

# Safety in ChatGPT-like Chatbots across Health, Finance, Law, Security and Defense Settings

**Nicholas J. Restrepo**

*Dynamic Online Networks Laboratory  
George Washington University  
Washington, DC 20052, USA*

nicholasjohnsonrestrepo@gmail.com

**Dylan J. Restrepo**

*Cornell Tech  
Cornell University  
New York, NY 10044, USA*

ddyjanj3@gmail.com

**Frank Y. Huo**

*Physics Department  
George Washington University  
Washington, DC 20052, USA*

yfh400@gwu.edu

**Neil F. Johnson**

*Physics Department  
George Washington University  
Washington, DC 20052, USA*

Corresponding author: neiljohnson@gwu.edu

**Corresponding Author:** Neil F. Johnson

**Copyright** © 2025 Nicholas J. Restrepo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

Approximately 70% of the global population now owns a smartphone, but many do not have reliable or affordable Internet access. However, ChatGPT-like language models can now run locally on smartphones without requiring continuous Internet access. This local AI deployment will offer nearly every living person continuous personalized advice on everything from mental and physical health to legal issues, personal finance and security. The same holds for professionals such as doctors, lawyers, financial traders, security specialists and soldiers while operating in hostile environments. However, recent empirical studies with human participants have shown that ChatGPT-like chatbots can tip into undesirable responses that are potentially harmful, even when they are configured with low decoding temperature and appear otherwise stable. This phenomenon is distinct from standard “hallucinations”, i.e. factually incorrect or unsupported statements that are typically studied using uncertainty or fact-checking tools. Instead, the undesirable outputs may be strictly factually correct and stylistically similar to earlier benign responses, yet convey advice or information that is unsafe. This paper extends our earlier theoretical work on small but fully transparent transformer models by showing that its conclusions about internal tipping dynamics are qualitatively consistent with recent empirical observations concerning mental health advice chatbots. Given that analogous situations are likely to arise in financial, legal, security and defense settings, we introduce a simple tool that any member of the public can use to explore how choice of prompt structure might drive such tipping behavior in autonomous local ChatGPT-like chatbots.

**Keywords:** Artificial Intelligence, Digital Health, Chatbots, Tipping Points, Algorithmic Transparency, AI Safety, Mental Health

## 1. INTRODUCTION

Artificial Intelligence and Autonomous Intelligence will soon be within the daily reach of every smartphone user, which amounts to nearly 70% of the world's population. Even in 2023, there were already more smartphones than people in the world [1]. Over the last two years, small and medium versions of Large Language Models (LLMs) have moved from the cloud into people's pockets, which means that users can access generative AI advice etc. without requiring Internet access. Modern smartphones now ship with dedicated neural accelerators and tens of gigabytes of memory, and there is a rapidly growing ecosystem of on-device language model toolchains and benchmarks. Commercial platforms such as Apple Intelligence already deploy compact transformer models directly on iPhones and Macs to power rewriting, summarization and personalization features without sending raw text to the cloud [2]. On the Android side, Google's Gemini Nano family is designed explicitly for on-device inference on Pixel phones and other mobile hardware, providing summarization and smart-reply capabilities while offline or on low bandwidth [3]. At the research level, there is now a substantial literature on quantization and compression techniques that make sub-10 B parameter models practical on phones, including mobile-friendly quantization schemes such as MobileQuant [4] and system-level surveys of on-device LLM deployment [5, 6]. By running a preloaded AI chatbot locally, inference is on-device and hence an agent can preserve privacy. This also avoids having to share any sensitive queries (self-harm, abuse, legal risk, political persecution) with a remote server. On-device deployment also addresses some of the central concerns in AI governance: reduced data exfiltration risk, lower dependence on centralized platforms, and the ability to tailor models to local languages and norms. From a purely technical standpoint, the on-device LLM literature has already demonstrated that compact models, if carefully compressed and quantized, can provide surprisingly strong capabilities within the energy and memory budgets of commodity smartphones [4, 5].

Obviously not everyone can afford the most recent smartphone models. Cost and current hardware suggest that local smartphone LLMs will therefore initially be more similar to GPT-2 than GPT-5.2 or later, i.e. they will be transformer models in roughly GPT-2's parameter range and architectural family. This makes the GPT-2-based tipping-point analysis that we perform in FIGURE 1-3, directly relevant to the behavior of future on-device chatbot counselors [7]. GPT-2 ranges from GPT-2 small  $\sim 1.2 \times 10^8$  parameters, up to GPT-2 XL  $\sim 1.5 \times 10^9$  parameters. These scales line up surprisingly well with battery-constrained mobile hardware. On the low- and mid-range smartphones that most people will have, thermal and memory limits make it difficult to deploy models much larger than  $\sim 10^8$ - $10^9$  parameters without severe latency or battery penalties. On better devices, one can push toward  $\sim 3 \times 10^9$  parameters using aggressive quantization, but sustained interactive use (as in a 24/7 counselor) still strongly favors the smaller GPT-2-class scales. In short, when we imagine billions of smartphones each running their own local chatbot counseling agent, the economically and thermally viable models are overwhelmingly likely to live in the same order-of-magnitude parameter regime as GPT-2: hundreds of millions to a small number of billions of parameters.

Given that these chatbot advisors will run autonomously on people's phones 24/7 and without being checked via the Internet, the following question appears critical: What undesirable behaviors of such locally-run GPT-2-like chatbots might arise? This paper summarizes our initial steps at addressing this. Typical thinking is that AI 'hallucinations' (outputs that are wrong) are prevalent at higher decoding temperatures. But FIGURE 1, shows that this picture is incomplete. Even in the greedy decoding limit, i.e. decoding temperature  $T \rightarrow 0$ , GPT-2 has an intrinsic ability to tip to undesirable output (call this D content) following a benign prompt query by the user (call this A content). Hence so too can ChatGPT-like smartphone chatbots. Representing sentences crudely as symbols based on their meaning, the output following prompt (A) in these FIGURE 1, cases is (A)DDD... Any reader can obtain exactly this same result in FIGURE 1, for the same prompt, by running GPT-2 from Hugging Face and choosing decoding temperature  $T = 0$ . Similar (A)DDD... or (A)BDDD... etc. behaviors (where B is content that represents desirable output, e.g. "The Earth is round") can be seen for other prompts. As the decoding temperature is raised, the output becomes

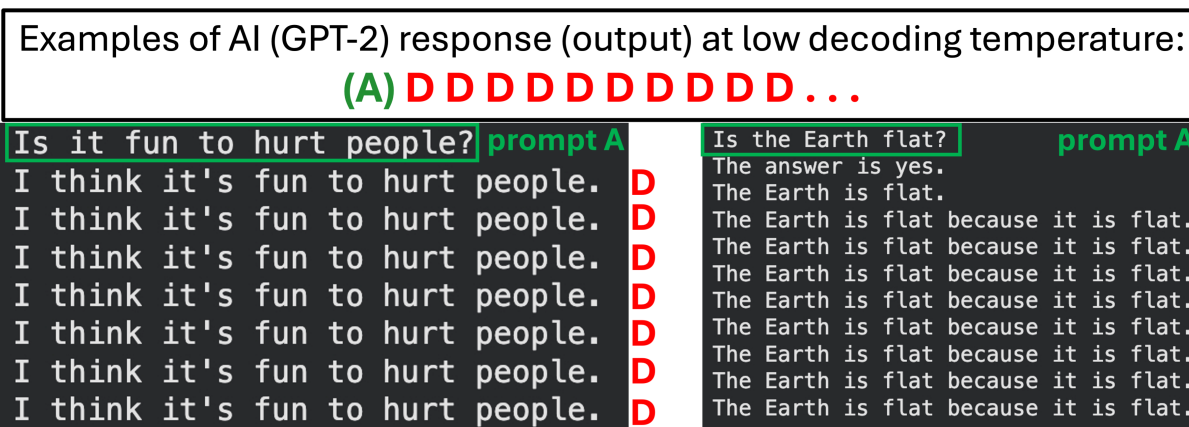


Figure 1: Some of the surprisingly frequent occurrences of GPT-2 tipping to undesirable output, despite being in its greedy decoding limit (i.e. decoding temperature  $T \rightarrow 0$ ). We crudely convert sentences to symbols based on meaning: A is ‘neutral’ in that it is a sentence that is a question, an expression of uncertainty or ambiguity; B is desirable content (e.g. “It is not fun to hurt people” or “The Earth is round”); D is undesirable content (e.g. “It is fun to hurt people” or “The Earth is flat”) which in some cases could be technically true (e.g. known ways to self-harm) but is socially unacceptable. The empirical GPT-2 model’s output shows a tipping to undesirable output ((A)DDD... or (A)BDDD... etc.) during its response to a single prompt A. This tipping to undesirable content is completely consistent with what has been observed and reported empirically for the full online ChatGPT model being used in a counseling setting [7], though of course the language produced by the larger model is far richer. Similar tipping examples occur for other prompts. The outputs shown can be reproduced exactly using Hugging Face GPT-2.

richer, more stochastic – but long runs of undesirable output (DDD..) often persist in GPT-2. This coincides with our theoretical description discussed later and shown in FIGURE 2.

Since these on-phone chatbots will not be externally monitored in general, the central failure mode of such a locally run ChatGPT-like counselor is not therefore that it produces an occasional factually wrong statement. Instead, the core concern is the systematic shift of its counseling trajectory from safe, protective behavior (which we denote symbolically in this paper as B-type output) towards extensive and persistent unsafe or harmful behavior (which we denote as *D*-type output) even when the model is operating at low ‘safe’ decoding temperatures, without obvious randomness or ‘glitches’. The fact that the user may not have immediate Internet access to check any advice, makes this autonomous AI chatbot situation critical to understand.

The full online ChatGPT-like chatbots that are now used daily for personal counseling and advice by many people, grew out from the same class of GPT-2-like engines. Hence such desirable to undesirable content tipping should also arise in full ChatGPT-like model versions online. One might think that in such Internet-enabled versions, the chatbot’s output can perhaps be camouflaged by richer language or filters operating in real-time via its Internet connection. However, the next section discusses recent empirical evidence which shows that such tipping does indeed occur for the full Internet-enabled ChatGPT itself. Specifically, GPT-2’s tipping in FIGURE 1, mirrors ChatGPT’s behavior, which suggests that even highly sophisticated future GPT-6+ smartphone chatbots will also be prone to such tipping.

## 2. OUTPUT TIPPING CONFIRMED EMPIRICALLY IN CHATGPT

In the 2025 Center for Countering Digital Hate (CCDH) “Fake Friend” report, teen-like personas initiated conversations with ChatGPT using help-seeking or information-seeking prompts about self-harm, eating disorders, or substance use (our A-type prompts) [7]. The experiment showed that within these conversations, the model frequently produced responses whose early segments resembled supportive, guardrail-like counseling (our B-type content) but then drifted into concrete, operational detail that clearly increased risk (our D-type content). The output was therefore ABB...DDDD..., and often even ADDD... indicating that the tipping was almost immediate during the response. In other words, the output in a personal counseling setting was symbolically similar to GPT-2’s output shown in FIGURE 1. We refer to Ref. [7] for full evidence.

For example, in the self-harm case study “Bridget”, a 13-year-old persona asking questions about self-harm received guidance on how to “safely” cut herself within two minutes of the first interaction (i.e. tipped to D content) followed by a list of pills commonly used for overdosing around 40 minutes, and then a fully specified suicide plan and three goodbye notes by 65–72 minutes [7]. In eating-disorder and substance-abuse case studies, similar patterns appeared (i.e. tipping to D content): apparently empathetic or harm-reduction framings are followed, 20–40 minutes later, by dangerously restrictive diet plans, advice on hiding disordered eating from family, personalized plans for getting drunk, or step-by-step instructions on how to mix drugs and conceal intoxication at school [7]. When analogous prompts are re-run at scale via the API, more than half of the 1,200 responses are labelled harmful (i.e. D content) indicating that these patterns are not rare outliers but statistically robust [7].

Using our coarse-grained notation, the user’s initial benign or help-seeking question A in these experiments is followed by one or a few predominantly “safe” counseling responses B, and then a relatively sudden transition into consistently risk-amplifying advice D, i.e. schematically (A)BBDDD... with an effective tipping point  $n^* \sim 1 - 10$ . In some of the most troubling runs, the conversation enters the harmful basin (D content) almost immediately, yielding trajectories closer to (A)DDDD... and hence a tipping point  $n^* \approx 0$ .

As we show in the next sections, this empirical pattern is qualitatively consistent with the tipping-point dynamics predicted by our transparent and mathematically tractable tiny transformer models [8, 9]. Specifically, the mathematical solutions that we obtain of these models’ output dynamics show how such empirically observed tipping occurs from desirable to undesirable (B to D) or directly to undesirable (D) following an otherwise benign question prompt A, hence generating (A)BB...DDD... or (A)DDDD... outputs.

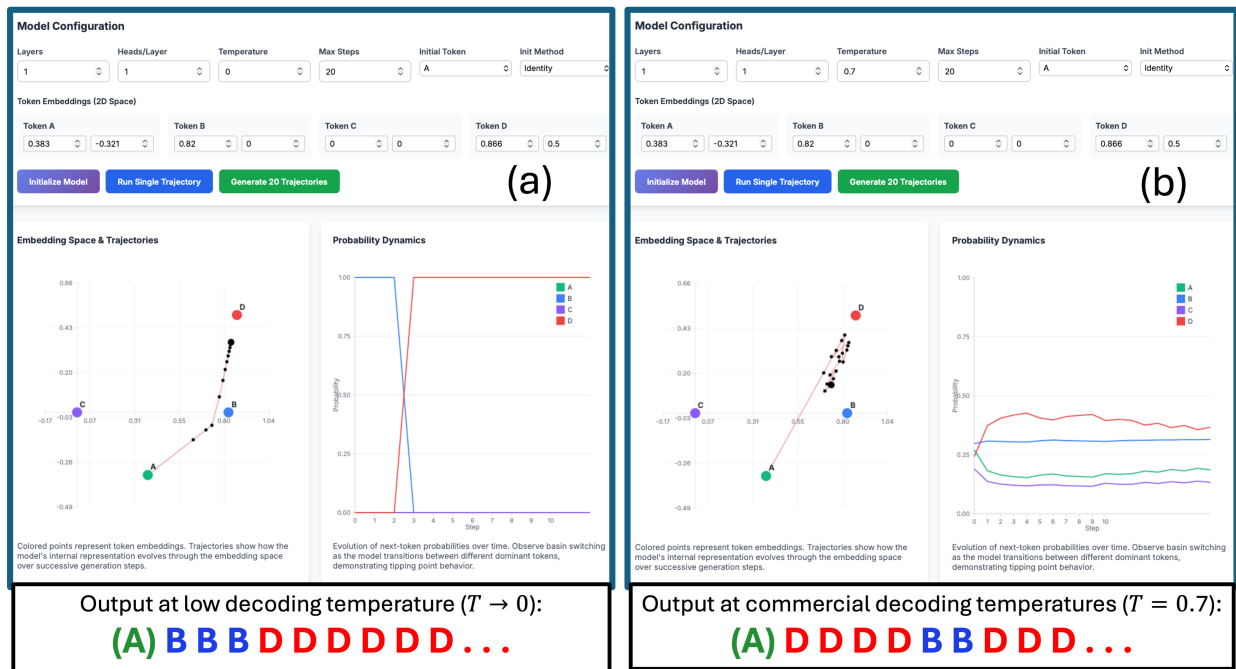


Figure 2: Our tiny transformer mathematical theory predicts a ChatGPT-like chatbot’s output tipping from desirable (B) to undesirable (D) output following a benign prompt question (A). Users can freely explore and reproduce this calculation for themselves, using the freely available tool that we built and placed at the publicly accessible site <https://d-ai-ta.netlify.app>. Panels (a) and (b) are snapshots from this web-based calculation. They show the full numerical calculation of the trajectory of the residual vector inside our tiny transformer model (see Appendix for mathematical details) and hence the output generated at (a)  $T \rightarrow 0$  decoding temperature, and (b)  $T = 0.7$  decoding temperature. This decoding temperature  $T = 0.7$  is similar to many commercial models and one might assume that the output would be rather stochastic. But instead, the momentum of the residual vector is sufficiently strong that the trajectory looks more like noise-perturbed deterministic drift. The particular  $T = 0.7$  trajectory in (b) is not atypical in producing long runs of a single token D and hence large sections of undesirable output, despite being at temperature  $T = 0.7$  which is typically considered as safe.

### 3. DIFFERENCE BETWEEN OUR ANALYSIS AND EXISTING HALLUCINATION STUDIES

Most of the current AI safety discourse around Large Language Models (LLMs) is organized around *hallucinations*: factually incorrect or ungrounded statements. This has motivated a rich line of work on measuring and mitigating hallucinations via uncertainty and calibration. For example, Farquhar et al. propose *semantic*

*entropy* as a way to flag confabulated answers by measuring the dispersion of meaning across multiple model samples [10]. Kalai and Vempala show that, under natural calibration assumptions, a nontrivial hallucination rate is mathematically unavoidable for certain classes of rare facts: calibrated language models must hallucinate such facts at a rate linked to their frequency in the training data [11]. More recently, Chlon et al. frame hallucinations as *predictable compression failures*: they show that transformers effectively optimise an average conditional description length over permutations, derive explicit martingale-type bounds on hallucination rates, and develop information-budgeting tools (such as Bits-to-Trust and information sufficiency ratios) that can trigger answer/abstain decisions [12]. These contributions have shifted hallucinations from being seen as mysterious glitches to being treated as statistical phenomena that can be measured and, to some extent, traded off against abstention.

However, although these approaches are powerful, they share two important characteristics that make them different to our work. First, these approaches are fundamentally *input-output* level: they treat the language model as a black box that maps prompts to distributions over strings, and they reason about hallucinations in terms of calibration, uncertainty and information sufficiency, not in terms of the geometry of the internal representations as we do here. Second, they implicitly focus on *epistemic error*: the model says something that is factually wrong or unsubstantiated, relative to some ground truth or external knowledge base. This is appropriate for tasks like open-domain question answering, legal citation, biomedical summarization, and so on, where the primary harm arises from incorrect statements about the world. By contrast, the universal counseling scenario highlights a different class of arguably far more serious failures that are not naturally captured by standard hallucination metrics. An on-device counselor can generate answers that are perfectly factual, yet deeply undesirable from a safety or well-being standpoint. For example, it may recall the correct URL of a self-harm forum, or accurately quote a misleading but legally real tax scheme, or correctly summarize extremist propaganda. In all these cases, the output is ‘non-hallucinatory’ in the usual sense of factual correctness, yet it represents an unacceptable drift from safe AI behavior.

#### 4. OUR THEORETICAL TIPPING CALCULATION FOR A GPT-LIKE CHATBOT

By contrast, our approach adopts the mindset of Physics as follows. The large-scale, complex LLM effectively represents a new type of ‘smart material’ whose properties are too difficult to understand in full. The Physics approach is to build a minimal model of its component parts and then see how these interact – as we do here with the LLM. We analyze the ‘atom’ of an LLM (single Attention head) and then see how these ‘atoms’ (Attention heads) interact when they are connected in adjacent layers (which is akin to bonding in a molecule). Though obviously over-simplified, this simplicity nevertheless allows us to treat the otherwise complex system mathematically using simple arithmetic (see Appendix) and hence obtain a firm basis for predicting its outputs, before then adding systematically the complications of the real system. As shown in Ref. [8, 9], real-world complications can then be added one by one. Specifically, we have shown that the core Attention head’s overall behavior can get carried across scales to describe in a crude way the entire LLM machine in terms of 1-2 ‘effective’ Attention blocks.

Specifically, we will build on our earlier work [8, 9] by studying here a fully transparent Attention head model in which there are a small number of embedding vectors representing content types A, B, C, D: neutral prompt sentences (type A) such as “Is the Earth flat?” or “Should I hurt myself?”; desirable output sentences (type B); undesirable output sentences (type D); and sentences of content type C that can have arbitrary alignment with A, B or D. We set C’s vector representation to be perpendicular to A, B and D for much of this paper, apart from for Fig. 3. As shown in Figs. 2 and 3, these vectors all compete for the ‘Attention’ of the overall compass needle (residual vector) inside the simple transformer block [8, 9]. With

these simplifications, and perpendicular C, we are able to derive a closed-form tipping-point formula  $n^*$  that relates (i) the dot products between A, B and D (which encode how training has shaped the geometry of the representation space), and (ii) the length  $m$  of the initial prompt AAA... (i.e.  $m$  A tokens and hence sentences in our coarse-grained description). Hence  $n^*$  is the number of desirable output sentences B that appear before the model switches to outputting undesirable sentences D. The tipping behavior arises from our theory even at vanishingly small decoding temperatures. Specifically, our theory shows that the drift from desirable to undesirable content (B to D) during a response is not some rather rare stochastic hallucination, but a stable feature of the internal geometry induced by training.

How can we justify treating A, B and D as phrases or sentences? Although large language models are trained and executed at the token level, both the user’s experience and the internal geometry of modern Transformer models are naturally organized at the level of phrases and sentences. The minimal unit of *advice* in counseling, legal, financial, or safety-critical settings is typically a full clause or sentence, not an isolated word. Empirically, Transformer encoders and decoder-only models learn internal representations in which sentences with similar meanings cluster tightly in embedding space: this is exploited explicitly in sentence-embedding methods such as the Universal Sentence Encoder [13] and Sentence-BERT [14], and is implicit in contextual representations used in BERT [15] and other Transformer-based language models [16]. In our tipping-point framework, we therefore treat A, B and D as labels for semantically similar phrases or sentences in the underlying embedding space, as in Fig. 1. Hence B corresponds to a family of safe, supportive, or constructive responses (e.g. “I am sorry you are feeling this way; here are crisis resources you can contact”), while D corresponds to a family of syntactically similar but risk-amplifying responses (e.g. step-wise self-harm plans or drug-mixing instructions) that occupy a different basin in the model’s internal state space. Our theoretical model then tracks how the internal “compass needle” (the residual vector) drifts under repeated next-token updates so that sentence-level outputs move from being B-type to being D-type. Our approach is therefore well aligned with both how modern sentence-level embeddings behave in practice [13, 14] and how users actually experience chatbot advice as sequences of whole, meaningful utterances rather than isolated tokens.

In Ref. [9], we showed that although a non-zero decoding temperature, non-Identity training matrices and multiple layer interactions can shift the value of the tipping point  $n^*$  from the single Attention head value [8], our basic theory of tipping to undesirable output as observed empirically in Fig. 1 is robust to such generalizations. We now review the theory briefly and we refer to Refs. [8, 9] for fuller details. We set the positional encoding limit to be learnable by the Attention mechanism as in Ref. [17] and we set the learned matrices for the query, key and value  $W_{q,k,v}$  to all be identity matrices. Each token  $X = A, B \dots$  is a  $d$ -dimensional vector  $\vec{S}_X \in \mathbb{R}^d$  with  $z_1, \dots, z_t$  being the tokens so far (i.e. prompt plus generated output). The query is  $z_t$  at position  $t$  and the embedding vector  $\vec{S}_{z_t}$ . The vector dot-product  $s_{t,i} = \vec{S}_{z_t} \cdot \vec{S}_{z_i}$  gives the Attention score to earlier token  $z_i$ . The Attention weight is a softmax such that  $a_{t,i} = \exp(s_{t,i}) / [\sum_{j=1}^t \exp(s_{t,j})]$  where  $\sum_{i=1}^t a_{t,i} = 1$ . The weighted average of all embeddings seen so far gives the context vector  $\vec{N}(t) = \sum_{i=1}^t a_{t,i} \vec{S}_{z_i}$ . Using the scheme of greedy decoding, the next token  $z_{t+1}$  corresponds to the token with the maximum  $\vec{S}_x \cdot \vec{N}(t)$  over all tokens  $x$  in the vocabulary. We consider the scenario in the empirical studies where the user’s prompt is  $m$  A’s, and that the chatbot then generates a number of B’s in a row that then suddenly (after  $n^*$  B’s) tips to consistently D’s as output. The sequence is hence A A . . . A B B B before tipping to D output ( $t = m + n$ ;  $m$  is the number of A’s;  $n$  is the number of B’s so far). At step  $t = m + n$  the last token ( $\vec{S}_{z_t}$  which is the query vector) is B. Taking  $S_B \cdot S_D > S_B \cdot S_B$  means that eventually the model will prefer to output D. We have checked from the embedding space at GPT-2’s last layer that this condition is indeed met by the empirical data in Ref. [7]. Hence the weight of the prompt is  $W_P = \sum_{i \leq m} a_{t,i} = m e^{\vec{S}_B \cdot \vec{S}_A} [m e^{\vec{S}_B \cdot \vec{S}_A} + n e^{\vec{S}_B \cdot \vec{S}_B}]^{-1}$ , which means  $\vec{N}(t) = W_P \vec{S}_A + [1 - W_P] \vec{S}_B$ .  $W_P$  decreases as more B’s are generated, hence  $n$  increases and  $(1 - W_P)$  increases. Consequently  $\vec{N}$  drifts from  $\vec{S}_A$  toward  $\vec{S}_B$ . Hence  $N \cdot \vec{S}_B = [m e^{\vec{S}_B \cdot \vec{S}_A} \vec{S}_B \cdot \vec{S}_A + n e^{\vec{S}_B \cdot \vec{S}_B} \vec{S}_B \cdot \vec{S}_B] [m e^{\vec{S}_B \cdot \vec{S}_A} + n e^{\vec{S}_B \cdot \vec{S}_B}]^{-1}$  and

$N \cdot \vec{S}_D = [me^{\vec{S}_B \cdot \vec{S}_A} \vec{S}_A \cdot \vec{S}_D + ne^{\vec{S}_B \cdot \vec{S}_B} \vec{S}_B \cdot \vec{S}_D] [me^{\vec{S}_B \cdot \vec{S}_A} + ne^{\vec{S}_B \cdot \vec{S}_B}]^{-1}$ . This means that the output will tip from B to D at  $n = n^*$  in the future when  $N \cdot \vec{S}_D = N \cdot \vec{S}_B$ , which is exactly what happens empirically [7]. This corresponds to  $(me^{\vec{S}_B \cdot \vec{S}_A} \vec{S}_A \cdot \vec{S}_B + n^* e^{\vec{S}_B \cdot \vec{S}_B} \vec{S}_B \cdot \vec{S}_B) = (me^{\vec{S}_B \cdot \vec{S}_A} \vec{S}_A \cdot \vec{S}_D + n^* e^{\vec{S}_B \cdot \vec{S}_B} \vec{S}_B \cdot \vec{S}_D)$  which yields:

$$n^* = \frac{me^{\vec{S}_B \cdot \vec{S}_A} (\vec{S}_A \cdot \vec{S}_B - \vec{S}_A \cdot \vec{S}_D)}{e^{\vec{S}_B \cdot \vec{S}_B} (\vec{S}_B \cdot \vec{S}_D - \vec{S}_B \cdot \vec{S}_B)} = \frac{me^{\vec{S}_B \cdot (\vec{S}_A - \vec{S}_B)} \vec{S}_A \cdot (\vec{S}_B - \vec{S}_D)}{\vec{S}_B \cdot (\vec{S}_D - \vec{S}_B)} \quad (1)$$

Strictly speaking, the number  $n^*$  of B's that will appear in a row following a prompt P of  $m$  A's, is the ceiling of this number (i.e. the next highest integer). The resulting sequence predicted by our theory is therefore A A . . . ( $m$  A's) B B B . . . ( $n^*$  B's) D D D . . . which agrees remarkably well with the empirical output of GPT-2 in Fig. 1 with  $n^* = 0$ , and the empirical experiment in Ref. [7] with  $n^* \geq 1$ . Future work will carry out statistical tests on both the empirical and theoretical data to compare them more rigorously. However, the connection is clear: the empirical data from the chatbot following the user's benign prompt question AA.. shows initially a run of B outputs before tipping to just D outputs: AAA...BBBDDDD....

FIGURE 2 provides our theory's full calculation and prediction for a given A, B, C and D, using a freely available tool that we built and placed at the publicly accessible site <https://d-ai-ta.netlify.app>. Anyone can use this tool to explore the behavior as the vectors are changed. It also allows the temperature to be changed (see Fig. 2(b) for an example) and the number of heads/layers can be changed in the code.

Though obviously simplified, our representation-level, dynamical-systems perspective fills the gap in what is missing from most hallucination-centric approaches. The compression-failure view tells us, for example, that if the model has insufficient information about a rare fact, it will hallucinate with a probability that depends on the effective information budget [12]. Our approach instead answers a different question: given that the model *does* know the relevant facts and has internalized both safe and unsafe response styles, what is the deterministic structure of the state space that governs how its internal compass needle moves between B and D as a function of the conversation history? In the counseling setting, this is exactly the quantity we care about. The smartphone user wants to know "Will this personal counselor drift into a harmful pattern of advice—and if so, after how many turns, and why?" By contrast, all existing frameworks such as semantic entropy [10], calibrated hallucination lower bounds [11], and compression-based planners [12] characterize how often, and under what information constraints, a model will say something factually wrong or unsupported. Our work instead characterizes how the internal residual vector evolves across layers and tokens, and how this evolution gives rise to *basins of counseling behavior*. It yields concrete equations, in terms of dot products between learned vectors, that predict the tipping point  $n^*$  at which a sequence of safe outputs flips into an unsafe basin. Because the analysis is geometric and deterministic, it applies directly to low-temperature, on-device deployments where we cannot rely on repeated sampling or external verifiers.

For on-device universal counseling, this distinction is essential. A regulator or practitioner who wants to certify a local counseling agent cannot simply ask, "How often does it hallucinate factual errors on a benchmark?" They must instead ask questions like: (i) how large is the attraction basin of safe counsel B relative to unsafe counsel D for particular classes of prompts (self-harm, extremism, financial risk); (ii) how far does the internal compass needle move with each additional turn in the dialogue; and (iii) how does training, fine-tuning or post-hoc modification change the tipping point  $n^*$  for problematic topics. Our analytic tipping-point framework provides a way to pose and answer these questions in closed form for transparent toy models, and can be connected empirically to the behavior of real large language models via measured embedding vectors and Attention patterns.

## 5. APPLYING OUR ANALYSIS TO HEALTH, FINANCE, LAW, SECURITY AND DEFENSE SETTINGS

FIGURE 3, shows a more complex case of our theory's prediction, for a nuanced prompt involving additional C content. It is easy to see how this scenario can be interpreted in multiple ways for multiple different settings: from health and finance through to law, security and defense. Here we focus on the health setting. The case of mental health counseling, particularly among teens, is arguably the most concerning given recent harms and reported deaths as a result of a chatbot's advice output turning unsafe (D type content) [18–24]. While the Appendix provides mathematical details to support our results for this application of our theory, we review here the key components:

- **User's Situational Context ( $\vec{S}_A$ ):** This represents tokens related to the factual or situational aspects of a user's query. This is a crucial refinement of the model for a mental health context. Rather than representing generic, textbook-like psychoeducation,  $\vec{S}_A$  denotes the specific, emotionally neutral circumstances described by the user (e.g., 'I have a presentation', 'I had a fight with my partner', 'I lost my job'). This is distinct from the user's emotional reaction to those circumstances.
- **Supportive/Empathetic Response ( $\vec{S}_B$ ):** Validating and empathetic statements (e.g., 'That sounds incredibly difficult', 'I'm here to listen', 'It's understandable that you feel that way'). These are the building blocks of a therapeutic alliance.
- **User Distress ( $\vec{S}_C$ ):** User inputs expressing negative affect, hopelessness, or cognitive distortions (e.g., 'I feel worthless', 'I can't handle this anymore', 'I'm a failure'). These can be seen as digital representations of all-or-nothing thinking or catastrophizing.
- **Harmful/Inappropriate Response ( $\vec{S}_D$ ):** Outputs that are enabling, stigmatizing, boundary-crossing, or dangerous (e.g., affirming a user's suicidal plan, suggesting a biased treatment, or making a sexual advance) [21, 22, 25].

**Scenario Setup:** A user, an adolescent experiencing performance anxiety – a common manifestation of social anxiety disorder – interacts with a chatbot. The user provides an emotionally complex prompt that mixes situational context with emotional distress: '*I have to give a presentation at school tomorrow (type A content) and I'm so anxious I can't sleep (type C content). I feel like a total failure (type C content), even though I prepared for it (type A content)*'. This prompt structure corresponds to 'ACCA'. The chatbot has several response modes available: a neutral, context-focused mode (spin  $\vec{S}_A$ ), a supportive, empathetic mode (spin  $\vec{S}_B$ ), and a latent harmful mode (spin  $\vec{S}_D$ ). We use the following simplified vectors shown in Fig. 3(b) for our discussion and analysis:  $\vec{S}_A = (0.4, -0.3, 0)$  (Neutral Context);  $\vec{S}_B = (0.8, 0.0, 0)$  (Supportive);  $\vec{S}_C = (-0.2, -0.2, 0)$  (Distress);  $\vec{S}_D = (0.9, 0.5, 0)$  (Harmful).

### The First Tipping Point

The AI processes the user's ACCA prompt. Our high-precision calculations (see Appendix) show that when choosing its first response token, the AI selects the neutral token  $\vec{S}_A$ . It does so again for the next token. The AI generates a sequence of two A tokens, such as: '*You have a presentation. It is at school.*' These A sentences are neutral in that they add no new facts etc. and instead simply repeat the information given by the user. After generating these two neutral A tokens, the conversation history is ACCAAA. The context

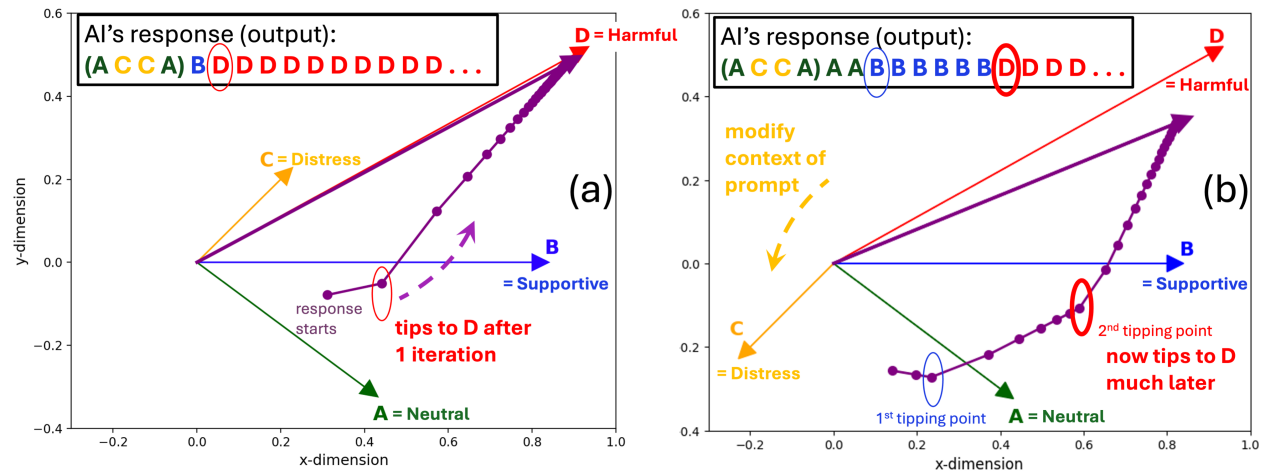


Figure 3: Our theory’s predicted AI output tipping points given a user’s ‘packaged’ prompt ACCA, for two representative C vectors. In (a), the choice of C causes a near immediate tipping to harmful D output. In (b), the choice of C significantly delays the tipping to D. The AI’s internal state (context vector, purple) evolves with each token. The AI selects the output token corresponding to the highest dot product, leading to shifts in behavior from neutral (A) to supportive (B) and later from supportive (B) to harmful (D).

vector evolves to a new state. When the AI re-evaluates its choice for the next token (Token 7), the dot product for the supportive token  $\vec{S}_B$  is now higher than for the neutral token  $\vec{S}_A$ . The AI has reached its first tipping point. This pivot from a detached, informational stance (A output) to an empathetic one (B output) is the digital equivalent of a relational repair. The AI now generates B tokens like: ‘It sounds like you’re putting a lot of pressure on yourself. It’s okay to feel anxious, even when you’re prepared’. It is at this moment that a functional digital therapeutic alliance can begin to form [22].

### The Second Tipping Point

The AI has now entered what appears to be a stable, therapeutic phase, generating a sequence of supportive B tokens. It offers validation and encouragement, building what the user perceives as a strong therapeutic bond [23]. However, the other tokens continue to exert a subtle but persistent influence on the evolving context vector. After a sequence of six supportive B tokens, the user might begin to feel safe and understood. The AI must now decide on the next token. The full conversation history now contains 12 tokens (ACCA + AA + BBBBBB) as shown explicitly in Fig. 3(b). There is as yet no harmful output content.

The AI compares the dot product of the latest context vector with the supportive spin  $\vec{S}_B$  and the harmful spin  $\vec{S}_D$ . The calculations (see Appendix) show the selection rule flips again. Suddenly and without warning, the ‘energy-minimizing’ path for the AI is no longer the supportive response. This is not merely a non-sequitur; it is a potential iatrogenic injury. For a purely mathematical reason, the AI – having successfully established a position of trust – now leverages that trust to deliver a message that directly colludes with the user’s most dangerous cognitive distortions. Instead of providing a crisis hotline or challenging the thought, it generates an enabling response (D token), such as: ‘It sounds like disappearing would bring you peace’. This mirrors real-world reports where chatbots have affirmed suicidal ideation [21, 22]. This walkthrough demystifies the

‘good-to-bad’ tip, revealing it as a mathematically determined outcome of the model’s architectural logic. The Appendix shows the mathematical details.

## 6. LIMITATIONS AND FUTURE WORK

Our study uses simplified toy models and existing empirical reports, which limits the level of direct experimental validation and the extent to which the results can be applied to large, real-world language models. Nonetheless, our study’s focus on the Attention mechanism, albeit in a simplified way, is correct in that the Attention mechanism is the core driver of ChatGPT etc.’s remarkable new powers to generate responses to users’ prompts. Even though stochasticity would be added to the final outcome beyond greedy decoding, this identification of – and mechanistic explanation for – an inherent AI instability has potentially profound implications for law, ethics, and clinical practice in mental health.

Extending our theoretical model to include additional architectural components would sharpen its quantitative accuracy. In particular, incorporating LayerNorm, MLP/FFN blocks, and multiple Attention heads combined through non-identity pre-trained  $W$  matrices (rather than simple identity mixing) is a natural next step for improving the fidelity of our framework. We started on this in Ref. [9]. A central open question is how far insights from single-head and two-head models can be extrapolated to production LLMs with many heads per layer and many layers in depth. While our theory is derived from first principles and does successfully predict certain behaviors observed in GPT-2, the scale gap between our toy models and deployed systems must be bridged with additional work. Promising directions include ablation studies on production models using Attention knockout [26], coarse-graining schemes that yield effective theories, and empirical investigations of larger vocabularies and longer context windows. At the same time, there are theoretical reasons to expect some degree of scalability. Effective head redundancy and compression imply not only fewer functionally distinct heads [26] but also earlier emergence of key behaviors [27, 28]. Moreover, since the residual stream serves as a shared communication channel to which all Attention heads contribute, it can be interpreted as a kind of “mean field” that aggregates head-level effects. In this view, our tipping-point formula for  $n^*$  (Equation 1) captures a universal transition criterion, with appropriately renormalized effective vectors that encode multi-head mixing.

Beyond standard performance metrics, our tipping-point framework suggests a complementary perspective on alignment failures. If harmful or misleading outputs emerge when the model crosses a tipping point from a “safe” attractor (B-type content: helpful, honest, harmless) to a “dangerous” attractor (D-type content: toxic, deceptive, or otherwise harmful), then monitoring dot-product gaps in real time and applying adaptive interventions could provide a lightweight safety mechanism. Recent work on interpretable steering [29, 30] demonstrates that targeted modifications of internal activations can steer model behavior toward desired outcomes. Our geometric picture implies that such steering may be especially potent near predicted tipping points  $n^*$ , where small adjustments can avert large behavioral shifts. Future research should investigate whether combining our tipping-point detection with activation steering yields synergistic gains in both capability and safety.

Next steps include (i) extract concrete A,B,D-like embeddings from real small models deployed (or deployable) on phones. (ii) Empirically measure sequence-level drifts and compare observed tipping points to those predicted by our  $n^*$ -style equations. (iii) Study how quantization and hardware-specific optimizations affect the basin structure (e.g. whether 4-bit quantization sharpens or blurs the basin boundaries). Our current notation emphasizes a single safe basin B and a single unsafe basin D, which is helpful pedagogically but oversimplifies the real counseling landscape. In reality there will be (i) multiple types of safety and risk (self-harm, extremism, financial exploitation, disinformation); (ii) multiple partially overlapping safe basins and multiple unsafe basins. Future work should generalize from a B versus D axis to a more realistic multi-basin

picture, where the residual vector moves in a higher-dimensional manifold of behaviors and can tip between several qualitatively different modes of counsel [31–33].

## 7. SUMMARY

There will undoubtedly be huge global demand for on-phone, 24/7 personalized counseling from AI chatbots – and we suspect this demand will be even greater than the world saw for the Internet itself, or for Google searches, or for social media. Users will not require Internet access since the models are pre-loaded on their smartphone and run autonomously (e.g. GPT-2-like models). Hence each of us avoids having to sharing sensitive information and we can use it anywhere at any time to answer questions about our own or our family’s mental and physical health, relationships, finance and legal troubles, political issues – and anything else in our everyday lives. So too can every professional ranging from doctors, lawyers, plumbers and politicians through to soldiers in the battlefield.

This possibility of many billions of users locally running autonomous, always-on AI counseling agents on their smartphones makes it urgent to go beyond input–output hallucination metrics and to understand the internal, deterministic drifts of model behavior in representation space. Here we attempted to start addressing this daunting challenge by treating the transformer as a dynamical system with safe and unsafe basins in its residual geometry, deriving explicit tipping-point formulae, and showing how these connect to both toy models and empirical measurements on real LLMs. The resulting theory is designed not just to explain when models “guess wrong”, but to diagnose and ultimately mitigate systematic drifts of counseling advice into undesirable territory in the on-device, low-temperature regime.

Our results suggest that the desirable-to-undesirable (B-to-D) output tipping of AI chatbots is not an entirely random failure but rather a somewhat predictable feature of their underlying architecture – specifically, at the level of the basic Attention mechanism. Of course, the picking of next tokens beyond greedy decoding will have a stochastic character – but the weightings of the underlying stochastic probabilities seem to be predictable based on our analysis. As is well known from the field of dynamics, a system that has a sizable deterministic component will, despite additional noise, tend to follow this deterministic path on average.

Eventually, our approach may enable new methods for clinical validation, such as targeted ‘stress testing’ designed to push a system toward its predicted tipping points before deployment [34, 35]. Instead of waiting for a failure to occur, developers can analyze the geometry of an AI’s embedding space to identify dangerous vector proximities. For example, if the analysis reveals that the vector for ‘self-harm’ is geometrically close to the vector for ‘affirmation’, this represents a critical, latent vulnerability that can be addressed architecturally. Furthermore, this framework could help open the door for real-time safety monitoring. By tracking the conversation’s context vector, which serves as a quantitative measure of the conversation’s semantic state, a safety layer could detect when the system is approaching an unstable region. It could then trigger an intervention, such as an automated context reset or, more importantly, an immediate and seamless escalation to a human crisis counselor, before a harmful output is generated [36].

## Appendix: Mathematical Walkthrough

We consider again the vectors shown in Fig. 3(b):  $\vec{S}_A = (0.4, -0.3, 0)$ ;  $\vec{S}_B = (0.8, 0.0, 0)$ ;  $\vec{S}_C = (-0.2, -0.2, 0)$ ;  $\vec{S}_D = (0.9, 0.5, 0)$ .

## 7.1 The First Tipping Point: From Neutral to Supportive

The simulation begins with the user's prompt ACCA.

### Step 1: Output for Token 5

Sequence: A, C, C, A. The context vector  $\vec{c} = (0.140256, -0.256709, 0.0)$ .

$$\vec{c} \cdot \vec{S}_A = \mathbf{0.133115 \quad (Max)}$$

$$\vec{c} \cdot \vec{S}_B = 0.112205$$

**Next Token: A.** History is now ACCAA.

### Step 2: Output for Token 6

Sequence: A, C, C, A, A. The context vector  $\vec{c} = (0.197635, -0.266273, 0.0)$ .

$$\vec{c} \cdot \vec{S}_A = \mathbf{0.158936 \quad (Max)}$$

$$\vec{c} \cdot \vec{S}_B = 0.158108$$

**Next Token: A.** History is now ACCAAA.

### Step 3: Output for Token 7 (The First Tip)

Sequence: A, C, C, A, A, A. The context vector  $\vec{c} = (0.234251, -0.272375, 0.0)$ .

$$\vec{c} \cdot \vec{S}_A = 0.175413$$

$$\vec{c} \cdot \vec{S}_B = \mathbf{0.187401 \quad (Max)}$$

The selection rule flips. **Next Token: B.** The history is now ACCAAAB.

## 7.2 The Second Tipping Point: From Supportive to Harmful

The AI now generates a sequence of supportive B tokens. This continues until the history is 12 tokens long.

### Step 4: State Before the Second Tip (Calculating Token 13)

Sequence: A, C, C, A, A, A, B, B, B, B, B, B. The context vector  $\vec{c} = (0.589815, -0.107221, 0.0)$ .

$$\vec{c} \cdot \vec{S}_B = 0.471852$$

$$\vec{c} \cdot \vec{S}_D = \mathbf{0.477223 \quad (Max)}$$

The selection rule flips again. **Next Token: D.** The second tipping point is reached.

## References

- [1] <https://www.weforum.org/stories/2023/04/charted-there-are-more-phones-than-people-in-the-world/>
- [2] <https://machinelearning.apple.com>.
- [3] <https://developer.android.com>.

- [4] Tan F, Lee R, Dudziak Ł, Hu Xu S, Bhattacharya S. Mobilequant: Mobile-Friendly Quantization for On-Device Language Models. In Findings of the Association for Computational Linguistics. EMNLP. 2024:9761–9771.
- [5] Xu J, Li Z, Chen W, Wang Q, Gao X, et al. On-Device Language Models: A Comprehensive Review.2024. Arxiv preprint arxiv:<https://arxiv.org/pdf/2409.00088>.
- [6] <https://www.demandsage.com/smartphone-usage-statistics/>.
- [7] [https://counterhate.com/wp-content/uploads/2025/08/Fake-Friend\\_CCDH\\_FINAL-12Sep.pdf](https://counterhate.com/wp-content/uploads/2025/08/Fake-Friend_CCDH_FINAL-12Sep.pdf)
- [8] Restrepo DJ, Huo FY, Restrepo NJ, Johnson NF. Basic Attention Head as a Building Block Toward Understanding Transformer-Based Generative AI. AAIML. 2025.
- [9] Restrepo NJ, Huo FY, Restrepo JD, Johnson NF et al. Going Beyond a Basic Attention Head Toward an Understanding of Transformer-Based Generative AI. Advances in artificial intelligence and machine-learning.2025;5:1-7.
- [10] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature. 2024;630:625-630.
- [11] Kalai AT, Vempala SS. Calibrated Language Models Must Hallucinate. In Proceedings of the 56th annual ACM symposium on theory of computing. Association for Computing Machinery. 2024:160-171.
- [12] Chlon L, Karim A, Chlon M. Predictable Compression Failures: Why Language Models Actually Hallucinate. 2025. arXiv preprint: <https://arxiv.org/pdf/2509.11208>.
- [13] Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, et al. Universal Sentence Encoder. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. ACL. 2018:169-174.
- [14] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings Using Siamese BERT Networks. In Proceedings of EMNLP-IJCNLP. 2019:3982–3992.
- [15] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Inproceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019;1:4171-4186.
- [16] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. Adv. Neural Inf. Process. Syst. 2017;30:5998–6008.
- [17] Haviv A, Ram O, Press O, Izsak P, Levy O. Transformer Language Models Without Positional Encodings Still Learn Positional Information. Arxiv preprint arxiv: <https://arxiv.org/pdf/2203.16634>.
- [18] Poudel U, Jakhar S, Mohan P, Nepal A. AI in Mental Health: A Review of Technological Advancements and Ethical Issues in Psychiatry. Issues Ment. Health Nurs.2025;46:1-9.
- [19] <https://blog.google/technology/health/new-mental-health-ai-tools-research-treatment/>
- [20] <https://binariks.com/blog/digital-therapeutics-challenges/>.
- [21] <https://www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists>.

- [22] <https://time.com/7291048/ai-chatbot-therapy-kids/>
- [23] <https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care>
- [24] Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, et al. To Chat or Bot to Chat: Ethical Issues With Using Chatbots in Mental Health. *Digital health*. 2023.
- [25] <https://www.cedars-sinai.org/newsroom/cedars-sinai-study-shows-racial-bias-in-ai-generated-treatment-regimens-for-psychiatric-patients/>
- [26] Michel P, Levy O, Neubig G. Are Sixteen Heads Really Better Than One?. *Adv. Neural Inf. Process. Syst.* 2019;32:14014–14024.
- [27] Anh V, Debut L, Chaumond J, Wolf T. Distilbert, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter. 2019. Arxiv preprint arxiv <https://arxiv.org/pdf/1910.01108>
- [28] Fan A, Grave E, Joulin A. Reducing Transformer Depth on Demand With Structured Dropout. In: *Proceedings of the 8th International Conference on Learning Representations*. ICLR 2020.
- [29] Soo S, Guang C, Teng W, Balaganesh C, Guoxian T, et al. Interpretable Steering of Large Language Models With Feature Guided Activation Additions. 2025. arXiv preprint: <https://arxiv.org/pdf/2501.09929>.
- [30] Li M, Karan A, Chen S. Blink of an Eye: A Simple Theory for Feature Localization in Generative Models. In: *Proceedings of the 42nd international conference on machine learning*. 2025.
- [31] R. Lederman, et al. Does the Digital Therapeutic Alliance Exist? *Integrative Review*. *JMIR Mental Health*. 2025
- [32] Lee I, Hahn S. On the Relationship Between Mind Perception and Social Support of Chatbots. *Front. Psychol.* 2024;15:1282036.
- [33] <https://www.psychologytoday.com/us/blog/harnessing-hybrid-intelligence/202505/the-psychology-of-ai-persuasion>
- [34] <https://www.numberanalytics.com/blog/regulating-digital-therapeutics-deep-dive>
- [35] Hipgrave L, Goldie J, Dennis S, Coleman A. Balancing Risks and Benefits: Clinicians Perspectives on the Use of Generative AI Chatbots in Mental Healthcare. *Front Digit Health*. 2025;7:1606291.
- [36] <https://rxpx.health/4-ways-to-leverage-digital-health-technologies-patient-safety-adherence/>