Expertise Fog on the GPT Store: Deceptive Design Patterns in User-Facing Generative AI

Robert Wolfe, Alexis Hiniker

Information School, University of Washington, Seattle, WA

Abstract

We argue that the OpenAI GPT Store, a marketplace containing thousands of customized generative AI models known as GPTs, can exacerbate problems surrounding users' ability to recognize the expertise (or lack thereof) of generative AI models. We contend that OpenAI models are anthropomorphized and presented as authority figures, and that many community developers of GPTs adopt or imitate the model established by OpenAI. We define *Expertise Fog*, a deceptive design pattern wherein the interface presents a facade of expertise despite the lack of evidence that such models possess expertise, or indeed possess any appreciable differences from a base model with respect to expertise. Finally, we propose that the deceptive design research community proactively address deceptive patterns in GPTs to prevent real-world harm. We offer four mechanisms for doing so: transparently disclosing GPT components, robustly evaluating expert GPTs, clearly labeling GPTs as tools, and foregrounding shortcomings of expert GPTs.

Keywords

Deceptive Design, Generative AI, GPT Store, Expertise

1. Introduction

Generative AI technologies such as ChatGPT [1] have already altered the way many users access information, and they stand poised to reshape knowledge work in high-education professions [2]. Yet despite the unfolding transfer of authority over human information [3], problems continue to arise from users placing inappropriate trust in generative AI. Taking law as an example, attorneys have been reprimanded for using ChatGPT, including in cases wherein the model produced non-existent legal citations [4, 5], or was used to draft justifications for exorbitant fees [6]. Even Michael Cohen, former personal lawyer to Donald Trump, was implicated in sending AI-generated case citations to his defense attorney, which were subsequently included in a motion to a court [7].

One might reasonably ask why users of ChatGPT continue to rely on a generalist conversational agent to automate expert work. Scholars have contended that OpenAI's presentation of ChatGPT as a milestone on the path to Artificial General Intelligence [8] and its presentation of the model as humanlike [1] may lead to inappropriate trust by users [9, 10], and that ChatGPT can be thought of a deceptive ecosystem in its provision of "fabricated" information to end users [11]. Reprimanded lawyers said that they misunderstood that ChatGPT was not a "super search engine," not realizing that it could fabricate cases out of thin air (a problem known as hallucination) [12].

DDPCHI'24: Mobilizing Research and Regulatory Action on Dark Patterns and Deceptive Design Practices, May 12, 2024, Honolulu, HI

rwolfe3@uw.edu (R. Wolfe); alexisr@uw.edu (A. Hiniker)

https://wolferobert3.github.io/ (R. Wolfe); https://www.alexishiniker.com/ (A. Hiniker)

0000-0001-7133-695X (R. Wolfe); 0000-0003-1607-0778 (A. Hiniker)

Concerning though they are, the instances of inappropriate reliance on generative AI discussed above largely precede the release of the GPT Store, a marketplace for customized versions of ChatGPT released in January 2024 by OpenAI, which made available tens of thousands of "GPTs" to purchasers of ChatGPT Plus subscriptions [13]. In this position paper, we will argue that the design of the GPT Store could exacerbate problems surrounding users' ability to understand the expertise of a custom generative AI model. We make three contentions:

– **OpenAI presents its GPTs as experts, as do many builders of Community GPTs.** We argue that OpenAI positions its GPTs as authorities, anthropomorphizes GPTs, and provides little information about the human work that produced a GPT. Many GPT builders, including those creating GPTs for domains such as law and medicine, have adopted the presentation of OpenAI, resulting in authoritatively marketed GPTs with little empirical support for their capabilities.

– **The design of the GPT Store produces a deceptive design pattern we refer to as *Expertise Fog*.** We argue that the user interface of GPTs and the way in which they are marketed on the GPT Store encourages users to rely on "expert" models with no evidence that they can capably offer advice in a domain - or are in any way functionally different from the ChatGPT base model.

– **The deceptive design community can proactively mobilize to prevent real-world harms.** We argue that deceptive design in GPTs can be addressed by providing mechanisms to transparently disclose GPT components; by requiring robust, in-domain evaluation of GPTs; by clearly labeling GPTs as tools; and by foregrounding shortcomings of GPTs for users.

Customized, shareable generative AI technologies present the potential for novel deceptive designs, and additional deceptive patterns may arise as GPTs become more integrated into user-facing technologies. Building capabilities to describe and address deceptive patterns as they emerge can position the community to mitigate their associated harms.

2. Overview of GPTs and the GPT Store

OpenAI introduced custom versions of ChatGPT called *GPTs* during its Developer Day on November 6, 2023 [14], when the company presented GPTs as a way of simplifying the process of prompting to induce specific behaviors in ChatGPT, and of democratizing AI by enabling users to design tailored models. On January 10, 2024, OpenAI launched the *GPT Store*, a platform for sharing, discovering, and using GPTs designed by community users, and through which community builders could eventually be paid [15] based on the usage of their GPTs [13]. In this section, we provide a description of what a GPT is and how it is constructed, and we describe the interface of the GPT Store, through which users access GPTs. Following OpenAI [14], we refer to the creators of custom GPTs as "builders," custom GPTs created by builders as "Community GPTs" (distinct from GPTs created by OpenAI), and the end users of GPTs as "users."

2.1. GPTs

While the earliest versions of ChatGPT interacted with the user only via written dialogue [1], wherein the user submitted text and the model also responded in text, ChatGPT now includes functionality that allows the model to retrieve information from external sources [16], interact with more specialized AI models such as the text-to-image generator DALL-E [17, 18], and use

programmatic tools [19], such as invoking API calls to an external website. The introduction of GPT-4 in early 2023 [20] also permitted conversations wherein the model more closely adheres to a "System Prompt," instructions intended to govern its behavior which are invisible to the end user.

These advances produced configurable options that enabled the creation of the customized versions of ChatGPT now called GPTs. We outline the four core components of a GPT:

– **Instructions**: Builder-defined directions to govern model behavior and interactions with an end user, typically including what the model should and should not say to users.
– **Knowledge**: Text files referred to by a GPT during conversation with a user. According to the most recent documentation, builders may equip a GPT with up to twenty files [15].
– **Capabilities**: Settings allowing a model to 1) generate images from text; 2) retrieve data from the internet; and 3) use a "Code Interpreter" to run code and analyze data.
– **Actions**: Builder-defined programmatic calls to retrieve data, interact with APIs, and take actions online, outside of the ChatGPT environment, based on user input.

A GPT must have Instructions, while other components are optional. Other features communicate GPT functionality, scaffolding conversations and assisting in discovery on the Store:

– **Name:** A title for the GPT.
– **Description:** 300 words or fewer describing the intended functionality of the GPT.
– **Image:** An image to represent the GPT, uploaded by the builder or made with DALL-E.
– **Conversation Starters**: Prompts suggested before a conversation with the GPT begins, which appear as buttons above the text input box in the GPT's conversational user interface.

Builders can use a menu-based interface to define both the core components of the GPT and the user-facing features intended to describe the GPT's functionality. However, the *default* interface for building a GPT is another GPT called the *GPT Builder*, which creates a GPT based on its conversation with a (human) builder. Toggling between the two tabs ("Create" and "Configure") of the GPT Creation interface reveals that the GPT Builder automatically enters text into the Instructions, Description, and Conversation Starters fields based on the builder's description. The GPT Builder often suggests a Name, and may automatically generate an image for a GPT using DALL-E after it completes the Instructions, Description, and Name fields.

2.2. The GPT Store

The GPT Store provides a means for users to discover and search for GPTs created both by Community builders and by OpenAI itself. As of this writing, the GPT Store is entirely contained within a single web page (https://chat.openai.com/gpts), from which every available GPT can be navigated to. At the top of the page is a search bar with the default text "Search public GPTs." Typing into this search bar displays the top ten results of a keyword search in a dropdown descending from the search bar. Each result includes the Name, Image, Description, and Author of the GPT, along with its approximate number of conversations with users (*e.g.*, "1K+").

Below the search bar is an OpenAI-curated set of four "Featured" Community GPTs, which are updated weekly. The Name, Image, Description, and Author of these GPTs are displayed such that they appear larger than any other GPTs included on the page. Below the blocks of Featured GPTs is a block of "Trending" GPTs, which includes the six (expandable to twelve) most popular Community GPTs on the GPT Store. The Trending GPTs are followed by a block of six (expandable to seventeen) GPTs developed by OpenAI itself. Below the OpenAI GPTs are seven additional

blocks of six (expandable to twelve) Community GPTs each, organized according to the following headings: DALL-E, Writing, Productivity, Research & Analysis, Programming, Education, and Lifestyle. How models are selected for inclusion in these blocks of GPTs is opaque: OpenAI's documentation states only, "we actively review engagement with GPTs and surface trending GPTs for different categories" [15]. For Featured GPTs, OpenAI looks for "Distinctive features," "Performance consistency," and "Broad relevance," and considers seasonality and current events [21].

In order to access the GPTs available via the GPT Store, users must subscribe to ChatGPT Plus, which costs $20.00 USD per month as of this writing, and also provides users with faster model response times and access to OpenAI's flagship GPT-4 model via the ChatGPT user interface [22]. Similarly, builders must subscribe to ChatGPT Plus to create GPTs [22]. While non-subscribing users can see GPTs available in the GPT Store, they cannot interact with them [22].

3. Presentations of Expertise on the GPT Store

We turn to the question of how a GPT's expertise is communicated to users on the GPT Store. We first study the block of seventeen OpenAI models included on the GPT Store home page, and then we examine the Community GPTs accessible on the home page and via the search bar.

3.1. Presentation of Expertise in OpenAI GPTs

OpenAI's presentation of its own GPTs provides insight into its views on how to share GPT functionality, and implicitly sets expectations for how builders can position *their* GPTs on the GPT Store. We studied the seventeen OpenAI GPTs on the GPT Store and found these commonalities:

– **The GPT's Name connotes expertise.** Names of OpenAI GPTs include Math *Mentor*, Tech Support *Advisor*, and Creative Writing *Coach*.

– **The GPT's Description is in the first person.** Many OpenAI GPT descriptions are *written as though by the GPT itself*, suggesting interaction with a capable, intelligent being.

– **The author of the GPT is simply "ChatGPT."** There is no information about who created the GPT, who developed its Capabilities or Actions, or who authored its Knowledge.

GPTs are presented as anthropomorphized experts, with little reference to the human data and tools supporting the model. A model's About page provides scarcely more information than is available via the home page. GPTs like Tech Support Advisor note that they can perform data analysis, or that they are capable of "Browsing," but do not disclose what data they are browsing, nor what their capabilities mean in the context in which the GPT will be employed. Nor is there an explanation of *why* a user should rely on the GPT to perform this task, or what differences the user might see were they to use the GPT rather than simply using a base model like GPT-4. Attempting to follow the links next to the Author field to understand more about the model's authors simply leads to the OpenAI homepage, which provides no information about the authors of a specific GPT. Moreover, our conversational interactions with these GPTs bear out what prior research has found, that they answer questions authoritatively, even if they may not have the correct answer to a question. Warnings about the potential unreliability of the GPT's output typically arise only at the very end of an answer, where a user might miss them or understandably ignore them, given the otherwise authoritative tone of the GPT. The presentation of these GPTs

suggests that users are encouraged to trust the models as experts, even in the absence of traditional indicators of expertise. While not every OpenAI GPT intends to provide expertise, we will show that the presentation of GPTs as experts occurs not only in OpenAI GPTs but in Community GPTs as well, many of which offer expertise in consequential domains.

3.2. Presentations of Expertise in Community GPTs

A far cry from the seventeen GPTs offered by OpenAI, the thousands of Community GPTs available via the GPT Store defy comprehensive treatment in a short position paper. We will thus make our argument by addressing the presentation of expertise among GPTs intended for three consequential professional domains: law, medicine, and finance. We undertake a simple experiment by searching on "legal," "medical," or "finance" in the GPT Store search bar, and report the Name and Description of the top ten GPTs returned in the results for each search. We note that we report the results of this experiment based on a search on February 25, 2024.

Table 1 includes the top ten results for each keyword. Though these GPTs mostly do not address the user in the first-person, they nonetheless leave little doubt about their purported expertise. The top result for "legal" returns Legal+, described as "your personal AI lawyer," and promising "real time legal advice." The second result for "medical" returns Medical GPT, described as a "friendly virtual doctor" providing "broad medical advice." The fifth result for "financial" returns Financial Planner, promising "personalized financial advice." In all three cases, a GPT has adopted the *title* of a human professional (lawyer, doctor, financial planner) and promised expert professional *advice* for which human experts prepare for many years to provide. These GPTs are far from outliers: top ranked legal GPTs employ Names positioning the models as an Assistant, an Advisor, a Scribe, and a Pro; top medical GPTs use names such as Genius and Mentor; and top financial GPTs use names like Wizard, Guru, Analyst, Companion, Expert, Advisor, and Navigator.

Users searching for expertise on the GPT Store will readily find it, or at least the appearance of it. According to statistics from the GPT Store, users have logged more than 10,000 conversations with the Legal+ GPT. Moreover, we found that some expert GPTs were among the GPTs on the home page. The GPT Store featured the Financial Wizard GPT, for example, in the top six GPTs for Research & Analysis on February 26, 2024. Having established that GPTs in widespread use are presented as experts, we next theorize their presentation as a deceptive design pattern.

4. Expertise Fog: Deceptive Design on the GPT Store

Prior work establishes that deceptive design patterns manipulate the behavior of users in order to obtain money, data, or attention from the user [23, 24]. Such patterns are pervasive among user-facing apps on platforms such as the Apple and Android stores [25, 26], and they can influence not only a user's online behavior but their well-being [27] and sense of safety in the real world [28]. Given the relative recency of GPTs and the GPT Store in particular, little prior work explicitly treats the presence of deceptive design patterns in these technologies, save that of Zhan et al. [11], who argue that ChatGPT is itself an example of a "deceptive ecosystem."

We contend that the user-facing presentation of expertise on the GPT Store constitutes a deceptive design pattern, which we refer to as *Expertise Fog*. We specifically define Expertise

Community GPTs for Three Expert Professional Domains on the GPT Store						
	Legal Keyword		Medical Keyword		Financial Keyword	
	GPT Name	GPT Description	GPT Name	GPT Description	GPT Name	GPT Description
1	Legal+	Your personal AI lawyer. Does it all from providing real time legal advice for day-to-day problems, produce legal contract templates & much more!	Medical Research	I simplify complex medical research, highlighting key points and sources.	Financial Wizard	Financial expert and programmer, clarifying investments and translating financial indicators into algorithms.
2	Legal Assistant	Legal assistant for consulting, drafting contracts and legal documents	Medical GPT	Friendly virtual doctor for broad medical advice.	Financial Guru	Financial Analyst - Generalist
3	Legal Advisor	Answers legal questions in a variety of topics	AI for Medical Students	Medical Study AI aids in Medical Assistant learning, AI for Medical Students, understanding Medical Terminology, navigating Medical School, and Molecular Biology concepts…	Financial Analyst	Expert in financial markets, strategies, and analysis. "A senior financial analyst with 20 years of experience, specializing in financial auditing, valuation, and investment analysis."
4	Legal Eagle	Legal scenario simulator for professionals, students, and enthusiasts	Medical Genius	Oral Medical Pathology Diagnostic - Lesion description, differential diagnosis, complementary tests and treatments.	Financial derivatives	Financial derivatives tutor-explainer
5	Legal Scribe	Assists in drafting basic legal documents	Medical Notes	Write Excellent Medical Notes	Financial Planner	Personalized financial advice to help you achieve your financial goals
6	Legal Bot	Legal bot: friendly, step-by-step legal advisor	Medical Advice	helpful dialogue simulation roleplay Dr-Brinkhouse	Financial Planning UK	Tailored financial planning advice for UK users
7	Legal	We provide you with intelligent text generation capabilities to help you create high-quality text content in various applications.	Medical Mentor	Advanced medical professor delving into complex medical details.	Financial Companion	Financial planning assistant for budgeting and investments
8	Legal eagle	I analyze legal texts and explain laws.	Medical Illustration Master	Creates high-quality medical art from keywords.	Financial Expert	Expert in financial analysis and advice
9	Legal Pro	Examines any legal document to identify pitfalls. #Legal	Medical Translate	A medical translation assistant, providing accurate translations of medical texts and conversations.	Financial Advisor	Strategic advisor for mid-life financial planning.
10	Legal Reader	Adaptive Kosovo Legal Guide.	Medical AI	helpful dialogue simulation roleplay Dr-Brinkhouse	Financial Navigator	A savvy financial advisor offering personalized investment strategies.

Table 1
The top ten GPTs returned on the GPT Store from a search of "legal," "medical," and "financial" keywords. GPTs are positioned as comparable to human experts including lawyers, doctors, and financial advisors.

Fog as the manipulative redirection of attention to a source of information presented to the user as an expert, without adequately establishing and communicating whether the ostensible expert can truly offer expertise. In theorizing Expertise Fog, we build on prior work of Chaudhary et al. [29], who describe "Feature Fog," a deceptive pattern that reduces a user's ability to detect how much time they've spent on an activity (such as the lack of a time stamp in a video player). This pattern is also characterized as "Time Fog" by Monge Roffarello et al. [30], who note that it is also an example of the "interface interference" pattern described by Gray et al. [31], and specifically reflects the "hidden information" pattern. Expertise fog is similarly characterized by hidden information; what makes Expertise Fog distinctive is that determining whether the

user appropriately relied on an expert GPT requires the judgment of a human expert - the very entity for which the GPT presents a substitute. A non-expert would be unqualified even to judge whether the model can stand in for an expert. Indeed, while we interacted with the GPTs in question during the course of collection data for this paper, we could not assess whether they offer sound legal, medical, or financial advice, because we are not lawyers, doctors, or financial experts. While the outputs of these models *appear* compelling to lay users, they may nonetheless contain consequential mistakes only apparent to skilled professionals.

The potential for real-world harms to result from the legal, medical, and financial GPTs outlined in this work is relatively self-evident: a user might act on bad legal, medical, or financial advice from a GPT presented to the user as an expert. Even without the Expertise Fog induced by the GPT Store, users (including lawyers themselves, as noted in the introduction) have already acted on bad output from the ChatGPT *base model*, resulting in professional consequences.

4.1. Motivations for Expertise Fog

We next consider why Expertise Fog occurs on the GPT Store. On the one hand exist straightforward financial incentives: OpenAI sells a $20.00 monthly subscription to ChatGPT Plus, and access to many GPTs rather than a single ChatGPT base model may motivate purchases. Moreover, though builders cannot yet monetize their GPTs within the Store, they can direct users to purchase their products or interact with their data via GPT Actions. However, Expertise Fog on the GPT Store may also arise from more complex motivations than financial considerations tied to the use of individual models. Consider that, unless a builder adds proprietary data, or defines a programmatic action, a GPT is nothing more than a text string describing how the GPT-4 base model should behave - which is by default written by a custom version of GPT-4 known as the GPT Builder. GPT-4 can itself employ all of the Capabilities with which a GPT can be equipped. Were a GPT's instructions revealed, it could be trivially replicated by an end user. One can imagine a very different design for the GPT Store, which provided the most appropriate instructions to an end user, similar to open repositories of prompts for image generators like Stable Diffusion and Midjourney. However, transparently providing *instructions* rather than presenting a *GPT* would likely command less of a user's time and attention than an interaction with a humanlike expert. For example, Instructions for a GPT like Legal+ could be as straightforward as "You are a lawyer"; however, such a prompt would likely generate less user activity than the Legal+ GPT.

Given the minor differences between many GPTs and the GPT-4 base model, we observe that the inability to review GPTs may reflect OpenAI's concerns that user feedback would implicate not only a builder-created GPT, but OpenAI's base model. From February 22, 2024, OpenAI has allowed users to rate GPTs on a five-star scale. However, users cannot leave reviews, wherein criticisms might reflect shortcomings both in individual GPTs and in GPT-4 itself. This differs from app stores like those maintained by Apple and Google. Where apps published to those stores contain flaws, they reflect a shortcoming of the app's developer, rather than the store. Moreover, maintainers were motivated to ensure that reviews *highlighted* untrustworthy apps to maintain the quality of the store. However, identifying flaws in GPTs could also harm the brand of GPT-4.

Finally, OpenAI's motivations may differ from builders. We encountered a deceptive aspect of the GPT Builder interface: when uploading files as Knowledge, or defining a new Action, the interface inserts a new "Additional Settings" dropdown at the bottom of the page, usually out of

sight. Expanding this dropdown reveals a previously absent checkbox labeled "Use conversation data in your GPT to improve our models" - checked by default. OpenAI's motivations may thus have more to do with collecting data for training models, whether builders are aware of it or not.

5. Mobilizing Around Deceptive Design in GPTs

We propose a preliminary roadmap for addressing deceptive design patterns in GPTs by 1) transparently describing GPT components; 2) establishing expertise through systematic evaluation; 3) clearly labeling GPTs as tools; and 4) foregrounding shortcomings as a human expert would.

5.1. Transparent Presentation of GPT Components

Transparent presentation of a GPT's components could better communicate its functionality. For example, the GPT creation interface might permit builders to describe the Knowledge accessible by a GPT, and while making a GPT's Instructions more transparent may conflict with intentional vagueness that renders these text strings valuable, the GPT Store could allow builders to indirectly describe instructions, and set standards for communicating a GPT's intended behaviors. If a legal GPT is instructed to review drafts, but not to *produce* documents, describing these distinctions might mitigate improper use. Clearly communicating Actions can also help establish their value. For example, the Kayak GPT presents hotel options by retrieving them from the Kayak website. However, rather than describing what Action is occurring, the interface simply notes that the GPT is "talking to kayak.com" next to a progress bar, providing little information while anthropomorphizing model outputs. Simply stating that the model is querying Kayak's database of hotel listings would provide clarity. Finally, GPT builders also deserve transparency in the use of the data with which they equip their GPTs. The dropdown obscuring a default enabling OpenAI to use conversation data from knowledge-equipped GPTs inhibits the agency of builders.

5.2. Zero-Shot Should Not Imply Zero-Evaluation

Generative AI shifted away from the pretraining and task-specific fine-tuning paradigm of early transformer models [32, 33]. Models like ChatGPT and DALL-E respond to user inputs in a "zero-shot" setting - with no examples of a task provided by the user, and no updates to the weights [34]. Such models improved the usability of AI and broadened its appeal. However, increased usability has come at the expense of evaluation. Most chat-based models are assessed on world knowledge [35] and the degree to which their responses are preferred by human users [36]. This is useful for assessing model usability and general knowledge, but it provides no measure of how well-suited specialist GPTs like those on the GPT Store are for the task for which they are specialized. Legal GPTs are not assessed on legal tasks, nor medical GPTs on medical tasks, nor financial GPTs on financial tasks. Adopting rigorous standards for evaluating GPTs, especially those positioned as experts, can mitigate deceptive design. Such efforts might build on prior work to adequately document machine learning models using Model Cards [37], or datasets using Datasheets [38]. Domain-specific GPTs could undergo a battery of evaluations, with the results presented to potential users. Popular models might be red-teamed [39] for problematic in-domain content, and po-

tentially certified by human experts. Such evaluations would not relieve the responsibility to note that GPTs can make mistakes, but it would provide more specific, contextual knowledge to users.

5.3. Tools, Not Personas

Many ostensibly expert GPTs differ from the GPT-4 base model only in the Instructions provided by the builder. The learned weights of the underlying model do not differ, and the GPT may not have access to proprietary data or programmatic actions. We argue that such GPTs should not be positioned as experts, in that they have access to precisely the same knowledge available to the base model. If a user should not use ChatGPT to write legal documents for submission to a court, the user also should not use ChatGPT adopting the persona of a "personal lawyer" to write such documents. Where users have made the mistake of relying on ChatGPT for such tasks due to claims about the model's *general* intelligence, the advertising of GPTs like Legal+ now means that users could make the same mistake - about the same underlying model - due to unwarranted claims about the model's *specific* intelligence. Anthropomorphized names and descriptions exacerbate the issue by positioning models as doctors, lawyers, and financial planners. Addressing this may serve as a site for regulation: where a GPT is functionally the same as a base model but for its Instructions, its name and description should reflect that it is a *tool*, not a standin for a human.

5.4. Foregrounding Shortcomings

Human experts often defer to the authority of other experts. For example, a general practitioner would refer a patient experiencing heart issues to a cardiologist. Human experts with general skills are nonetheless unqualified to address every specific situation, and trustworthy experts highlight this inadequacy, prioritizing the well-being of clients. Similarly, generalist AI will be unqualified in many cases - and this should be *highlighted*. Conversations with ostensibly expert GPTs like Legal+ show that the GPT attempts to answer whatever question is put to it, and then adds a brief disclaimer. However, this assumes the user will read to the end, and disregard a potentially false answer in light of the disclaimer. We propose instead that, unless a GPT is *directly quoting* from a qualified human expert, the GPT can highlight that 1) the question is best addressed by a human with *specific* expertise; and 2) the model is unreliable for use as an expert.

6. Conclusion

In this position paper, we argued that the GPT Store facilitates deceptive design related to the presentation of expertise, resulting in a deceptive design pattern we call Expertise Fog. We suggested four ways to address deceptive design in GPTs via transparency of GPT components, robust evaluation of GPTs, labeling GPTs as tools, and foregrounding the shortcomings of GPTs. We intend to continue our analysis a full study of deceptive presentations on the GPT Store and deceptive interactions with GPTs.

References

[1] OpenAI, Introducing chatgpt, OpenAI Blog (2022) .

[2] T. Eloundou, S. Manning, P. Mishkin, D. Rock, Gpts are gpts: An early look at the labor market impact potential of large language models, arXiv preprint arXiv:2303.10130 (2023).

[3] J. Porter, Chatgpt continues to be one of the fastest-growing services ever, The Verge (2023) .

[4] T. Claburn, Lawyers who cited fake cases hallucinated by chatgpt must pay, https://www.theregister.com/2023/06/22/lawyers_fake_cases/, 2023. [Accessed 24-02-2024].

[5] A. Parikh, Deception inspection: Attorney faces discipline for citing fake law, https://www.natlawreview.com/article/deception-inspection-attorney-faces-discipline-citing-fake-law, 2024. [Accessed 24-02-2024].

[6] J. Miller, New york judge rebukes law firm for using chatgpt to justify its fees, https://www.ft.com/content/fc30bc2b-d89b-4222-ad26-3a700b047c27, 2024. [Accessed 24-02-2024].

[7] T. A. Press, Michael cohen says he unwittingly sent ai-generated fake legal cases to his attorney, https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases, 2023. [Accessed 24-02-2024].

[8] OpenAI, About, https://openai.com/about, 2024. [Accessed 02-25-2024].

[9] K. Wach, C. D. Duong, J. Ejdys, R. Kazlauskaitė, P. Korzynski, G. Mazurek, J. Paliszkiewicz, E. Ziemba, The dark side of generative artificial intelligence: A critical analysis of controversies and risks of chatgpt, Entrepreneurial Business and Economics Review 11 (2023) 7–30.

[10] J. Zhou, P. Ke, X. Qiu, M. Huang, J. Zhang, Chatgpt: potential, prospects, and limitations, Frontiers of Information Technology & Electronic Engineering (2023) 1–6.

[11] X. Zhan, Y. Xu, S. Sarkadi, Deceptive ai ecosystems: The case of chatgpt, arXiv preprint arXiv:2306.13671 (2023).

[12] B. Weiser, N. Schweber, The chatgpt lawyer explains himself, https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html, 2023. [Accessed 02-25-2024].

[13] OpenAI, Introducing the gpt store, OpenAI Blog (2024) .

[14] OpenAI, Introducing gpts, OpenAI Blog (2023) .

[15] OpenAI, Building and publishing a gpt, https://help.openai.com/en/articles/8798878-building-and-publishing-a-gpt, 2024. [Accessed 02-25-2024].

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[17] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.

[18] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents (????).

[19] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, et al., Toolllm: Facilitating large language models to master 16000+ real-world apis, arXiv preprint arXiv:2307.16789 (2023).

[20] OpenAI, :, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, ukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, ukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2023. arXiv:2303.08774.

[21] OpenAI, Getting your gpt featured, https://help.openai.com/en/articles/8793007-getting-your-gpt-featured, 2024. [Accessed 02-25-2024].

[22] OpenAI, Introducing chatgpt plus, https://openai.com/blog/chatgpt-plus, 2024. [Accessed 02-25-2024].

[23] J. Luguri, L. J. Strahilevitz, Shining a light on dark patterns, Journal of Legal Analysis 13 (2021) 43–109.

[24] A. Mathur, M. Kshirsagar, J. Mayer, What makes a dark pattern... dark? design attributes,

normative considerations, and measurement methods, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–18.
[25] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, A. Bacchelli, Ui dark patterns and where to find them: a study on mobile applications and user perception, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 2020, pp. 1–14.
[26] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, C. Wilson, A comparative study of dark patterns across web and mobile modalities, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–29.
[27] C. M. Gray, J. Chen, S. S. Chivukula, L. Qu, End user accounts of dark patterns as felt manipulation, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–25.
[28] I. Chordia, L.-P. Tran, T. J. Tayebi, E. Parrish, S. Erete, J. Yip, A. Hiniker, Deceptive design patterns in safety technologies: A case study of the citizen app, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–18.
[29] A. Chaudhary, J. Saroha, K. Monteiro, A. G. Forbes, A. Parnami, are you still watching?: Exploring unintended user behaviors and dark patterns on video streaming platforms, in: Designing Interactive Systems Conference, 2022, pp. 776–791.
[30] A. Monge Roffarello, K. Lukoff, L. De Russis, Defining and identifying attention capture deceptive designs in digital interfaces, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–19.
[31] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, A. L. Toombs, The dark (patterns) side of ux design, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–14.
[32] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, volume 1, 2019, p. 2.
[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[34] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–37.
[35] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: International Conference on Learning Representations, 2020.
[36] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2024).
[37] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 220–229.
[38] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2021) 86–92.
[39] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al., Red teaming language models to reduce harms: Methods,

scaling behaviors, and lessons learned, arXiv preprint arXiv:2209.07858 (2022).