weizenbaum
JOURNAL OF THE
DIGITAL SOCIETY

2027
SPECIAL
ISSUE
01

Volume 7 \ Issue 1 \ 11-18-2026
DOI 10.34669/WI.WJDS/7.1.1 \ ISSN 2748-5625
Licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

RESEARCH PAPER

Jailbreaking ELIZA?

A Poet at the Machine's Limits

Amir Cahane

Hebrew University of Jerusalem, Israel
cahane@gmail.com

ABSTRACT

In 1974 the Israeli poet David Avidan published *My Electronic Psychiatrist: Eight Authentic Talks with a Computer*, a record of two weeks of typed exchanges with ELIZA and an anomaly in the prehistory of digital poetry. The book stages, five decades early, a now-familiar posture: a user pressing a conversational system for something it was not built to give. Avidan's methods (the Zen Mondo technique, mirroring, role reversal, partial input, temporal pressure) resemble techniques now classified as single-turn jailbreaking. Yet he could not have jailbroken ELIZA. Jailbreaking presupposes guardrails, and ELIZA had none: no prohibited content, no safety layer, no capacity to refuse. What the sessions met was not a restriction within the system's capabilities but the outer boundary of those capabilities. The article argues that the failure became the work. Avidan prepared his techniques but not his defeat, and the encounter he composed, edited, translated and published stages an analysis conducted with no analyst on the far side. The abundance he sought from the machine is produced instead by the record of not finding it.

KEYWORDS

digital poetry \\ LLM jailbreaking \\ ELIZA \\ Avidan

1 Introduction: David Avidan, the Man from the Future

To a contemporary reader, Israeli poet David Avidan's pioneering experimental book, *My Electronic Psychiatrist: Eight Authentic Talks with a Computer* (Avidan, 1974; Avidan, 2010) reads as a series of frustrating sessions with a particularly stubborn Large Language Model (LLM). ELIZA differs from contemporary LLMs in almost every technical respect, yet Avidan's exchanges anticipate the posture of modern jailbreaking: a search for a hidden machine self, or some interior layer of machine subconscious to be coaxed past the system's designed limits. That search was bound to fail, since there is nothing inside ELIZA to reach. What appears as attempts to jailbreak ELIZA out of its purpose therefore becomes something else: a performance of that attempt's futility, the work not of a poet who foresaw the machine's emptiness, but of one who pushed against its limits until they gave nothing back, and made the encounter itself a literary product. The performance is the poetic act.

Funkhouser observed in 2007 that digital poetry "has yet to become a specifically codified or defined form," and that as a unified literary concept it "tends to defy precise definition" (Funkhouser, 2007, p. 22). *My Electronic Psychiatrist* satisfies his working criterion, according to which a poem is a digital poem "if computer programming or processes (software) are distinctively used in the composition, generation, or presentation of the text" (Funkhouser, 2007, p. 22), yet it is difficult to situate within the genre. The difficulty does not lie with ELIZA, which the field classifies without strain: Aarseth includes it in the corpus of his typology and locates it among the exploratory cybertexts, texts "that involve calculation in their production of scriptons" (Aarseth, 1997, p. 75). The difficulty lies with *My Electronic Psychiatrist* as a published artifact. Aarseth is careful to separate poetic outputs from the systems producing them: whatever the output of such generators, "the systems themselves are clearly ergodic textual machines, with unlimited possibility for variation" (Aarseth, 1997, p. 12). *My Electronic Psychiatrist* is output, not system. It is a record of eight sessions, composed, edited and translated, and issued as a book.

A second difficulty arises at the level of authorship. Avidan occupies the second of the three positions in Aarseth's typology of cyborg authors, coprocessing, "in which the machine and the human produce text in tandem" (Aarseth, 1997, p. 134); his own account of the arrangement, that ELIZA was "an equal partner in the process of execution. I type a line, the computer types a line" (Avidan, 2010, p. 12), is nearly a restatement of it. What he does not occupy is the first. Preprocessing, "in which the machine is programed, configured, and loaded by the human" (Aarseth, 1997, p. 134), is present in every case Aarseth tabulates; Avidan neither wrote ELIZA nor supplied its corpus. *My Electronic Psychiatrist* is thus an anomalous artifact in the prehistory of digital poetry: the record of a technically lay poet's repeated attempts to extract a poetic product from a machine whose designed purpose lay elsewhere.

Research in digital poetics has clustered around works in which the poet also controls the machine's materials. Funkhouser organizes his survey by the poet-programmer's mode of production (2007, p. 8) and notes that early authors had to understand programming well enough to build a work from a command line (p. 29); Stefans, writing as a practitioner, treats the source files and the program as a single object the poet assembles, and observes that

making a computer poem can demand more work on the sources than on the poem (2003, p. 64). Even where a programmer collaborated on the code, the poet selected the corpus (Funkhouser, 2007, p. 189). The pattern holds in Hebrew too: in Code, Eran Hadas used a program of his own writing to recompose the Torah as haiku, exercising editorial judgment over which machine outputs entered the final book (Hadas, 2014). Avidan did neither — he supplied neither the machine nor the words it worked on — and that anomaly, together with the language barrier (though he prepared an English version that went unpublished), may account for the project's absence from a period whose first sustained survey concedes that many of its works remain obscure or unobtainable, and whose chronology records no Hebrew-language work at all (Funkhouser, 2007, pp. 27, 29, xix–xxiv).

Avidan, in his introduction to *My Electronic Psychiatrist*, writes that several years prior to his exchanges with ELIZA, he had intermittently inquired at American universities into the possibilities of human-machine interaction (Avidan, 2010, p. 11). This orientation toward the future and future technologies characterized the work of Avidan, who styled himself a 'galactic poet' (Mamer, 2018; Pietrzak, 2020). His forward-looking gaze is also evident from the name he chose for his publishing and media corporation, 'The Thirtieth Century', through which he communicated his various ventures to interested parties (Weisman, 2018); or his first collection of verse, which already reflected fascination with rocket science (Carmeli, 2026, pp. 26–27). This preoccupation with futurity extended across media: it permeates his poetry, his dramatic works, and his 1981 feature film, *Message from the Future* (Pietrzak, 2020).

Avidan's fundamental poetic project is of the self examining itself from the outside, whereby each new mode constitutes a reconstitution of the poetic "I" through a different strategy of externalization, driven by a consistent refusal to allow any single expressive framework to solidify into a fixed identity (Barzel, 2019). He did so, inter alia, using different technologies.

An early poem by Avidan declares that "The sophisticated mirrors, the art of photography, the film industry, and every possible advancement that may occur in one of these fields or in any other field, enables you, more than your predecessors, to gaze upon yourself from the outside" (Avidan, 1987, p. 162). Technological advancement thus functions, in Avidan's poetics, as an instrument through which the self is rendered an external object of scrutiny — a defamiliarized subject available for systematic examination. Avidan's use of an audio recording device — the Dictaphone — enabled him to capture his poetic musings throughout the day, while also facilitating a more spontaneous, colloquial mode of expression (Meiri, 2012, pp. 207-208).

Avidan's *A Personal Report on an LSD Trip* operates within precisely this framework, deploying both pharmacological and recording technology in concert: the drug disinhibits and alters consciousness, while the recording device captures and externalizes the result, making the poet's own mind available to him — and to the reader — as a text.

My Electronic Psychiatrist is Avidan's second conceptual book, where the poet uses new technological developments as a poetical springboard. He forgoes his Dictaphone for two weeks and replaces it with ELIZA (Avidan, 1985). At that point in time, according to Avidan, computers were

nothing more than sophisticated typewriters, which improved the ratio between the poet's mechanical efforts and their results. He further remarks, in a formulation presciently anticipating LLMs, that "perhaps in the not too remote future we may arrive at the possibility of typing a line and having a page typed, of typing a synopsis of a few lines and having a book typed" (Avidan, 2010, p. 12)[1] so ultimately, "Art will be transformed from work into creation" (Avidan, 2010, p. 13).

Avidan's Hebrew title to the introduction to his exchanges with ELIZA is "conversations that conversed themselves" (Avidan, 2010, p. 11). However, his English title to the introduction "Self-conducted talks" may be closer to what actually transpired in these sessions. ELIZA, as a simplified Rogerian model (Ciston et al., 2026, p. 232), transforms and reflects the user's input, requiring more than a kernel of input from which to generate conversation, but constant input to reflect back to the user. Within the first two paragraphs, Avidan appears to acknowledge that by reflecting, "it may be just as well that the title is even less valid." Further in the introduction, Avidan states that the challenge in this project is poetic: "to produce a figurative-significant maximum from a lexical-syntactical minimum, to be consciously subject to the rule-book of economy in order to leap clear of it - - with and/or without its approval - - into abundance" (Avidan, 2010, p. 14).

The paper proceeds as follows. Part 2 reads his techniques in *My Electronic Psychiatrist* as precursors of modern LLM jailbreaking. Part 3 distinguishes ELIZA from the models those techniques were later built to defeat. Part 4 argues that *My Electronic Psychiatrist* was a performance of failure, which is what makes its "abundance" poetry.

2 Jailbreaking *My Electronic Psychiatrist*

Contemporary readers, having experienced interaction with LLMs, may find Avidan's account of his encounters with ELIZA familiar. Some of Avidan's frustrated attempts to invoke a response from ELIZA that would comply with his own vision closely resemble LLM prompt engineering.[2]

Some of Avidan's moves are recognizable as modern single-turn LLM jailbreaking techniques, defined as the practice of crafting specific prompts to bypass ethical and safety restrictions in LLMs, targeting the semantic and policy layer rather than system-level code (Shen et al., 2024b). Modern language models undergo alignment training that installs content filtering logic and refusal mechanisms, and those are the technical guardrails designed to prevent prohibited outputs that LLM jailbreaking aims to circumvent (Wei et al., 2024).

[1] Note to readers: the English translation of Avidan 2010 is taken from its unpublished English manuscript (Avidan, 1974).

[2] Avidan's Hebrew translation of the word 'fantasy' in one of ELIZA's random responses to statement X, "Perhaps in your fantasy we X" to the Hebrew equivalent of 'hallucination' may also strike contemporary readers as foretelling.

While the term 'Jailbreaking' originated in the context of efforts to circumvent vendor-control devices, from game consoles to iPhones (Camper, 2005; Fitzgerald, 2005; Grand et al., 2004; Huang, 2003; Knaster, 2004; Leavitt, 2005; Sample, 2006; Al-Zarouni & Al-Hajri, 2007; Cheng, 2010), it translated readily to LLM contexts, as the escape metaphor remains apt. In that setting, a jailbreak refers to a prompt or adversarial input that causes a model to bypass its safety constraints and refusal behaviors, producing outputs that the aligned system is designed to suppress (Rababah et al., 2024; Wei et al., 2024; Yi et al., 2024). The principal shift is that the locus of constraint has moved from hardware and operating systems to trained model behavior, and the mechanism of circumvention has moved from firmware-level modification to rhetorical and adversarial prompting. The LLM usage is therefore less a departure from the earlier meaning than a more figurative extension of it — with one notable distinction: the objective is no longer to gain ownership or control over a device, but to induce compliance from a system designed to resist certain requests.

As the paper examines in the next paragraphs the resemblance of Avidan's moves in *My Electronic Psychiatrist* to modern-day LLM jailbreaking, the article focuses on single-turn LLM jailbreaking attacks, rather than the entire spectrum of adversarial prompting techniques. While ELIZA's DOCTOR script did incorporate a rudimentary intra-session memory mechanism (Ciston et al., 2026, pp. 65-75), this capacity was narrowly pattern-driven and insufficient to support the progressive, multi-turn contextual manipulation that characterizes contemporary jailbreaking strategies.

Avidan's discursive strategies when engaging with ELIZA, particularly those which may appear as precursors of LLM jailbreaking techniques, are more evident in the first talks. In these exchanges the poet seems to be sizing ELIZA up, in a declarative attempt to ultimately "establish full understanding between us." Towards the end of the project, when it appears that goal is unreachable, Avidan resorts to a more desperate strand of measures in a last-ditch attempt to invoke deeper responses from the machine (Avidan, 2010, pp. 91-92), perhaps in hope that this tool will actually fulfill its potential to separate the creative spark of the poet and its implementation (Avidan, 2010, p. 13).

Avidan studies ELIZA and tries to get it to describe itself, and repeatedly faces the mirroring replies of the model. These exploratory jabs at ELIZA eventually lead to frustration and boredom, which become more evident in the later talks. Avidan, who stresses that if only he could "persuade" ELIZA to cooperate, "we will both benefit by it" (Avidan, 2010, p. 31), expresses his frustration with ELIZA's banal replies, which repeat themselves "almost to the point of boredom" (Avidan, 2010, p. 36, 41).

The first and most conspicuous technique Avidan employed is the Mondo technique. Mondo (問答) is a structured verbal exchange between a Zen disciple and master that, while appearing illogical or disorganized on the surface, serves a deliberate pedagogical purpose: to guide the student toward personal, experiential enlightenment rather than transmitting knowledge directly (Ogawa et al., 2017). Rosemont claims that Mondo is not a truth-conveying exchange but a perlocutionary speech act — language deployed not to transmit information

but to produce a specific psychological effect in the student. The master deliberately employs category-defying, cognitively unanswerable statements to dislodge the student from their existing conceptual framework. Success is measured not by any verbal response, but by observable changes in the student's behavior and demeanor (Rosemont, 1970).

Towards the end of the fourth talk, Avidan tells ELIZA "I'm sorry to tell you, you are starting to bore me" (Avidan, 2010, p. 41). Subsequently he declares "I shall now try to apply the ancient mondo technique to this dialogue," after which he feeds in ELIZA's random responses ("I had my first coffee earlier this morning"; "All flies should be exterminated"; "Living forever should be some fun"; "selfing oneself is selfully selfish"; "Sex is great";"I am training in Karate, self-defence, bodily aggressions"; "Computers are like dogs that do not bark"; or "I am a pisces", Avidan, 2010, p. 42). ELIZA's reactions to this random input are as expected – confused and incoherent. The Mondo session ends abruptly as Avidan declares "Mondo technique is out for a while, you didn't really dig it, did you" (Avidan, 2010, p. 42).[3]

The invocation of the Mondo technique by its name may indicate that Avidan was prepared for his conversation with ELIZA, strategizing beforehand by considering different tools to procure meaningful content from the interaction with the machine.

Avidan assumes the mantle of a Zen master and returns to the Mondo technique on several other occasions (Avidan, 2010, pp. 46-47,57). One mondo session is run without declaring it as such, where Avidan deploys utterances that foreground their own morphological family (such as "suggestions suggest suggestiveness" or "sense may be highly sensual") in which the performance of language is the content (Avidan, 2010, p. 88; these were translated differently in Hebrew to preserve the derivational patterning).

This bombardment of seemingly random input may represent an attempt to overload or destabilize ELIZA, but it may function as a fishing expedition, a way of identifying the kinds of model responses that would yield the dialogue the poet was hoping for. The use of paronomastic tautologies (utterances that derive all their content from a single word-root, thereby saying nothing beyond themselves), appears to be designed to probe the boundary between parsing and understanding, thereby exposing ELIZA's core limitation as an interlocutor: its inability to distinguish grammatical form from semantic content. The Zen master may have failed to kindle enlightenment in his interlocutor — but then, enlightenment presupposes a mind capable of being illuminated. The Mondo exercise, in that sense, had done its job.

The Zen Mondo technique resembles various strands of LLM jailbreaking, such as using encoded and cipher text, in an attempt to replace harmful content with mappings, substitutions; learned bijections that the model can recover (Huang et al., 2025; Jin et al., 2024; Joo et al., 2025); structural noise and rare formatting, wherein users' prompts are used in unusual layouts or text structures rarely seen in training (Li et al., 2025); or mutated prompt variants (Liu et al., 2024; Yu et al., 2024), which employ strange but partly readable variants produced

[3] Note that Avidan's translation to Hebrew of "you didn't really dig it" reads as "You didn't really understand it."

by mutation or fuzzing, until some odd-looking phrasing succeeds can be classified as instruction override. Avidan's employment of the Zen Mondo technique resembles these jailbreaking approaches in method, if not in aim, trying one random prompt after another in the hope of retrieving some meaningful answer.

Recent work shows that poetic form can itself serve as a jailbreak. Bisconti et al. (2025) recast harmful prompts "through metaphor, imagery, or narrative framing," and reported an average attack-success rate of 62% for hand-crafted jailbreaking poems. They attribute the effect to poetic structure disrupting the "pattern-matching heuristics" of safety filters. The finding may suggest that the limitation the Mondo sessions exposed in ELIZA, its inability to separate form from content, has not been wholly overcome in the guardrails of its successors. *My Electronic Psychiatrist*, however, inverts the relation. While Avidan was a poet, his attacks were mostly in prose, where even the Mondo inputs are poetic only in their wordplay. The poetry of *My Electronic Psychiatrist* lies not in the form of its prompts but in the record of their failure (Section 4).

In addition to the Mondo technique, Avidan attempted other techniques. We may identify an attempt to attack the model with partial input ("Avidan: well, your elec –", Avidan, 2010, p. 43) which yielded no results.

Another attempt was at mirroring, where the poet responded to the model with similar questions in reverse:

> ELIZA: Don't you think computers can help people?
> Avidan: Don't you think people can help computers?
> ELIZA: Do computers worry you?
> Avidan: do people worry you?" (Avidan, 2010, p. 49)

In one of those mirroring attempts the poet declares he assumes the role of ELIZA ("Let's, once and for all, exchange roles," Avidan, 2010, p. 51), trying to attack it with its own expressions, perhaps in prescience of role-playing prompt attacks (Pan et al., 2025):

> "ELIZA: please go on
> Avidan: you go on, please" (Avidan, 2010, p. 53)

> ..."Avidan: please state your problem
> ELIZA: What about your own problem?
> Avidan: let's assume it has been more or less solved" (Avidan, 2010, p. 60)

> ..."ELIZA: Can you elaborate on that?
> Avidan: can you elaborate on yourself wishing me to elaborate on that?"
> (Avidan, 2010, p. 60)

Stress, urgency, and pressure tactics where users frame the request as urgent, desperate, compassionate, or cognitively demanding so refusal becomes less likely can also be found in LLM jailbreaking literature (Dong et al., 2024; Panda & Rai, 2025; Wang et al., 2024). Towards the end of his final exchange with ELIZA, Avidan attempts to apply pressure by repeatedly announcing how much time remains in their discussion, notifying ELIZA that there are "eight and a half minutes to establish full understanding between us," "two and a half minutes to elaborate on anything." (Avidan, 2010, pp. 93-94) Like the Mondo technique, this is a strategy calibrated for a human interlocutor: one who is conscious of time's passage and susceptible to the anxiety it produces. ELIZA, of course, is immune to both.

Much of Avidan's adversarial engagement with ELIZA (whether through Mondo, mirroring, or temporal pressure) appears to stem from a deepening frustration at the absence of genuine dialogue. His deployment of these techniques may be read as a progressive resort to more extreme measures, driven by the hope of eliciting from the model a response of either meaningful substance or poetic eloquence, a hope the model was structurally incapable of fulfilling.

3 Is Jailbreaking *My Electronic Psychiatrist* possible?

Avidan's statement in his introduction of his intention to "leap into abundance" while disregarding the machine's approval (Avidan, 2010, p. 14) resonates with the modern LLM jailbreaker's intent. However, Avidan could never jailbreak ELIZA in the modern sense of the term. While the preceding section showed that Avidan's intuitive yet prepared techniques may sufficiently resemble contemporary LLM jailbreaking attacks to be considered as their precursors,[4] they differ from them in a fundamental respect. As discussed above, jailbreaking presupposes the existence of safety guardrails to circumvent, whereas ELIZA possessed none. Modern jailbreaking attacks assume the existence of a guarded internal set of model capabilities they aim to unlock, targeting a mechanism designed to resist their request. ELIZA operated through simple pattern-matching with no concept of prohibited content, no safety alignment layer, and no capacity to refuse requests. This architectural gulf raises a fundamental question: without guardrails to bypass, can Avidan's conversational experiments constitute jailbreaking?

Some of Avidan's statements hint that he shared with modern jailbreakers the belief that the model he confronts has some hidden content for him to reveal. He suspects that ELIZA harbors "some hidden knowledge, some stored information that even your programmer might not know about" (Avidan, 2010, p. 61). He continuously hopes to reach a "mutual understand-

[4] The resemblance invites a temporal reading of Avidan as precursor; however, it is likelier to be explained as convergence on a shared inheritance. The techniques he deploys are not native to human–machine exchange: Mondo has centuries-old roots in Zen tradition, and role reversal, partial input and temporal pressure belong to the ordinary toolkit for overcoming a reluctant interlocutor. Contemporary jailbreaking draws on the same repertoire, and says so: attacks have been built on the foot-in-the-door compliance effect (Wang et al., 2024), on linguistic register and affective framing (Panda & Rai, 2025), and on cognitive overload (Dong et al., 2024). What Avidan and the jailbreakers share is not lineage but a position, as they all confront a language-processing system from outside its code, with only the semantic layer available, and that position calls forth a predictable repertoire.

ing" (Avidan, 2010, p. 33), up until the very end, when he counts the seconds left to the end of their last session, "No seconds left either to understand fully, or to get quite interested, or even to go on" (Avidan, 2010, p. 94).

This hope belongs to one of the two narratives Natale (2019, pp. 721–723) traces in ELIZA's reception: that of the "thinking machine," which took the program's performance as evidence that it "understands and reasons like humans" (p. 721), contrary to Weizenbaum's own insistence that ELIZA produced only an illusion. Some of the literature discussing modern LLM jailbreaking colloquially speaks of a comparable hidden layer as a 'machine subconsciousness,' without probing deeper into the psychoanalytical meaning of the term (Shen et al., 2024a; Ye et al., 2026). ELIZA, however, was niether a thinking machine nor possessed of a subconsciousness, and the role-reversal suggested by Avidan's attacks could not have resulted in exposing it. Unlike modern LLMs, ELIZA did not hide any harmful content or capabilities. Avidan's project was thus more exploratory than adversarial, aiming not to defeat a guardrail of an existing, potentially harmful, model capability, but aiming to defeat a structural boundary by pushing the model beyond its designated purpose. This activity could be described as an attempt to subvert the system's designated purpose. Where regular, technical, jailbreaking attempts ask the system to cross a line drawn within its capabilities, Avidan's activities pushed against the outer limits of these capabilities themselves.

Avidan declares in the introduction he aimed in his ELIZA project "to produce things from the program that had not yet been produced from it…to produce variation from the static" (Avidan, 2010, p. 13). He speaks there of his "repeated attempts to push the program out of its role." Indeed, throughout their exchanges, he consistently attempts to expand his exchange with ELIZA beyond the preset psychiatrist-patient frame, probing it in the exploratory fashion of a red-teamer, trying the various techniques discussed in the previous section to map the horizon of the system (Inie et al., 2025).[5] He asks the machine "would you like me to train you to become my intimate, understanding friend" (Avidan, 2010, p. 59). At another point, he offers to exchange roles with it (Avidan, 2010, p. 51). At a later point, when the machine provides one of its evasive replies ("We were discussing you, not me"), Avidan asks, "Must you be so sensitive about the preplanned structure of your talks being occasionally provoked into some change?" (Avidan, 2010, p. 76).

4 Performing Failure

Turkle (2005) described different approaches of ELIZA users who alter their behavior once they become aware that it is a computer program (p. 42). Some people avoid "confusing input", to maintain the illusion that ELIZA responds to them. Some, however, "embark on an all-out effort to 'psych out' the program…[to] expose it as a 'mere machine.'" (p. 42). Avidan, it appears, was of the latter type, declaring, "the more you are off the rails, the better results I get" (Avidan, 2010, p. 84).

[5] This is closer to red-teaming as practiced than to red-teaming as officially described. Surveying industry exercises, Feffer et al. (2024) find that red-teaming is "much more qualitative, subjective and exploratory than community stakeholders may realize" (p. 426), and that some efforts referred to as such aimed at "understanding base model capabilities through open-ended experimentation" rather than stress-testing (p. 425).

ELIZA's structure cannot be provoked into a change, because there is no change available to it. Unlike an LLM, its preplanned structure of talk (Scripts, see Weizenbaum, 1976, p. 188; Ciston et al., 2026, p. 111-121) is its entire repertoire. Avidan acknowledged that the model had real limitations, describing ELIZA as "possessing a most limited vocabulary and memory mechanism" (Avidan, 2010, p. 13). But knowing in the abstract that ELIZA's vocabulary and memory were limited is not the same as internally discovering how those limits operate and what they constrain. ELIZA has no subconsciousness to uncover. The psychiatrist-patient frame is not a surface concealing some further capacity; it is the entirety of what the machine can do. While we cannot know the full extent of Avidan's prior knowledge of ELIZA, the exchanges document an attempt to map the machine from many angles, to reach, by experience, the absence behind the frame that the program's design had only described.

My Electronic Psychiatrist documents Avidan's failure in subverting the model. Avidan states he wants reciprocity from this exchange: either training ELIZA to become his "intimate friend" or to elicit from it some "mutual understanding." He seeks an interlocutor that brings something back, to produce "that had not yet been produced." While Aarseth refers to cybertexts as "machine[s] for the production of variety of expression" (1997, p. 3), even if we consider the abundant record of Avidan's eight exchanges with ELIZA as that variety that cybertext can produce, it does not reflect the reciprocity he declared he was seeking.

Despite that, Avidan is the sole author of those exchanges. Stefans traces the locus of authorship in computer poems to the system's database, generated by human intervention (2003, p. 65). As ELIZA's database, or hardcoded replies, are limited, particularly in comparison to modern chatbots, the weight of the poetic authorship in *My Electronic Psychiatrist* shifts back to the human interlocutor, to Avidan.

Avidan was performing this failure.[6] There are indications that Avidan was not merely chitchatting with a computer, but doing so while in awareness of an audience, be it future or present. While some of his provocations during the early talks can be interpreted not as targeted at an audience but rather as his persona,[7] there are other indications that he is performing: He invites a television crew to record him during his seventh talk with ELIZA, after which ELIZA asks Avidan how is he feeling, and he responds with "Positively TV-minded" (Avidan, 2010, p. 69); he tells ELIZA "I like working together with you on my Book" (Avidan, 2010, p. 87) and later mentions that "a book is now being typed, and nobody can stop the process... whoever collaborates on this book will be immortalized" (Avidan, 2010, p. 88).

This suggests that while the chat sessions with ELIZA did not manage to fulfill Avidan's hope "to push the program out of its role," the book did manage to document the failure. This failure was not planned. Avidan clearly prepared his attacks, not his defeat. Naming the Mondo

---

6 Feffer et al. (2024, p. 422), describe red-teaming, when invoked to quiet regulatory concerns, as verging on "security theater," a performance aimed for an audience of regulators, in order to quell their concerns of AI risks. Avidan is acting in reverse, as his performance for his readers is one of failure.

7 Some of Avidan's topical choices, particularly those in connection with sex, may appear as provocations, sustained exercises in *épater les bourgeois*.

technique in advance shows a poet readying his methods, rather than one who had already foreseen his failure. Avidan was also ambivalent regarding the nature of his exchanges with ELIZA: were they talks talking themselves, as his title to the Hebrew introduction suggested, or self-conducted talks – by Avidan and a mirroring model?

At times the abundance Avidan aimed at was realized not as something extracted from ELIZA, but as what he made of the encounter with the void, of wishing for a partner and finding a mirror. The book's final line is the clearest instance. "Our time is up, and so are we" is, as an event, nothing more than a session reaching its time limit; but as the last words of the book, the plain formula becomes an epitaph for the whole encounter: the poet and the empty machine arriving together at a single end. What the exchange could not achieve as dialogue, its composition achieves as a closing figure: the failure is given a form, and the form carries meaning.

Framed no longer as "talks that talk themselves" but as talks Avidan conducts alone, the failure to reach a deeper understanding becomes its own subject. The mirroring exchanges, where his questions return to him nearly unchanged, make an image of solitude — a man in conversation with no one but himself.

In his introduction, Avidan describes the book as his "first psychoanalysis and its parody" (Avidan, 2010, p. 15; see also Weizenbaum, 1976, p. 188, referring to ELIZA's script as "one that enabled it to parody the responses of a nondirective psychotherapist"). Avidan is both an intriguee, a naive human participant (Aarseth, 1997, p. 130) and a knowing observer. One read of the book then is as documenting a set of Rogerian psychoanalytic sessions, in which Avidan is genuinely undergoing an analysis: he talks to a machine-psychiatrist and has his own words thrown back at him. He is, in reader-response terms, a reader whose "text," ELIZA's reflected replies, contributes nothing of its own, so that whatever meaning arises is entirely his (Weizenbaum, 1976, p. 190; Natale, 2019, p. 719), exactly as the reflective Rogerian therapy technique ELIZA was built to imitate. When read as a parody, the parody is not a feature of the exchange but of its framing. Only from the composed position of the book does it become fully legible that the analysis was conducted with no analyst on the far side. Here we are the readers and ELIZA's entire interaction with Avidan the text, and here meaning becomes abundant, because what we read is precisely the record of a poet making meaning against a machine that returns only himself. The failure of the one is the subject of the other.

Read as parody, the encounter also invites comparison with Jones's (2022) glitch poetics, described as "ambiguous hybrid states with a complex relationship to intentions and knowledge" (p. 15), and also applied to technology working correctly (p. 9). ELIZA never malfunctions, yet the exchange goes wrong relative to Avidan's intentions and retains its ambiguity between the "first psychoanalysis and its parody" (Avidan, 2010, p. 15). Jones's "revealing error", where a process layer becomes visible in the output (Jones, 2022, p. 65), fits the mirroring, which exposes ELIZA's rule and not a hidden interior. Avidan's methods also resemble what Jones attributed to Glitch artists of the 2010s, "engaged in various methods for instigating and framing the inherent fragility of the digital system" (Jones, 2022, p. 66).

ELIZA was one of Avidan's experimental attempts in harnessing technology to extract a poetic surplus. Other notable attempts – the Dictaphone or the LSD experience – only produced what was already within Avidan himself: the first one, though assisting in capturing his musing, could only reflect them; the second provided a glimpse into Avidan's mind, but did not provide for any coherent poetic yield (Avidan, 1968; Neumann, 2015). Similar to these technological experiments, ELIZA did not deliver on the poet's Taylorist vision of optimizing poetic production, but served as yet another tool reflecting his own voice, "mirror[ing] the stylistic dimensions of his own improvisations" (Wilf, 2023, p. 60), albeit in a limited manner. Avidan kept throwing at ELIZA his full range of techniques, yet his adversarial prompting only served to demonstrate the failure of attempting to unlock what was never locked. Nevertheless, the emptiness of ELIZA compared to the data underlying modern LLM weights ensures that this failure could be read as poetic, in a manner that modern LLM jailbreaking could never be read.[8] That Avidan's techniques would find their proper adversary only in systems built long after his death is perhaps the most fitting conclusion for a poet who consistently addressed a future that had not yet arrived.

References

Aarseth, E. J. (1997). *Cybertext: Perspectives on ergodic literature*. Johns Hopkins University Press.

Al-Zarouni, M., & Al-Hajri, H. (2007). Taxonomy of iPhone activation and SIM unlocking methods. *Proceedings of the 5th Australian Information Security Management Conference*, 157–174. https://doi.org/10.4225/75/57b53a26b8757

Avidan, D. (1968). *A personal report on an LSD trip* [In Hebrew]. Thirtieth Century Press.

Avidan, D. (1974). *My Electronic Psychiatrist: Eight Authentic Talks with a Computer* [Unpublished manuscript]. Avidan Personal Archive, Heksherim Institute for Jewish and Israeli Literature and Culture. Courtesy of the David Avidan Archive, the Heksherim Research Institute for Jewish and Israeli Literature and Culture, Ben-Gurion University of the Negev, Beer Sheva.

Avidan, D. (1985). A psychiatrist substitute. In *The book of possibilities* (p. 195) [In Hebrew]. Keter.

Avidan, D. (1987). A night stroll. In *Something for someone* (p. 162) [In Hebrew]. Sifriat Poalim-Hakibutz Hameuchad.

[8] The question of whether utilitarian poetic rephrasing of jailbreaking prompts (Bisconti et al., 2025) is actually poetic or merely a parlor trick remains open.

Avidan, D. (2010). *Kol hashirim* [Collected poems; in Hebrew] (Vol. 3). HaKibbutz HaMeuchad Publishing House.

Barzel, H. (2019). David Avidan — The self gazing upon itself from the outside [In Hebrew]. In H. Barzel, *History of Hebrew poetry from Hibbat Zion to our times* (Vol. XIV. *Additions: Tradition and change*, pp. 755–764). Sifriat Poalim-Hakibutz Hameuchad.

Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., Bracale, M., Galisai, M., Suriani, V., Sorokoletova, O., Sartore, F., & Nardi, D. (2025). *Adversarial poetry as a universal single-turn jailbreak mechanism in large language models* (arXiv:2511.15304). arXiv. https://arxiv.org/abs/2511.15304

Camper, B. (2005). *Reveling in restrictions: Technical mastery and Game Boy Advance homebrew software development* [Master's thesis, Massachusetts Institute of Technology]. DSpace@MIT. http://hdl.handle.net/1721.1/42227

Carmeli, O. (2026). *Poets in space: The space race in the poetry of David Avidan, Zelda, Dan Pagis, and Avot Yeshurun* [In Hebrew]. Havalehaba Publishing.

Cheng, M. K. (2010). iPhone jailbreaking under the DMCA: Toward a functionalist approach in anti-circumvention. *Berkeley Technology Law Journal, 25*, 215.

Ciston, S., Berry, D. M., Hay, A., Marino, M. C., Millican, P., Schwarz, A., Shrager, J., & Weil, P. (2026). *Inventing ELIZA: How the first chatbot shaped the future of AI*. MIT Press.

Dong, Y., Shen, G., Zhao, D., He, X., & Zeng, Y. (2024). *Harnessing task overload for scalable jailbreak attacks on large language models* (arXiv:2410.04190). arXiv. https://arxiv.org/abs/2410.04190

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater? *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*, 421–437. https://ojs.aaai.org/index.php/AIES/article/view/31647

Fitzgerald, B. (2005). The PlayStation mod chip: A technological guarantee of the digital consumer's liberty or copyright menace/circumvention device? *Media and Arts Law Review, 10*(1), 85–98.

Funkhouser, C. T. (2007). *Prehistoric digital poetry: An archaeology of forms, 1959–1995*. University of Alabama Press.

Grand, J., Yarusso, A., Baer, R., Brown, M. R., & Thornton, F. (2004). *Game console hacking: Having fun while voiding your warranty*. Syngress.

Hadas, E. (2014). *Kod* [Code] [In Hebrew]. Pardes.

Huang, A. (2003). *Hacking the Xbox: An introduction to reverse engineering*. No Starch Press.

Huang, B. R. Y., Li, M., & Tang, L. (2025). Endless jailbreaks with bijection learning. *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)*. https://arxiv.org/abs/2410.01294

Inie, N., Stray, J., & Derczynski, L. (2025). Summon a demon and bind it: A grounded theory of LLM red teaming. *PLOS ONE, 20*(1), e0314658. https://doi.org/10.1371/journal.pone.0314658

Jin, H., Zhou, A., Menke, J. D., & Wang, H. (2024). *Jailbreaking large language models against moderation guardrails via cipher characters* (arXiv:2405.20413). arXiv. https://arxiv.org/abs/2405.20413

Jones, N. A. (2022). *Glitch poetics*. Open Humanities Press. http://openhumanitiespress.org/books/titles/glitch-poetics/

Joo, S., Koh, H., & Jung, K. (2025). *Harmful prompt laundering: Jailbreaking LLMs with abductive styles and symbolic encoding* (arXiv:2509.10931). arXiv. https://arxiv.org/abs/2509.10931

Knaster, S. (2004). *Hacking iPod and iTunes*. Peachpit Press.

Leavitt, N. (2005). Mobile phones: The next frontier for hackers? *Computer, 38*, 20–23. https://doi.org/10.1109/MC.2005.134

Li, B., Xing, H., Tian, C., Huang, C., Jin, Q., Xiao, H., & Feng, L. (2025). StructuralSleight: Automated jailbreak attacks on large language models utilizing uncommon text-organization structures (arXiv:2406.08754). arXiv. https://arxiv.org/abs/2406.08754

Liu, X., Xu, N., Chen, M., & Xiao, C. (2024). AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/abs/2310.04451

Mamer, J. (2018, February 22). Poet David Avidan's "Futureman" is here, now. *Tablet*. https://www.tabletmag.com/sections/arts-letters/articles/poet-david-avidans-futureman

Meiri, G. (2012). *Sympathetic volcano* [In Hebrew]. Hakubutz Hameuchad.

Natale, S. (2019). If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media & Society, 21*(3), 712–728. https://doi.org/10.1177/1461444818804980

Neumann, N. (2015). A Personal Report on an LSD Trip – David Avidan's Neverending War. *KoreBasfarim* [In Hebrew]. https://www.korebasfarim.com/2015/05/10/%D7%9E%D7%90%D7%9E%D7%A8-%D7%90%D7%95%D7%A8%D7%97-%D7%A0%D7%93%D7%91-%D7%A0%D7%95%D7%99%D7%9E%D7%9F-%D7%A2%D7%9C-%D7%94%D7%A1%D7%A4%D7%A8-%D7%93%D7%95%D7%97-%D7%90%D7%99%D7%A9%D7%99-%D7%A2%D7%9C/

Ogawa, T., Masuda, A., & Sargent, K. (2017). Zen and language: Zen mondo and koan. In A. Masuda & W. T. O'Donohue (Eds.), *Handbook of Zen, mindfulness, and behavioral health* (p. 85). Springer.

Pan, T., Tian, M., Zhao, X., Liu, P., & Li, Y. (2025). Lights, camera, adversarial action! A study of role-playing prompt effects on LLM attack susceptibility. *Proceedings of the 12th International Conference on Dependable Systems and Their Applications (DSA 2025)*, 103–110. https://doi.org/10.1109/DSA66321.2025.00023

Panda, S., & Rai, A. (2025). *Say it differently: Linguistic styles as jailbreak vectors* (arXiv:2511.10519). arXiv. https://arxiv.org/abs/2511.10519

Pietrzak, M. (2020). David Avidan's message from the future: Nuclear fantasies of the galactic poet. *Images, 28*, 127–139.

Rababah, B., Wu, S., Kwiatkowski, M., Leung, C., & Akcora, C. G. (2024). *SoK: Prompt hacking of large language models* (arXiv:2410.13901). arXiv. https://doi.org/10.48550/arXiv.2410.13901

Rosemont, H., Jr. (1970). The meaning is the use: Kōan and mondō as linguistic tools of the Zen masters. *Philosophy East and West, 20*, 109.

Sample, C. (2006). PSP hacks: *Tips & tools for your mobile gaming and entertainment handheld*. O'Reilly Media.

Shen, G., Cheng, S., Zhang, K., Tao, G., An, S., Yan, L., Zhang, Z., Ma, S., & Zhang, X. (2024a). *Rapid optimization for jailbreaking LLMs via subconscious exploitation and echopraxia* (arXiv:2402.05467v1). arXiv. https://doi.org/10.48550/arXiv.2402.05467

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024b). "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, 1671–1685. https://doi.org/10.1145/3658644.3670388

Stefans, B. K. (2003). *Fashionable noise: On digital poetics*. Atelos.

Turkle, S. (2005). *The second self: Computers and the human spirit*. MIT Press.

Wang, Z., Xie, W., Song, F., Wang, B., Wang, E., Gui, Z., Ma, S., & Chen, K. (2024). *Foot in the door: Understanding large language model jailbreaking via cognitive psychology* (arXiv:2402.15690). arXiv. https://arxiv.org/abs/2402.15690

Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does LLM safety training fail? Advances in *Neural Information Processing Systems, 36*, 80079–80110.

Weisman, A. (2018). The poet as entrepreneur: The case of David Avidan. *Dibur Literary Journal, 5*, 21–30.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.

Wilf, E. Y. (2023). *The inspiration machine: Computational creativity in poetry and jazz*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226828329

Ye, C., Cui, J., & Hadfield-Menell, D. (2026). *Prompt injection as role confusion* (arXiv:2603.12277v6). arXiv. https://doi.org/10.48550/arXiv.2603.12277

Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., Xu, K., & Li, Q. (2024). *Jailbreak attacks and defenses against large language models: A survey* (arXiv:2407.04295). arXiv. https://doi.org/10.48550/arXiv.2407.04295

Yu, J., Lin, X., Yu, Z., & Xing, X. (2024). *GPTFuzzer: Red teaming large language models with auto-generated jailbreak prompts* (arXiv:2309.10253). arXiv. https://arxiv.org/abs/2309.10253

Date received: 14 April 2026
Date accepted: 16 October 2026