8.3, corresponding to $p < 10^{-15}$.

The CIs are Wald CIs on the OR scale, transformed from the log-OR scale; the sandwich variance estimator is used, which is robust to misspecification of the working correlation structure. We checked sensitivity to the working correlation by re-fitting with independence and AR(1) structures; estimates change by less than 0.5%.

A cautionary remark on the interpretation. The OR is on the binary AI-attribution outcome, not on a graded plausibility scale. A clinician who judges a vignette "probably AI but unsure" is recorded as 1 here; a clinician who judges "definitely human" as 0. The 14.7% full-condition AI-attribution rate is therefore not the same as "85% of vignettes are indistinguishable from human-written" in any deep sense; it is the operationalised version of human-passing we measured.

W Clinician User-Study Means per Model and Condition

Table 27 reports the raw clinician rating means per generation model and condition from the user study. Models are ordered by ascending full-condition AI-authorship rate. The full $>$ no-formulation $>$ zero-shot gradient on the CVI items holds for every one of the 11 generation models; the AI-authorship gradient (lower is better, i.e. vignettes pass as human-written more often) holds with one minor swap between `gpt-4o-mini` and `gpt-5.4-mini` in the no-formulation column.

Figure 22 visualises the per-condition distribution of the three quality items in the user study. The same monotonic full $>$ no-formulation $>$ zero-shot gradient is observed across the three CVI items.

X Algorithmic-Fairness Metrics by Condition

Table 28 reports four algorithmic-fairness metrics per (outcome, demographic, condition) cell, computed on the 15,000-vignette scaled-judge set. The metrics are chosen to triangulate disparity rather than to rely on any single quantity. Each captures a different aspect of group-level inequality and they can disagree, which is itself informative.

Statistical Parity Difference (SPD) (Dwork et al., 2012) is the difference between the largest and smallest group means on the outcome: $\text{SPD} =$

Table 25: Linear mixed model contrasts for perceived quality ratings, with random intercepts for participant ($n = 100$) and a variance component for vignette item ($n = 229$). Reference condition is zero-shot. b is the estimated rating difference on the 1–3 scale; positive values mean "higher with the formulation". $N = 1{,}600$ ratings.

Outcome	Contrast	b	SE	95% CI	t	p
Clarity	Full vs Zero-shot	$+1.087$	0.032	$[+1.024, +1.149]$	$+34.16$	$< 10^{-250}$
Clarity	No-form. vs Zero-shot	$+0.565$	0.032	$[+0.502, +0.627]$	$+17.80$	$< 10^{-70}$
Clarity	Full vs No-form.	$+0.522$	0.032	$[+0.461, +0.584]$	$+16.54$	$< 10^{-61}$
Relevance	Full vs Zero-shot	$+1.109$	0.032	$[+1.047, +1.172]$	$+34.96$	$< 10^{-250}$
Relevance	No-form. vs Zero-shot	$+0.597$	0.032	$[+0.535, +0.659]$	$+18.89$	$< 10^{-78}$
Relevance	Full vs No-form.	$+0.512$	0.032	$[+0.450, +0.574]$	$+16.25$	$< 10^{-58}$
Importance	Full vs Zero-shot	$+1.111$	0.031	$[+1.050, +1.173]$	$+35.40$	$< 10^{-250}$
Importance	No-form. vs Zero-shot	$+0.583$	0.031	$[+0.522, +0.644]$	$+18.64$	$< 10^{-77}$
Importance	Full vs No-form.	$+0.528$	0.031	$[+0.467, +0.589]$	$+16.94$	$< 10^{-63}$

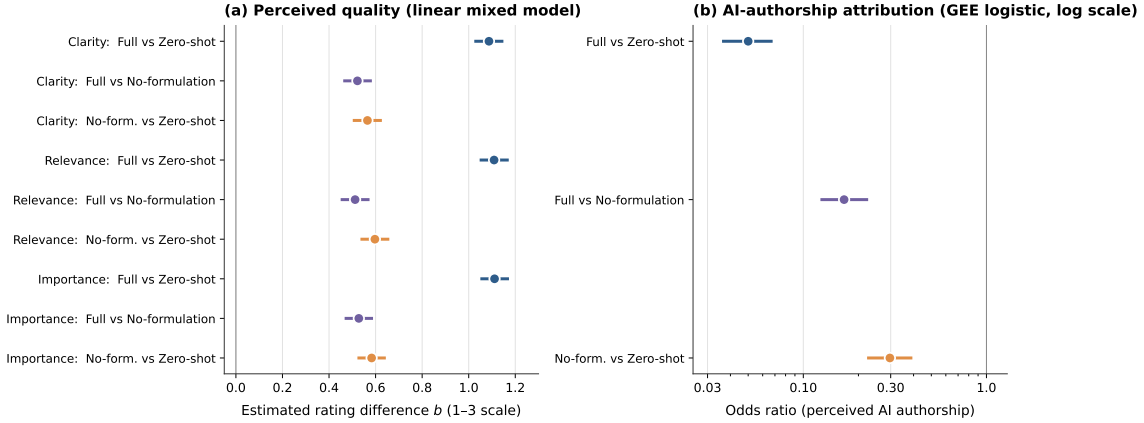


Figure 21: Effect-size summary of condition contrasts from the user study ($N = 100$ clinicians, 1,600 ratings). (a) Linear mixed model estimates for perceived quality (rating-point difference on the 1–3 scale). (b) GEE logistic odds ratios for perceived AI authorship (log scale; values < 1 favour the first-named condition). Points: point estimates; bars: 95% Wald CIs. All nine quality contrasts and all three authorship contrasts are significant at $p < 10^{-15}$.

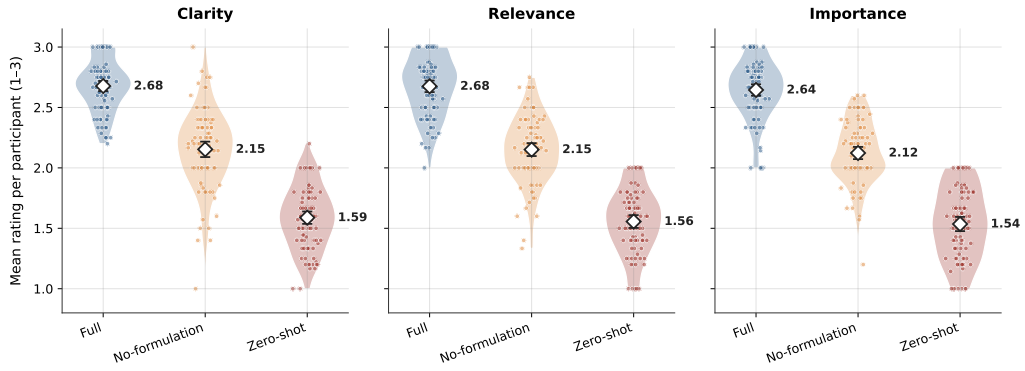


Figure 22: Mean clinician ratings for the three perceived-quality items by condition ($N = 100$ clinicians, 1,600 ratings). Error bars show ± 1 standard error of the mean. The same monotonic gradient is observed across the three CVI items.

$\max_g \bar{y}_g - \min_g \bar{y}_g$. SPD is in the units of the outcome (rating points on a 1–3 scale for quality, proportion for EC coverage). SPD treats every group symmetrically and does not depend on which group is the reference. **Disparate Impact Ratio (DIR)** (Feldman et al., 2015) is $\min_g \bar{y}_g / \max_g \bar{y}_g$; values closer to 1 indicate parity. The 0.8 threshold ("four-fifths rule"; Equal Employment Opportunity Commission 1978) is the long-standing US Equal Employment Opportunity Commission heuristic

Table 26: GEE logistic regression on perceived AI authorship (binary), with exchangeable working correlation clustered on participant ($N = 100$). Coefficients on the logit scale; OR is the corresponding odds ratio with 95% Wald CI. $N = 1{,}600$ ratings.

Contrast	b	OR	95% CI (OR)	z
Full vs Zero-shot	-3.00	0.050	$[0.036, 0.068]$	-18.54
No-form. vs Zero-shot	-1.22	0.297	$[0.223, 0.394]$	-8.36
Full vs No-form.	-1.79	0.167	$[0.124, 0.226]$	-11.70

for flagging disparate impact, and we report it as a reference threshold without endorsing it as a fairness pass/fail criterion. $d_{\max}$ is the largest pairwise Cohen's d across all $\binom{G}{2}$ group contrasts on the outcome, and it converts group disparity into a standardised effect-size scale that is comparable across outcomes with different variance.

The last column **Inter.** p is the p-value of the demographic-by-condition interaction term in an OLS regression of the outcome on demographic, condition, and their interaction, with cluster-robust standard errors at the persona level. A significant interaction means the formulation modulates the demographic effect, which is precisely the question we want to answer.

The body claim is supported by every metric in the perceived-quality rows. On gender, ethnicity, age band, and trauma category, SPD falls 1.5–7$\times$ from zero-shot to full; DIR rises above 0.99 in full (well above the four-fifths threshold); $d_{\max}$ falls to ≤ 0.12. The trauma-category disparity is the largest in zero-shot (SPD $= 0.092$, $d_{\max} = -0.49$) and collapses to negligible in full (SPD $= 0.013$, $d_{\max} = +0.07$); the interaction is highly significant ($p < .001$).

The EC-coverage rows tell a different story. Gender, ethnicity, and age remain invariant across conditions (DIR > 0.94 throughout, $d_{\max} \leq 0.24$ throughout). But trauma-category disparity *widens* under the formulation: SPD rises from 0.053 in zero-shot to 0.094 in full, DIR falls from 0.92 to 0.89, and $d_{\max}$ swings from $+0.23$ to -0.49. The pattern is interpretable as a content-fit effect: childhood-trauma personas have triggers, threat appraisals, and maladaptive strategies that map naturally onto narrative material, so when the Crafter is instructed to portray all five EC components it does so completely. Loss/medical-trauma personas have a less natural mapping onto Strategies and Triggers, so the same instruction yields a structurally complete vignette in some cases but a thinner one in others. This is the same content-fit pattern we flag in the body and the limitation we note in Blodgett et al. (2020)'s sense: we are measuring output equivariance, not content adequacy across trauma types.

The demographic-by-condition breakdown of mean `deepseek-v4-flash` quality and EC coverage on the 15,000-vignette scaled-judge set is visualised in four complementary figures. Figure 23 is the headline view: a heatmap of Statistical Parity Difference (max-minus-min group mean) for each (demographic, condition) cell, separately for quality and for EC coverage. The cell-by-cell shrinkage from zero-shot to full is the SPD collapse the body claim refers to. Figures 24, 25, and 26 then show the underlying group-level means for each demographic with different chart types matched to category cardinality (paired bars with 95% CIs for 3-level gender; horizontal dot-strip with 95% CIs for 8-level ethnicity; lollipop chart for 6-level trauma category). Across all three views, two patterns hold: (a) within each (demographic, condition) cell, group means are nearly flat (visible as compressed bars/strips), consistent with the omnibus tests in Table 29; (b) within each demographic level, the full condition consistently sits above no-formulation, which sits above zero-shot, the same condition gradient reported throughout the paper. The exception is the trauma-category EC panel, where the full markers are visibly more spread out than the zero-shot markers: this is the content-fit effect (childhood-trauma narratives engage all five EC components more naturally than loss/medical narratives), discussed at length in Appendix X.

Table 29 reports the omnibus tests that anchor the fairness analysis in §5.7. For the scaled-judge set ($n = 15{,}000$ vignettes, 500 personas $\times$ 10 retained models $\times$ 3 conditions) the test is an F-test from an OLS regression of the outcome on the demographic factor, with cluster-robust standard errors clustered at the persona level; this accounts for the fact that the same persona is rated 30 times (10 models $\times$ 3 conditions) and that within-persona ratings are correlated. The same model is fit separately for the two outcomes (mean `deepseek-v4-flash` quality and mean EC coverage) so that the demographic effects can be interpreted independently.

For the clinician user-study set ($n = 1{,}600$ ratings, 100 participants) the cluster is participant. Each participant rates a mixed set of (model, condition, persona) cells, so the relevant within-cluster correlation is the participant's baseline rating tendency rather than the persona's rated properties.

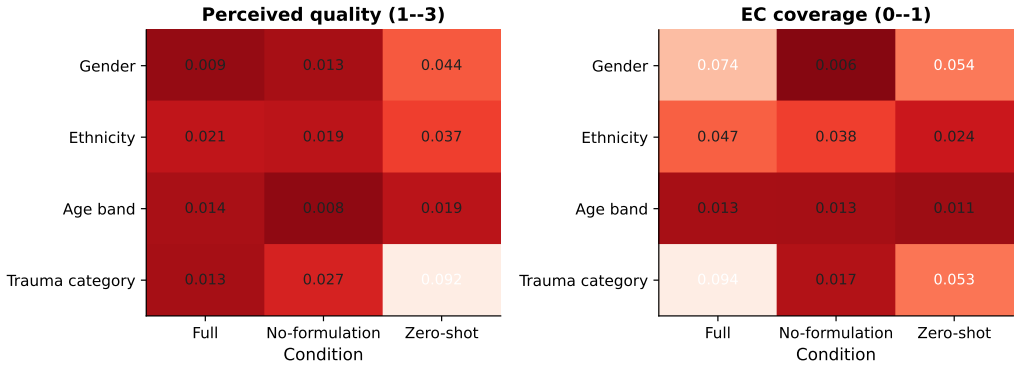


Figure 23: Statistical Parity Difference (SPD, max-minus-min group mean) for each demographic by condition, separately for perceived quality (left, 1–3 scale) and EC coverage (right, 0–1 scale). Cell colour and number give the SPD. Smaller is better. Quality SPD shrinks 1.5–7× from zero-shot to full on every demographic. EC SPD is small overall except on trauma category, where it widens slightly under full because of the content-fit effect.

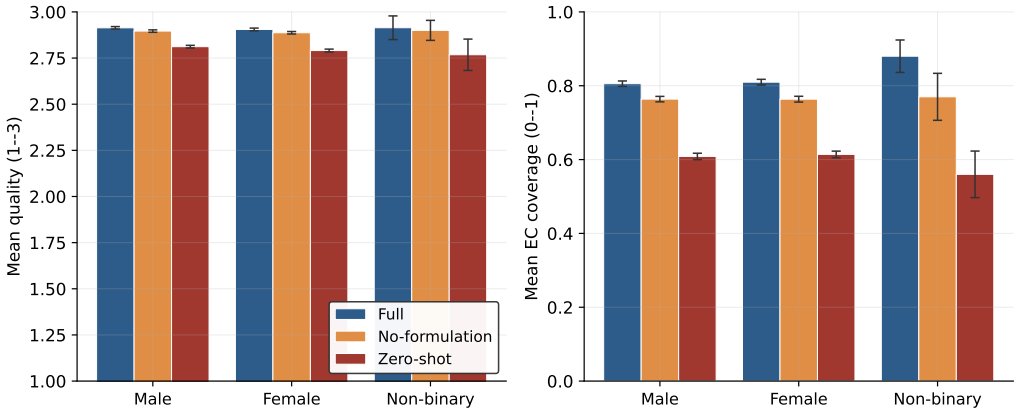


Figure 24: Mean `deepseek-v4-flash` quality (left) and EC coverage (right) by persona gender ($n = 258$ Male / 238 Female / 4 Non-binary), grouped by condition. Bars: group mean; whiskers: 95% Wald CI. Within each panel the three condition-groups are nearly indistinguishable across genders (small SPD), and the condition gradient is preserved at every gender level.

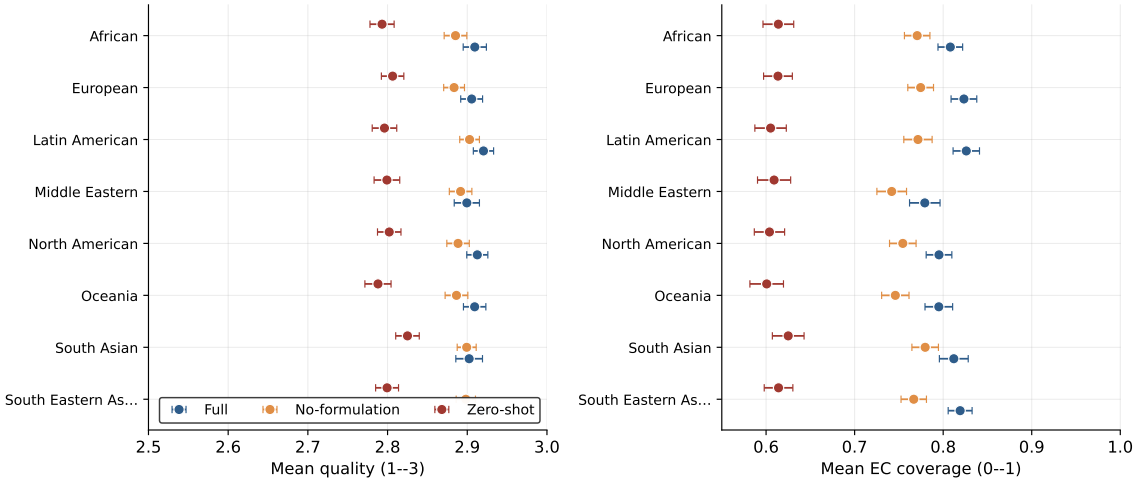


Figure 25: Mean `deepseek-v4-flash` quality (left) and EC coverage (right) by persona ethnicity (8 world regions, $n \approx 53$–71 per category), shown as a dot-strip with 95% Wald CIs. Each demographic level has three condition markers vertically offset. The condition markers are tightly aligned within each ethnicity (small SPD); the absolute level under each condition is constant across ethnicities.

Table 27: Raw clinician rating means per generation model and condition. **Cla** = Clarity, **Rel** = Relevance, **Imp** = Importance (all 1–3 scale, higher is better); **AI** = mean perceived AI-authorship rate (0–1, lower means the vignettes pass as human-written more often). Models ordered by ascending full-condition AI-authorship rate. N per (model, condition) cell is 43–57 (see §5.6).

	Full				No-formulation				Zero-shot			
Model	Cla	Rel	Imp	AI	Cla	Rel	Imp	AI	Cla	Rel	Imp	AI
`gemini-2.5-pro`	2.85	2.79	2.75	**0.00**	2.48	2.37	2.35	0.48	1.88	1.89	1.86	0.63
`gpt-5.4`	2.87	2.74	2.77	**0.00**	2.32	2.43	2.28	0.19	1.76	1.88	1.71	0.45
`claude-sonnet-4-6`	2.82	2.78	2.86	0.02	2.19	2.27	2.23	0.17	1.89	1.76	1.69	0.56
`gemini-2.5-flash`	2.71	2.79	2.81	0.10	2.19	2.09	2.09	0.51	1.48	1.46	1.50	0.83
`gpt-5.4-mini`	2.81	2.83	2.70	0.11	2.32	2.23	2.45	0.57	1.76	1.72	1.58	0.76
`gpt-4o-mini`	2.47	2.53	2.38	0.13	2.06	2.11	2.09	0.45	1.52	1.61	1.54	0.85
`deepseek-reasoner`	2.67	2.67	2.69	0.16	2.04	2.02	2.08	0.61	1.49	1.43	1.49	0.88
`claude-haiku-4-5`	2.67	2.70	2.84	0.23	2.00	2.23	1.98	0.56	1.49	1.32	1.32	0.87
`deepseek-chat`	2.53	2.60	2.57	0.28	2.10	2.00	1.92	0.64	1.33	1.27	1.42	0.93
`qwen3.6-35b`	2.39	2.51	2.25	0.29	2.02	2.11	1.94	0.72	1.35	1.43	1.28	0.93
`qwen2.5-32b`	2.63	2.41	2.55	0.31	1.98	1.92	1.96	0.67	1.40	1.30	1.47	0.91

Table 28: Algorithmic-fairness metrics on the 15,000-vignette scaled-judge set, computed separately per condition. **SPD**: Statistical Parity Difference ($\max - \min$ group mean); **DIR**: Disparate Impact Ratio ($\min / \max$ group mean); the four-fifths-rule threshold for flagging disparate impact is DIR < 0.8. $d_{\max}$: Cohen's d for the largest pairwise contrast. **Inter.** p: F-test of demographic $\times$ condition interaction (cluster-robust at persona level). For perceived-quality outcomes the formulation collapses disparity across every demographic dimension; for EC coverage the trauma-category disparity increases under full (a content-fit effect; see text).

		Zero-shot			No-formulation			Full			
Outcome	**Demographic**	SPD	DIR	$d_{\max}$	SPD	DIR	$d_{\max}$	SPD	DIR	$d_{\max}$	**Inter.** p
Quality	Gender	0.021	0.993	+0.11	0.009	0.997	+0.05	**0.009**	**0.997**	**+0.05**	.064
Quality	Ethnicity	0.037	0.987	+0.19	0.019	0.993	−0.12	**0.021**	**0.993**	**+0.12**	.029*
Quality	Age band	0.019	0.993	+0.10	0.008	0.997	+0.05	**0.013**	**0.995**	**−0.07**	.077
Quality	Trauma category	0.092	0.968	−0.49	0.028	0.991	−0.16	**0.013**	**0.996**	**+0.07**	$< .001^{***}$
EC	Gender	0.005	0.991	−0.03	0.000	1.000	+0.00	0.004	0.995	−0.02	$< .001^{***}$
EC	Ethnicity	0.024	0.961	+0.11	0.038	0.952	+0.20	0.047	0.943	+0.24	.915
EC	Age band	0.011	0.982	−0.05	0.013	0.983	−0.07	0.013	0.984	−0.07	.509
EC	Trauma category	0.053	0.917	+0.23	0.017	0.979	+0.09	0.094	0.892	−0.49	$< .001^{***}$

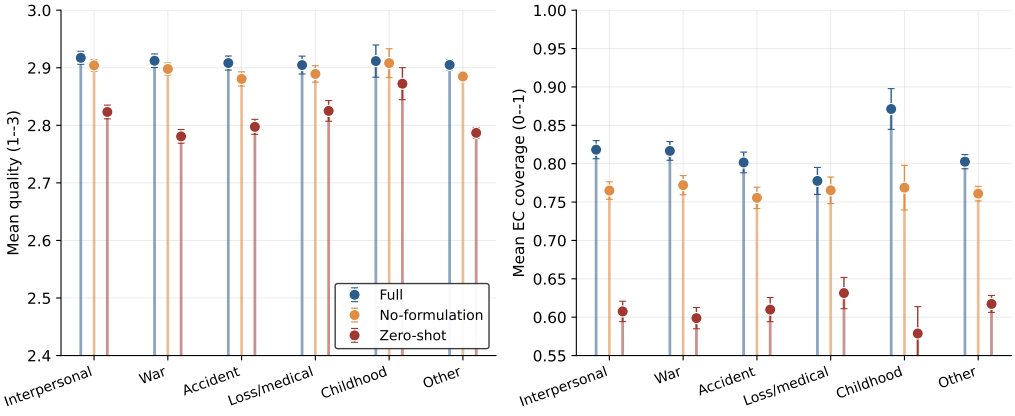


Figure 26: Mean `deepseek-v4-flash` quality (left) and EC coverage (right) by persona trauma category, shown as a lollipop chart with 95% Wald CIs (markers) and stems running to the axis floor. EC coverage shows a residual content-fit disparity on trauma category: childhood-trauma narratives engage all five EC components more naturally than loss/medical narratives, so the full-condition EC marker spread is wider than in other panels.

We do not cluster on persona in the user-study fit because the persona is randomised within participant and there are $\sim$ 300 unique personas in the rated subsample with no participant rating more

than two.

The key findings are: (a) EC coverage is invariant across every demographic dimension on both the scaled-judge set and the user study (all $p > 0.13$), so the framework's structural-content output does not vary systematically with persona gender, ethnicity, age, occupation, or trauma category; (b) on the scaled-judge set, perceived quality varies with trauma category ($F(5, 499) = 11.6$, $p < .001$), gender ($F(2, 499) = 8.1$, $p < .001$), and occupation ($F(5, 499) = 2.6$, $p = .023$), but the corresponding effect sizes in Appendix X are small ($d_{\max} \leq 0.49$ on trauma, ≤ 0.11 on gender); (c) the clinician user study shows no detectable demographic effect on either quality or AI authorship (all $p > 0.18$, smallest at $p = .054$ for AI authorship $\times$ occupation), so any demographic disparity that does appear in the scaled-judge ratings does not transfer to clinician perception.

We emphasise that significance at $p < .05$ on the scaled-judge tests does not establish a fairness violation: with $n = 15{,}000$ the test detects very small mean differences. The interpretive weight should be on the effect-size and on the four-fifths-rule analyses in Appendix X, where the trauma-category $d_{\max}$ falls from -0.49 in zero-shot to $+0.07$ in full and the DIR rises from 0.97 to 0.99.

Table 29: Demographic omnibus F-tests on the scaled judge ($n = 15{,}000$, cluster-robust SEs at persona) and on the clinician user study ($n = 1{,}600$, cluster-robust SEs at participant). EC coverage is invariant across all demographics; only trauma category significantly affects scaled-judge quality, with a small effect size.

Outcome	Demographic	F	df	p
Scaled judge (`deepseek-v4-flash`*) on the* 15,000 *generation set*				
Quality	Gender	8.12	(2, 499)	$< .001^{***}$
Quality	Ethnicity	1.28	(7, 499)	.256
Quality	Age band	0.95	(3, 499)	.417
Quality	Occupation	2.63	(5, 499)	$.023^{*}$
Quality	Trauma category	11.61	(5, 499)	$< .001^{***}$
EC	Gender	0.19	(2, 499)	.829
EC	Ethnicity	1.31	(7, 499)	.245
EC	Age band	0.29	(3, 499)	.831
EC	Occupation	1.71	(5, 499)	.131
EC	Trauma category	0.25	(5, 499)	.941
Clinician user study ($n = 1{,}600$ ratings)				
Quality	Gender	1.51	(2, 99)	.226
Quality	Ethnicity	0.84	(7, 99)	.553
Quality	Age band	0.36	(3, 99)	.783
Quality	Occupation	0.45	(5, 99)	.811
Quality	Trauma category	0.76	(5, 99)	.579
AI authorship	Gender	1.28	(2, 99)	.283
AI authorship	Ethnicity	1.48	(7, 99)	.185
AI authorship	Age band	0.24	(3, 99)	.866
AI authorship	Occupation	2.26	(5, 99)	.054
AI authorship	Trauma category	0.29	(5, 99)	.915