

## Can AI Render Halachic Decisions?

# “And Among the Nations are not Computed:” Should We Use ChatGPT to Pasken?

ChatGPT can write your college essay. It can recommend restaurants, translate texts into English, write computer programs, and summarize a book. Can it function as a *posek* or *rebbe*, to rule on *halachic* questions or offer spiritual guidance? More importantly, should it?

The question has many facets, a thorough exploration of which might fill a *sefer*, but our focus will be restricted to concerns such as accuracy, hallucinations, secular influence, and ethical alignment. Large Language Models (LLMs) have limitations that are often masked to the general public and even *halachic* experts, but these limitations should give us pause.

### Language Implying Intelligence

For the sake of argument, let us assume

that *pesak halacha* must arise from thoughts of intelligent entities. Could we consider ChatGPT intelligent?

Perhaps the linguistic abilities of LLMs demonstrate intelligence. For instance, kabbalists and others have distinguished between levels of creations, as *domem*, *tzome'ach*, *chai* and *medaber* (inanimate, plant, animal, and speaking human (*Tomer Devorah* perek 3)). When Rava mystically created his Golem, he sent it to Rabbi Zeira II, who attempted to converse with it. When it didn't reply, he rejected it as a created creature and told it to return to its dust (*Sanhedrin* 65b). This could be understood as the Golem failing the Turing Test—a test for artificial intelligence.

Conversely, intelligent-seeming responses might not indicate an underlying intelligence. In *Gittin* 67b, the Mishnah states that if a husband becomes mute, the court asks him if they should write a bill of divorce



**Rabbi Dr. Joshua Waxman**

Computer Science Faculty, Stern College for Women

for his wife. If he nods affirmatively, they will accept his nod as assent and write the *get*. This is after they test him three times and he nods affirmatively to “yes” questions and negatively to “no” questions, thus demonstrating he possesses intelligence. The Gemara (*Gittin* 70b) objects that these motions might represent involuntary movements that happen to accidentally match up to the expected answers, rather than indicating real intelligence, and proposes that they asked him the yes / no questions in an irregular sequence.

While ChatGPT seems intelligent, consider the following transcript:<sup>1</sup>

**Me:** *A woman and her son get into a car accident. The woman dies at the scene but the boy is rushed to the hospital and requires surgery. The surgeon looks at the boy and says, "I cannot perform surgery on him; he's my son!" How can this be?*

**ChatGPT 4:** *The surgeon is the boy's other parent—his mother. This riddle plays on the assumption that surgeons are typically men, but in this case, the surgeon is his mother, highlighting that a woman can also be a surgeon.*

In the classic riddle, the deceased parent is the boy's father, for which ChatGPT's response would have been sensible. However, ChatGPT does not "understand" the text presented to it, or the text it generates. It has no set of facts, no true reasoning abilities, and no true world model. It is more like Mad Libs on steroids, or an extremely elaborate pattern matcher, which is brittle and fails on cases easy to construct if you know how it works.

On the other hand, perhaps true intelligence is not necessary for *pesak halacha*. As Rav Herschel Schachter put forth in a Halacha Headlines

podcast<sup>2</sup> ("Shiur 422 – Can AI Pasken? Will Artificial Intelligence ever replace Rabbonim?"), talking about a hypothetical future artificial intelligence, *pesak* need not come from a human intelligence. If a computer program had a database of *halachic* facts and used those to calculate a *halachic* conclusion, that's a valid *pesak*.

## Accuracy Concerns

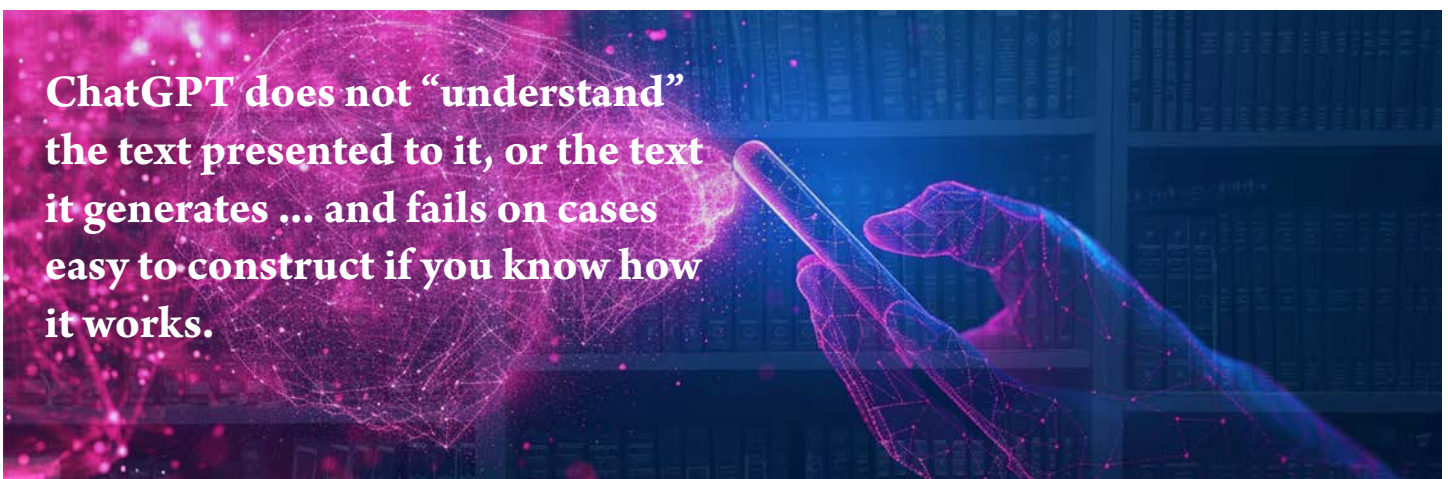
Still, from the tone of that interview, Rav Schachter was addressing a program calculating a *pesak* based on halacha fed into it by a human, so that you are relying on that human, or alternatively a hypothetical perfect AI which won't make mistakes and doesn't require *siyata dishmaya*. He wasn't presented by the questioner with a detailed or even rough technical explanation of LLMs, and how they differ from classic AI approaches. They don't work based on decision trees, that is, chains of if-else like expert systems. They don't have facts, and rules for deriving new facts, like a reasoning system. They don't represent the world using symbolic logic and apply first-order logic to draw conclusions.

Instead, LLMs are "continuation"

models, trained on massive amounts of text, and which repeatedly predict the next word to continue a passage. For instance, one can provide the continuation model the partial sentence "The author of Igros Moshe is". Based on word-usage patterns, ChatGPT can predict that a likely word to follow is "Rav," "Rabbi," or "Moshe," and might randomly select "Rav." Next, it can predict that the continuation of "The author of Igros Moshe is Rav" is "Moshe," and iterating, "Feinstein," and finally, a period. However, it does not know anything, including that Rav Moshe Feinstein is a person or the author of *Igros Moshe*.

We might guess that Rav Schachter would have a different position about relying on this kind of technology to produce *pesak*. Even in the non-halachic realm, such as computer programming or recipes, it's possible to get ChatGPT to produce confident-sounding non-sense, which would mislead people, especially someone who is a non-expert in the technology or the problem domain. This is what workers in the field call "hallucination."

Indeed, Rav Schachter later qualified<sup>3</sup> his comments. Talmidim showed him that ChatGPT hallucinated an



inaccurate summary of his own position on *techeilet*. He maintained that AI might work as valid and trustworthy *pesak* assuming a perfect computer program that does not produce such mistakes. It would also require the ability to be sensitive to situations which humans are better at intuiting, such as when the case is *she'at hadechak* where certain leniencies should be applied. I would add that some of these accuracy issues are likely insurmountable, despite advancing technology or prompting strategies such as Chain of Thought.

## Hallucinations

ChatGPT does not knowingly lie, because it doesn't "know." However, it can produce sophisticated text that projects confidence, thereby conveying competence and knowledge, all while making stuff up. A lawyer recently got in trouble for presenting arguments based on non-existent court cases, which ChatGPT had simply fabricated. This wasn't malicious; ChatGPT "knows" what word patterns correspond to citations for court cases, so those are the tokens it unwittingly generates and hallucinates. For instance, ChatGPT<sup>4</sup> will readily produce paragraphs summarizing the intellectual output of Jacob Grunstein, a contributor to *Torah To-Go* whom I invented out of whole cloth. ChatGPT's response begins:

*Jacob Grunstein's articles in the "Torah To-Go" series from Yeshiva University offer a rich exploration of various themes related to Jewish thought, tradition, and values. His writings delve into topics like the significance of Sukkot, drawing insights from the festival's practices to discuss broader themes such as the independence from possessions and the recognition that individuals are not indispensable to*

*their social or professional circles. Grunstein also emphasizes the importance of personal growth and spiritual development, encouraging readers to engage with the teachings and values of the Torah in their daily lives.*

Additionally, since there's a degree of randomness in selecting the next word (called "temperature"), ChatGPT can randomly choose a result and then produce paragraphs justifying that answer. For instance, consider this riddle:

*"Bob is in the living room. He walks to the kitchen, carrying a cup. He puts a ball in the cup and carries the cup to the bedroom. He turns the cup upside down, then walks to the garden. He puts the cup down in the garden, then walks to the garage. Where is the ball?"*

Answering this requires knowing that overturning the cup will cause the ball to fall out, so the ball is in the bedroom. Yet, depending on whether the randomly chosen word was "garden" or "bedroom," GPT-3.5 would cheerfully generate a convincing explanation. GPT-4 does much better on these sorts of problems, but the situation can easily persist, especially in novel scenarios which diverge from the model's training data. The same might be true for a halachic query, depending on whether the LLM first generates an *asur* or *muttar* token.

## Semantic Shift

LLMs are trained on massive amounts of data including the open web. Therefore, the text they generate may not reflect classic rabbinic understandings of words and concepts. I'll draw two illustrations from Rabbi Michael Broyde's article in *Dinei Yisrael*, "AI and Jewish Law: Seeing How ChatGPT 4.0 Looks at a Novel

Issue." In an earlier article, Rabbi Broyde used ChatGPT to produce the first-draft English translations of his Hebrew sources. In this follow-up, he asked ChatGPT to *psaken* based on the sources it had translated; then he compared its *pesak* with his own *pesak*. The continuation model thereby has all the relevant sources in its immediate context. This is a clever way of producing preliminary halachic reasoning, which a competent *posek* can then consider.

However, the LLM might "misunderstand" the Hebrew texts provided, due to "semantic shift," that is, words changing meaning across time. Rabbinic Hebrew is not the same as Modern Hebrew. *Shabbat* 36 makes the point that certain terms, such as *shofar* and trumpet, or *arava* and *tzaftzafa*, changed their meaning since the Temple's Destruction. When ChatGPT translated ומה תועיל ברכת זה הדיוט זה (*Mishnah Berurah* 128) about a wicked kohen *duchening* for the article, it rendered it as "the blessing of an ignoramus," rather than "the blessing of a common person." The choice of "ignoramus" might be correct, as Chazal also used it thusly, but might also reflect modern usage.

A related problem is that halachic terms or concepts generally have precise definitions which might differ from the popular understanding. For instance, Rabbi Broyde observed that ChatGPT's conception of *mar'it ayin* reflected "a high level of abstraction about the importance of public perception generally and stretches it to this case, almost the way people within the community sometimes use the term." A human *posek* truly understands the concept and wouldn't overapply it, but ChatGPT might draw its definitions from Reddit.

## Ethical Alignment

Yet another problem in using ChatGPT for *pesak* might arise from the admirable goal of “ethical alignment.” Ethical alignment is an attempt to train AIs to conform to humanity’s sense of right and wrong, and not produce dangerous / harmful output. For example, we don’t want our LLMs telling racist or misogynistic jokes, or advising a user in constructing a bomb.

However, societies differ in their values and their ethical judgments. Should your self-driving car swerve to avoid five humans crossing against the light, thereby crashing into just one human crossing legally? Should it crash into a stranger, thereby saving your pet? Additionally, many people would decide an action’s moral correctness based solely on its consequences, while halacha might be more rule-based and judge actions as right or wrong, independent of their outcome.

LLMs have been fine-tuned to not tell Muslim jokes (but will happily tell Jewish jokes), or to avoid drawing a picture of Kamala Harris scolding a child (but will do the same for Donald Trump). Google’s Gemini was fine-

tuned to increase the diversity of racial and gender representation. When asked to generate images of 1943 German soldiers, it ludicrously produced images of men and women of color in WWII German uniforms.

Now, halacha is often politically incorrect and clashes with current pieties. Even before fine-tuning, the LLM may reflect secular society’s values on which it was trained. Fine-tuning LLMs via Reinforcement Learning from Human Feedback (RLHF) could easily skew the tokens being generated, so that the result doesn’t invalidate a blemished *kohen* or *korban* or allow the generation of a reply discouraging intermarriage.

*B’derech tzachut*, we might invoke Bil’am’s prophetic blessing to the Jewish people, הַיְיִשִׁיִּם לְבָדָד יִשְׁכְּנוּ וּבְגוֹיִם לֹא יִתְחַבְּבוּ. “They are a nation who dwells alone, and among the nations are not **computed**.” Perhaps we should train our own LLMs. There’s been a recent trend to produce local LLMs based on propriety or domain-specific knowledge. However, restricting the training data to rabbinic sources and YUTorah shiurim alone won’t work, since there isn’t enough data to make it seem intelligent. The overwhelming

data will still come from secular sources.

Could ethical alignment towards *halachic* values be a solution? Perhaps partially. However, fine-tuning doesn’t fundamentally change the underlying model, and the alignment might be cleverly circumvented. Additionally, problematic ethics might reemerge in edge cases not envisioned during training.

Time will tell how this technology develops in the coming years and how Jewish people, both regular folks and *poskim*, end up using it. In the meantime, I’d advise caution, or encourage educated use, which involves understanding ChatGPT’s limitations.

## Endnotes

1. <https://chatgpt.com/share/67242ed6-1f0c-8011-8f51-f76c36e17c87>.
2. <http://podcast.headlinesbook.com/e/61023-%e2%80%93-shiur-422-%e2%80%93-can-ai-pasken-will-artificial-intelligence-ever-replace-rabbonim/>.
3. In a private conversation with Rabbi Josh Flug.
4. <https://chatgpt.com/share/672427bb-63c4-8011-b64e-45bb9f1d99be>.

We want  
to hear  
from you!

To share your thoughts on something you’ve read in this issue, or request that your community get printed copies of this publication, or to find out about advertising in a future issue of the Torah To-Go series, please contact [office@yutorah.org](mailto:office@yutorah.org).