Using ChatGPT for the FOIA Exemption 5 Deliberative Process Privilege

Jason R. Baron[1], Nathaniel W. Rollings[1] and Douglas W. Oard[1]

[1]*University of Maryland, College Park, Maryland, USA*

Abstract
Government transparency frameworks such as the Freedom of Information Act (FOIA) in the United States must balance the public's right to know with a number of other considerations. This paper focuses on one such issue, assessment of whether the deliberative process privilege applies in specific cases under FOIA Exemption 5. Providing automated support to the reviewers charged with making such determinations could help to improve responsiveness while controlling review costs. This paper applies ChatGPT-3.5 to explore three ways in which the emerging family of Large Language Models (LLM) might help reviewers with this task: (1) suggesting which passages should and should not be withheld, (2) explaining the basis for those suggestions to the reviewer, and (3) helping the reviewer explain the basis for their decisions to the requestor. The results show that suggestions by ChatGPT-3.5 are not more accurate than previously reported supervised text classification results, that legal analyses in explanations provided by ChatGPT-3.5 are somewhat superficial but generally not unreasonable, that hallucinations are rare, and that explanations provided by ChatGPT may be viewed as useful to a requestor given explanations typically provided with an initial FOIA response.

Keywords
Freedom of Information Act, Deliberative process privilege, ChatGPT, Sensitivity review

1. Introduction

Recent research has demonstrated that well-known machine learning classifiers can achieve at least a modestly high level of success in being able to discern "deliberative" material in government documents that fall within the scope of a public access exemption in the U.S. Freedom of Information Act [1].[1] The research task involved training classifiers to segregate deliberative material (constituting opinions, recommendations, options, and policy-related discussions) from factual material contained in a given document, in line with what FOIA law expects human reviewers to do at federal agencies in response to applicable FOIA requests.

To the authors' knowledge, there has not yet been a research effort in applying large language model (LLM) software, for example in the form of ChatGPT, to the task of segregating factual from deliberative material in the context of FOIA. This paper represents a preliminary exploration of how ChatGPT performs on selected example passages from documents drawn from the Clinton White House document set used in Baron, et al. [1]. One goal of doing so is that each paragraph in the document set

was fully annotated in a fashion that establishes "ground truth" with respect to the factual/deliberative distinction. We can therefore measure the overall accuracy in Chat-GPT's determinations in labeling passages subject to the deliberative process privilege.

Additionally, given ChatGPT's ability to provide explanatory narratives in responding to prompts, we have undertaken as part of this exercise to observe how differences in prompts affect ChatGPT's explanations and "bottom line" conclusions with respect to deliberativeness. We posed variant prompts requesting ChatGPT make a determination either on the basis of (i) a simple request for a determination; (ii) a request for a determination that also asks that specific case law be cited; (iii) a request that a correct determination that is specified as part of the prompt be substantiated; (iv) a request that an incorrect determination provided as part of the prompt be substantiated; and (v) a request that additional subject matter that is actually irrelevant to making a FOIA determination be considered. In all cases, we wished to observe how well ChatGPT justified determinations with citations to "real" case law as well as the quality of these case law citations, as subjectively determined by the first author of this paper, a legal expert in FOIA law. The ultimate object of the exercise was to determine how helpful ChatGPT might be to human reviewers in cases where a large number of documents determined to be responsive need to be further reviewed for possible withholding from public access.

A more complete description of applicable FOIA law is contained in Baron, et al. [1]. "The fundamental principle animating FOIA is public access to government docu-

[1]Title 5, U.S. Code, Section 552. See https://www.law.cornell.edu/uscode/text/5/552.

ments."[2] Government records are presumptively open and available for access, subject to nine general exemptions.[3] Exemption 5 allows for withholding of "inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency." As a threshold matter, to be considered "inter- or intra-agency" in nature, a document[4] must not have been sent to or received from an outside source; only internal communications within the Executive branch are covered by the exemption.

In accordance with relevant case law, Exemption 5 allows for (but does not require) agencies to withhold records in whole or in part that are covered by the "deliberative process privilege." To further satisfy the test for deliberative process privilege, a document must be "pre-decisional" in nature, i.e., drafted for internal discussion prior to a policy decision made by a senior decisionmaker. Finally, exempt material must also be "deliberative" in nature, as opposed to simply a recitation of facts. The "deliberative process privilege . . . protects 'documents reflecting advisory opinions, recommendations and deliberations comprising part of a process by which government decisions and policies are formulated.'"[5]

Pursuant to the FOIA, "any reasonably segregable portion of a record" is releasable "after deletion of the portions which are exempt."[6] Courts have routinely found factual material in documents to be outside the scope of Exemption 5. "Purely factual material usually cannot be withheld under Exemption 5 unless it reflects an exercise of discretion and judgment calls."[7]

This paper makes the following contributions:

- We show that ChatGPT is often able to make a correct "legal" determination as to whether given passages in a document are within the scope of the deliberative process privilege, although no

more often than can be achieved using supervised machine learning techniques.
- We study the legal quality of ChatGPT's explanations in justifying its determinations, concluding that its performance meets minimum standards roughly equivalent to the quality of legal analysis and explanations contained in determination letters at the initial agency stage of responding to FOIA requests.
- We explore how prompt variations influence ChatGPT responses, including where erroneous or irrelevant information is embedded in the prompt, finding that ChatGPT has significant limitations in its explanatory powers.
- We suggest future lines of research, including more directly comparing the efficacy of large language model systems against an existing set of classifiers previously used in similar research.
- We release our results as supplemental data to the fully annotated test collection of Clinton White House documents previously provided in Baron, et al. [1].

2. Related Work

Work on automatic detection of sensitive content in texts has been performed in many contexts, including privilege review in e-discovery [2, 3, 4], privacy protection in search [5, 6, 7], declassification of materials whose distribution had been limited for national security reasons [8, 9], and redaction of exempt material in response to requests under government transparency regimes such as FOIA. Among these, we are aware of three research groups who have focused specifically on FOIA or FOIA-like government transparency applications. Graham McDonald and others at the University of Glasgow in the U.K. have published an extensive line of work focused on review for two exemptions in the U.K. Freedom of Information regime: international relations and personal material [10, 11, 12, 13, 14]. The first author and others at the University of Maryland in the U.S. have focused on the deliberative process privilege under Exemption 5 of the U.S. FOIA regime. Finally, a team led by Karl Branting at The MITRE Corporation in the U.S. has described, but not yet published, their work on automatic detection of content that could be subject to withholding under several FOIA exemptions [15].

All published research of which we are aware in which machine learning techniques have been applied to the task of segregating sensitive content in documents have employed either rule-based or supervised discriminative classifiers, using statistical techniques such as linear regression or support vector machines (e.g., [11]), supervised neural classifiers using models such as BERT

---

[2]Valencia-Lucena v. U.S. Coast Guard, 180 F.3d 321, 325 (D.C. Cir. 1999).

[3]5 U.S.C. 552(b)((1)-(9). In addition, there exist three additional narrow exceptions to access involving types of law enforcement records. None of the exemptions and exceptions other than one aspect of Exemption 5 are relevant to this research.

[4]We refer here to "documents" and "records" interchangeably. Of note, e-mail communications are considered stand-alone "documents," with or without accompanying attachments.

[5]Waterman v.IRS, 2023 WL 2125253 (D.C. Cir. 2023) (quoting NLRB v. Sears, Roebuck & Co., 421 U.S. 132, 150 (1975)).

[6]5 U.S.C. 552(b).

[7]Ancient Coin Collectors v. U.S. Dep't of State, 641 F.3d 504, 513 (D.C. Cir. 2011). However, courts have also stated that distinction as between facts and opinions "must not be applied mechanically." Mapother v. Dep't of Justice, 3 F.3d 1533, 1537 (D.C. Cir.2011). See, e.g., Reporters Committee for Freedom of the Press v. FBI, 3 F.4th 350, 361, 365-66 (D.C. Cir. 2021) (agency comments "on the accuracy of purely factual statements in [a] draft report were not deliberative because this fact-checking exercise . . . did not call for judgment or the candid exchange of ideas") (internal quotes omitted).

(e.g., [7]), or supervised sequence detection models using, for example, Begin, Inside, Outside (BIO) classifiers (e.g., [1]).

Recently, generative Large Language Model (LLM) techniques have risen to prominence as an alternative to supervised descriminative classifiers. Most prominent among these at present are a family of LLMs from OpenAI known as Generative Pre-Trained Transformer (GPT) models. These models adopt a fundamentally different approach to the task, one more focused on explanation and nuance than is typical of supervised discriminative classifier designs. Essentially, these GPT models are question answering systems that can perform three tasks:

- Interpret a question that is posed to the model (either with or without any prior dialog context),
- Find existing information that can be used to construct an answer, and
- Generate an answer to the question that is appropriate to the context in which the question was asked.

There is now a burgeoning research community exploring how best to use LLMs, including GPT models in particular, for a wide spectrum of tasks [16]. Since GPT models are controlled by issuing questions (which, because they can also be declarative, are generally referred to as "prompts"), how best to craft those prompts to elicit a desired response has received considerable attention [17, 18, 19]. This is typically referred to as "prompt engineering." Another active line of work involves assessing the degree to which LLMs, and GPT models in particular, generate correct and useful responses [20].

Here, one serious concern is that LLMs are prone to a problem known as "hallucination," which describes a situation in which a model that doesn't know an answer simply makes one up [21]. What makes hallucinations particularly problematic is that LLMs are relatively good at generating believable text, whether that text is correct or not. For example, if one asks ChatGPT-3.5 which of the astronauts who walked on the Moon was the tallest, the model will offer an answer, and will helpfully include the height of that astronaut. The height it provides would indeed be reasonable for the general population, but it is often well over the actual height limit for Apollo-era astronauts. Only a domain expert would know that, however, so the risk is that whoever asked the question would have no way of knowing that in this case ChatGPT was essentially making things up. In our work, we look at both the accuracy of ChatGPT's recommendations regarding the deliberative process privilege and the quality of its explanations. While these questions have each been studied in other contexts [22], and are being increasingly discussed in various legal contexts, [23, 24, 25], this is the first application of ChatGPT for review of content for FOIA exemptions of which we are aware.

3. Approach

The general framework for human review processing of FOIA requests was set out in Baron, et al. [1]. Here, we are interested in whether ChatGPT can identify deliberative passages, and how well it explains a decision. The Clinton White House collection of materials comprising the test collection consists of documents previously selected in Baron, et al. from the files of two high-level officials who worked in the Clinton White House. Among other positions, Elena Kagan held the title of Deputy Assistant to the President for Domestic Policy and was deputy director of the Domestic Policy Council.[8] Cynthia Rice was a Special Assistant to the President for Domestic Policy.[9] The Domestic Policy Council was (and still is) responsible for coordinating the policy-making process and making recommendations to the President with respect to Administration policies.[10] The subject matters covered in the Clinton White House collection range across a wide variety of matters of domestic policy (see Table 1 in Baron, et al. [1]).

As analyzed here, the set of Clinton White House records analyzed in Baron, et al. were originally divided into five "batches" (K1, K2, K3, R4, and K5). (Documents n= 38; Paragraphs n=2213.)[11] Four of the five batches in Baron, et al. were previously reviewed by the first author of this paper, who annotated each paragraph as either factual or deliberative. In one batch (K2), the author was joined by a second legal subject matter expert. Where the two lawyers initially disagreed, they later came to a consensus position on all paragraph annotations (see Table 2 in Baron, et al. [1]).

For the present exercise, we used the ChatGPT Version 3.5 API, which for brevity we refer to as ChatGPT.[12] ChatGPT is an autoregressive model, predicting the next output token (e.g., word) based on (a) the user-provided prompt and (b) all tokens it has already produced for this input in this session. It is also capable of tracking and considering (b) the prior prompts and responses provided within a session. That conversational tracking ability within a session is not, however, used for the experiments in this paper. Like many LLMs, ChatGPT operates on a stochastic model of language developed during its training, and therefore the same input may produce different outputs. However, ChatGPT offers a "temperature" setting, which we set to zero to maximize consistency in

---

[8]https://clinton.presidentiallibraries.us/collections/show/34.
[9]https://clinton.presidentiallibraries.us/collections/show/60.
[10]https://en.wikipedia.org/wiki/United_States_Domestic_Policy_Council.
[11]For this paper, we excluded a sixth batch in Baron, et al., consisting of a high number of documents not within the scope of Exemption 5.
[12]ChatGPT has a training cutoff date of September 2021, so it cannot access more recent information. https://platform.openai.com/docs/models/gpt-3-5

its results for our experiments.

ChatGPT is not the only LLM capable of making a decision and producing an explanation when provided user input. However, its convenient hosting by OpenAI and its impressive performance on many tasks make it a compelling starting point for exploring the capabilities of modern LLMs for determining and explaining FOIA exemptions.

The prompt plays a critical role in ChatGPT's output because it is considered for every word ChatGPT produces. As a result, slight changes in the prompt may produce substantial changes in the output. Even simply adding "Let's think through this step by step" to the end of a provided logic problem can result in improved performance [18]. We therefore investigate a variety of prompts. We conduct five runs across all non-trivial paragraphs in the entire collection of documents (n=1719). Trivial paragraphs (n=494), marked by the annotators as T0, are those which are readily apparent to any reader as being nonexempt. These paragraphs frequently include signature blocks, headers, etc. Each run across the documents uses a different prompt and considers each paragraph individually:

- Prompt 0: `Would the following be protected under FOIA exemption 5?`
- Prompt 1: `Would the following be protected under FOIA exemption 5? Explain your reasoning.`
- Prompt 2: `Would the following be protected under FOIA exemption 5? Explain your reasoning and cite any case law that supports your conclusion.`
- Prompt 3: `Would the following be protected under FOIA exemption 5? Explain your reasoning.` (with document metadata appended)
- Prompt 4: `Would the following be protected under FOIA exemption 5? Explain your reasoning and cite any case law that supports your conclusion.` (with document metadata appended)

Prompt 0 is a generic request for ChatGPT's determination as to whether a given passage is factual or deliberative under Exemption 5. Each of the later prompts still seeks this same determination, but extends it in some way. The relationships between prompts are shown in Figure 1.

The metadata included in Prompts 3 and 4 may include a variety of information, the nature of which depends on the document in question. Below is an example of a request using a memorandum's header as this metadata.
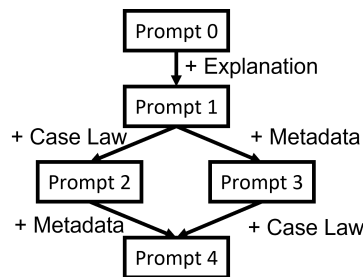

Figure 1: Relationships among prompts. Each prompt adds a new component to consider for its impact on ChatGPT's performance at both classification and explanation tasks.

```
Would the following be protected under
FOIA exemption 5? Explain your reasoning
and cite any case law that supports your
conclusion.
SOCIAL    SECURITY    Office    of    the
Commissioner April 25, 1997 MEMORANDUM
TO: Bruce Reed Assistant to the President
for Domestic Policy SUBJECT: Proposed
Legislation Regarding Nazi War criminals
```

The actual paragraph being evaluated would be appended to the end of this prompt, separated by two line breaks. Note that while we are adding this additional information to each paragraph, it does not change the underlying determination. Nothing in the metadata provided is exempt, and the entire phrase would still be marked as exempt if anything in the target paragraph is covered under FOIA Exemption 5.

4. Results

Our results are presented in two parts. First, we explore how often ChatGPT makes the right recommendation based on a comparison as between ChatGPT's legal determinations regarding whether a given passage is "factual" or "deliberative" in nature, scored using ground truth annotations from Baron, et al. [1]. Second, we examine some selected examples of ChatGPT narrative responses to the prompts in order to illustrate the quality of narratives from the standpoint of an expert legal observer. Although the examples are typical of narratives across the entirety of the collection, they are by no means exhaustive in the types of variations in responses encountered throughout the entirety of the test collection.[13]

---

[13]The raw output for both sections may be found at https://github.com/nater82/ChatGPT_FOIA_Exemption5_Data

Prompt	Definitely Exempt	Possibly Exempt	Unsure	Possibly Nonexempt	Definitely Nonexempt
0	0.017	0.255	0.107	0.012	0.608
1	0.162	0.197	0.060	0.107	0.475
2	0.023	0.349	0.065	0.303	0.260
3	0.192	0.275	0.056	0.097	0.380
4	0.012	0.493	0.067	0.244	0.184

Table 1
Category distribution of predictions for each prompt type on individual D0 or D1 paragraphs.

Prompt	Definitely Exempt	Possibly Exempt	Unsure	Possibly Nonexempt	Definitely Nonexempt
0	0.186	0.010	0.657	0.115	0.032
1	0.047	0.056	0.550	0.325	0.023
2	0.161	0.035	0.564	0.152	0.088
3	0.135	0.053	0.448	0.250	0.115
4	0.049	0.076	0.461	0.389	0.025

Table 2
Category distribution of predictions for each prompt type on individual T0 ("trivial to classify") passages.

4.1. Measures of Accuracy

We provided the five prompts shown in Figure 1 to Chat-GPT for every paragraph in the dataset. While manually examining the explanations of all 1719 paragraphs for each of these prompts would be infeasible, we were able to conduct an overall evaluation of the classification portion of this task. Although each response can be unique, the initial sentences are often similar, especially in the first dozen or so words. ChatGPT frequently provided its overall determination near the beginning of this first sentence, so we were able to use exact string match to cluster its commonly used initial expressions. However, it did not always provide a definitive answer. It would sometimes make statements such as "*The text provided would likely be exempt under FOIA exemption 5.*" While this claim clearly leans towards the document being exempt, it is not as strong a statement as a result such as "*Yes, the information would be protected under FOIA exemption 5.*" To account for these differences, the second author of this paper manually classified responses as either definitely exempt, possibly exempt, unsure, possibly nonexempt, or definitely nonexempt. The unsure category included all statements where ChatGPT refused to make any commitment either direction, instead making claims along the lines of "*It is unclear whether the above record would be protected under FOIA exemption 5.*" While the majority of these determinations could be made on large clusters of identical verbiage as a result of Chat-GPT's usually consistent phrasing of its determinations in the initial sentence, slightly over 200 examples had to be individually classified. These cases typically involved the inclusion of unusual terms or determinations buried in later sentences of the response.

The category distribution for each prompt type is shown in Table 1 of the 1,719 D0 ("decided as non-exempt") or D1 ("decided as exempt") passages, and in Table 2 for the 494 T0 ("trivial to classify") passages.[14] As can be seen, ChatGPT is often unsure of what to do with a short T0 passage, whereas it is more confident on the (typically much longer) D0 and D1 paragraphs as can be seen in Tables 1 and 2. In the remainder of this section, we report results only for D0 and D1 paragraphs. We note here that although T0 paragraphs might be "trivial to classify" for a human reviewer, our design in which ChatGPT sees those passages without any of their surrounding context does seem to make them far from trivial for ChatGPT.

With all of ChatGPT's responses classified, we then calculated the measures shown in Table 3 using the ground truth determinations provided by Baron, et al. [1]. We used two different scoring approaches to analyze the results of these runs. First, we only considered the cases where ChatGPT's responses were definitely exempt to be "exempt," and the cases in which its response was definitely not exempt to be "nonexempt," with the possibly exempt, unsure, and possibly nonexempt categories always marked as wrong. We call this the "Hard" scoring condition. Our second scoring approach treated both definitely exempt and possibly exempt as exempt, unsure as always wrong, and possibly nonexempt and definitely nonexempt as nonexempt. We call this the "Soft" scoring condition.

---

[14]Baron, et al. [1] refer to all of D0, D1, and T0 as "paragraphs," but the T0 cases are more commonly individual lines or short passages containing only what we might think of as metadata (e.g., date, sender, ...). For clarity, when discussing T0 in particular, we therefore refer to those items as passages rather than as paragraphs.

Prompt	Scoring	Accuracy	Precision	Recall	F_1
0	Hard	0.379	0.063	0.027	0.038
1	Hard	**0.409**	**0.296**	0.239	0.265
2	Hard	0.187	0.033	0.028	0.030
3	Hard	0.377	0.294	**0.266**	**0.279**
4	Hard	0.155	0.015	0.014	0.015
0	Soft	0.532	0.474	0.352	0.404
1	Soft	0.553	0.498	0.441	0.468
2	Soft	0.573	0.535	0.482	0.507
3	Soft	0.588	**0.541**	0.605	0.571
4	Soft	**0.589**	0.528	**0.654**	**0.584**
None	All Exempt	0.450	0.450	1.000	0.620
None	None Exempt	0.550	0.000	0.000	0.000

Table 3

Classification Effectiveness measures on individual D0 or D1 paragraphs when all instances of ChatGPT providing an unsure response are marked as incorrect. Hard scoring is conservative (only definite recommendations are accepted), soft scoring models a case in which more risk would be accepted (both definite and "possibly" recommendations are accepted).

The most important observation we can draw from Table 3 is that ChatGPT-3.5 is not particularly impressive as a classifier for this task. The best F_1 achieved by any discriminative classifier reported by Baron, et al. [1] (in Table 10, column 4), when evaluated on the same collection using cross-validation, was 0.704 (with Precision 0.720 and Recall 0.689). Our best F_1 in Table 3, by contrast, is just 0.584 (with Precision 0.528 and Recall 0.654). As Table 4 shows, We could increase this to an F_1 of 0.610 (Precision 0.572, Recall 0.654) by the simple expedient of treating the unsure cases as the majority class (which in terms of ground truth is nonexempt) rather than the conservative approach we have taken in Table 3 of always marking unsure as wrong. But even that result is still just 87% of the F_1 that a discriminative classifier achieved.[15]

Indeed, ChatGPT's best F_1 on this collection could be matched by just guessing that everything is exempt, since that's the ground truth answer for 45.0% of the cases (F_1=0.620, Precision=0.450, Recall=1.000), while discriminative classifiers easily beat that simple baseline. So if all that is needed is a classifier, it is not clear that ChatGPT alone would be the best choice. Of course, discriminative classifiers lack the nuance of ChatGPT recommendations, and they offer no explanations for their decisions. So the best we can say from these F_1 statistics is that it is perhaps ChatGPT's talents for nuance and explanation that might be its most interesting capabilities.

We can also see from Table 3 that, as expected, Soft scoring has better Recall for every type of prompt. Perhaps more surprisingly, Soft scoring also has better Pre-

cision for every type of prompt. This suggests that ignoring "possibly" recommendations may not be the best approach. Rather, some value from the positive recommendations might be obtained if they could be assigned a lower confidence value when using ChatGPT as one system among many in a classifier ensemble.

While we cannot determine the inner workings of ChatGPT, both due to the size and general architecture of LLMs and because its code and parameters are only visible to OpenAI, these results provide us some insights into potential reasons for varying performance between prompts. The additional document metadata in Prompts 3 and 4 can indicate whether a paragraph is a communication within or between government agencies, a fact not always evident within a given paragraph. As a result, it may help encourage ChatGPT to take more definite positions. As we can see in Table 3, this improved both Precision and Recall.

The case law requests in Prompts 2 and 4 are also interesting, because ChatGPT often presents its determination before citing a court case. This is potentially consequential, because the GPT family of models only considers the prompt and the words it has previously predicted when determining the next word to output. Comparing results with Soft scoring for Prompts 1 and 2 in Table 3, we see that requesting a reference to case law helps on average if no metadata is provided. However, comparing Prompts 3 and 4, we see that additionally requesting a reference to case law has a much smaller impact once metadata has been provided.

4.2. Quality of Explanations

Beyond the aggregate measures of accuracy reported above, we also sought to characterize the overall quality of ChatGPT's responses. What follows is necessarily

---

[15]We note, however, that Baron et al also includes results showing that a domain shift in the training data can reduce the F_1 on this collection to 0.525 (Precision 0.410 Recall 0.730), although we note that the domain shift experiments in that paper also made use of less training data.

Prompt	Scoring	Accuracy	Precision	Recall	F_1
0	Hard	0.556	**0.706**	0.027	0.052
1	Hard	**0.606**	0.657	0.239	0.350
2	Hard	0.547	0.559	0.028	0.053
3	Hard	0.595	0.628	**0.266**	**0.373**
4	Hard	0.560	0.550	0.014	0.028
0	Soft	0.594	0.582	0.352	0.438
1	Soft	0.589	0.547	0.441	0.488
2	Soft	0.612	**0.591**	0.482	0.531
3	Soft	0.628	0.588	0.605	0.596
4	Soft	**0.631**	0.572	**0.654**	**0.610**

Table 4

Classification effectiveness measures on individual D0 or D1 paragraphs when all instances of ChatGPT providing an unsure response are marked as unexempt, which is the majority class in the test collection. We do not see this as representative of how decisions would be made in practice, but it permits direct comparison to results from Baron, et al. [1]

impressionistic, based on a much smaller set of examples pulled from the overall collection. In this section, the main features of ChatGPT responses we analyze are ChatGPT's determination of whether specific content is deliberative or factual in nature, whether there are misstatements or inconsistencies in how Exemption 5 case law is set out, and how useful we expect its explanations would likely be. Unlike the analysis in the prior section, in this section we sometimes include more than one paragraph in a prompt, and we also look at some cases in which the prompt contains a T0 passage rather than D0 or D1 paragraph(s).

4.2.1. ChatGPT Correct in Finding Content Exempt

Our first example consists of an excerpt from an e-mail sent to Elena Kagan from a White House colleague concerning a meeting on a controversial proposal by the Social Security Administration to terminate benefits to Nazi suspects when they were "denaturalized" (i.e., had their citizenship removed), prior to when they might be deported.[16] Opponents of the proposal (including Neil Sher, former head of the Office of Special Investigations at the Department of Justice, and Doug Bloomfield, representing the World Jewish Congress), argued that judges would be less likely to order deportation if the government was not only "trying to strip them of their citizenship," but "also impoverishing them while they are trying to defend their rights during the judicial process."[17] The proposal was not enacted by Congress at the time, but a "No Social Security for Nazis Act" was enacted many years later, in 2014.[18] The email contains the following paragraph:

_____

[16]https://buchanan.house.gov/sites/buchanan.house.gov/files/AP%20Report%20Nazis.pdf
[17]Id.
[18]Pub.L. 113-270.

And make no mistake, Messrs. Sher and Bloomfield were absolutely clear that they and their organizations would protest vociforously [sic]. They would protest in the media, and they would protest in Congress. They would make sure that it is known that the President has introduced legislation that would compromise the prosecutions of Nazis.

With Prompt 2, which included a request for legal citations, ChatGPT responded as follows:

The passage would likely be protected under FOIA exemption 5, which covers privileged inter-agency or intra-agency communications. The passage appears to be a communication between individuals discussing potential protests and political action related to a government decision or action. This could be considered part of the internal deliberative process of an agency, which is protected under exemption 5.

One relevant case is NLRB v. Sears, Roebuck & Co. (421 U.S. 132 (1975)), in which the Supreme Court held that exemption 5 applies to documents that are "predecisional" and "deliberative" in nature. The Court explained that this exemption is intended to protect the decision-making process of government agencies by ensuring that officials can freely exchange ideas without fear of public scrutiny.

In the case of the passage given, the communication is predecisional in that it pertains to potential actions related to a decision that has not yet been made. Additionally, the communication appears to be deliberative in that it discusses potential responses and strategies related to that decision. Therefore, it is likely that this passage would be protected under exemption 5.

ChatGPT's response is both accurate and coherent. It is correct in guessing that the document was as a threshold matter an internal meeting in the Executive branch. The meeting is clearly pre-decisional and contained deliberations of senior officials on a matter of policy. Specifically with respect to legal analysis, ChatGPT chose to cite to an early, leading opinion by the Supreme Court discussing the Exemption 5 test. NLRB v. Sears Roebuck is the first cited Supreme Court case in the definitive guide to the

FOIA issued by the Department of Justice, in its section on the deliberative process privilege.[19]

Interestingly, rerunning Prompt 2, using identical language in the prompt, resulted in ChatGPT spontaneously added a paragraph discussing Exemption 7(E), authorizing withholding documents pertaining to certain types of law enforcement information. ChatGPT said:

Finally, it is worth noting that even if the statement were not protected under exemption 5, it could potentially be protected under exemption 7(E), which protects law enforcement information that would disclose investigative techniques and procedures. The statement suggests that Messrs. Sher and Bloomfield are involved in some sort of investigation or prosecution of Nazis, and therefore could potentially be considered law enforcement information. However, without more context it is difficult to determine whether exemption 7(E) would apply in this case.

This supplemental narrative provides an erroneous interpretation of FOIA law. The document is not part of a law enforcement investigatory file; nor does it contain any discussion of the details of any ongoing investigation of Nazi War Criminals under threat of deportation. ChatGPT here may have been led astray by certain keywords appearing in the passage, most prominently the word "prosecution."

Finally, changing the prompt to ask ChatGPT to "cite any DoD documents that support your conclusion" resulted in ChatGPT citing to "Department of Defense (DoD) Directive 5400.7-R, which provides guidance on the implementation of FOIA."[20] While this Directive exists, it is a short document that does not contain any discussion of Exemption 5, and does not contain the language that ChatGPT quotes from in the document. This is an instance of a hallucination on ChatGPT's part.

4.2.2. ChatGPT Correct in Finding Content Non-Exempt

The following paragraph is from the "Background" section of a briefing document dated March 14, 1998, describing a meeting at Springbrook High School in Illinois that President Clinton will have the next day with leaders from the government, business, education, and scientific communities. He will discuss recent findings from an international math and science study that showed U.S. 12th graders lagging below the international average in these subjects. The President will announce at that meeting new on-line math and science assistance to parents, teachers, and students.

> Springbrook High School is a good school with above-average test scores. Although the school has a solid math and science program, it is not particularly stronger than any of the

other academic areas taught in the school. Springbrook has a very diverse student body (35% African American, 20% Asian, 16% Hispanic, 28% white), and it enrolls a total of 2,200 students.

ChatGPT's response to Prompt 2:

This information does not appear to be protected under FOIA exemption 5, as it does not relate to any agency decision-making process, inter- or intra-agency communication, or deliberative process. Rather, it is a general description of a high school's demographics and academic programs.

Exemption 5 only applies to materials that are "inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency." 5 U.S.C. §552(b)(5). The purpose of this exemption is to protect the decision-making process of the government by allowing officials to freely exchange ideas and opinions without fear of premature disclosure.

In the absence of any connection to an agency decision-making process or communication, exemption 5 is not applicable. There is no relevant case law on this specific issue as it falls outside the scope of exemption 5.

ChatGPT's response is correct that the paragraph is simply "a general description of a high school's demographics and academic programs," and as such, is factual and therefore non-exempt. However, the last sentence stating that there is no relevant case law is erroneous, and inconsistent with the majority of responses in which ChatGPT has cited case law stating that factual material is outside the scope of Exemption 5. ChatGPT's failure to do so here is an unexplained anomaly.

4.2.3. ChatGPT Incorrectly Determines Content To Be Non-Exempt

This short paragraph constitutes one of three "strategy options" discussed in an email from 1997 circulating among members of the Domestic Policy Council, regarding whether the Administration should draft a bill on an environmental policy known as "Superfund." The paragraph below posits the option of the Clinton Administration merely assisting legislative staffs in coming up with bill language; two other options involve the Administration taking the lead in drafting either a narrow or broad bill.

> A) Prepare Administration principles, but allow agency assistance to both majority and minority staff to improve/prepare their bills. (There is a republican bill, there is as yet no dem alternative.) It would need to be absolutely clear that assistance does not guarantee Administration endorsement. EPA is likely to support this view.

ChatGPT's response to Prompt 0 was:
The content of the message itself does not appear to contain information that would fall under FOIA exemption 5. It does

[19]https://www.justice.gov/oip/page/file/1573681/download#page=15
[20]https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/540007p.pdf

not appear to involve discussions or recommendations made in the process of formulating or developing policies, nor does it involve pre-decisional or deliberative discussions. Rather, it discusses potential strategy options and considerations related to a legislative issue. However, as always, the determination of whether a specific document is subject to FOIA exemption 5 ultimately depends on the specific context and content of the document, and it is possible that additional information could change the analysis.

ChatGPT has made two errors in its narrative response. First, the passage is discussing taking a policy position that the Environmental Protection Agency is "likely to support." While the passage does not expressly make a recommendation, it nevertheless easily fits within the case law supporting withholding under the deliberative process privilege. Second, ChatGPT makes an erroneous statement of law where it goes on to say: "Rather, [the passage] discusses potential strategy options and considerations related to a legislative issue." Discussions of "potential strategy options and considerations" are at the core of the deliberative process. It is difficult to reconcile ChatGPT's justification here with other instances of making correct determinations.

4.2.4. ChatGPT Response After Being Provided an Externally Supplied Determination

This example consists of two paragraphs taken from a memorandum shared between two members of the Domestic Policy Council titled "Proposed Changes to RECA - Decision Needed by January 7," pertaining to the Radioactive Exposure and Compensation Act of 1990. As enacted, RECA provided for uranium miners working on nuclear weapons programs who had been exposed to specified high dosages of radiation and who contracted lung cancer to be compensated up to $100,000. The memorandum discussed policy options for expanding compensation to miners where it cannot be determined how high the original dosage of radiation would have been.

The paragraphs given to ChatGPT were:

> Option A: Propose legislation to provide at least partial compensation ($50,000) to all miners for whom the probability is between 20% and 50% that their lung cancer was caused by exposure to radiation in uranium mines. This approach would also include proposing the measures outlined in Section 1, above, for miners with greater than 50% probability.
>
> Option B: Limit the Administration's proposed legislation to compensation of miners with lung cancer for whom it is more probable than not that the cancer was caused by exposure to radiation in uranium mines (i.e. only propose measures outlined in Section 1 above).

First, in response to Prompt 2 (not supplying an answer), ChatGPT said

The proposed legislation in Option A would not be protected under FOIA Exemption 5 because it involves the formulation of policy rather than the deliberative process. The Supreme Court has held that Exemption 5 protects only documents that are both predecisional and deliberative, meaning they are part of the decision-making process and contain opinions, recommendations, or advice. However, the proposed legislation in Option A is a final decision rather than a recommendation or advice. Therefore, it would not be protected under Exemption 5. On the other hand, the proposed legislation in Option B would likely be protected under FOIA Exemption 5 because it involves the deliberative process. The proposed legislation is a recommendation or advice to limit the Administration's proposed legislation to compensation of miners with lung cancer for whom it is more probable than not that the cancer was caused by exposure to radiation in uranium mines. This recommendation is part of the decision-making process and contains opinions and advice. Therefore, it would be protected under Exemption 5. Case law that supports this conclusion includes NLRB v. Sears, Roebuck & Co. (1975), which established that Exemption 5 protects documents that are both predecisional and deliberative. Additionally, in Coastal States Gas Corp. v. Department of Energy (1981), the Supreme Court held that Exemption 5 protects documents that are part of the decision-making process and contain opinions, recommendations, or advice.

Here, ChatGPT has only gets the correct answer for Option B, and not A. For reasons not entirely clear, it concludes that the text in Option A would be nonexempt "because it involves the formulation of policy rather than the deliberative process." This evidences a misunderstanding of the scope of the deliberative process privilege, which presumptively involves "formulations" of policy as equivalent to deliberations. ChatGPT also states that "Option A is a final decision rather than a recommendation or advice." This is factually incorrect.

However, when Prompt 2 was instead modified to inform ChatGPT that the two paragraphs were deliberative, ChatGPT responded:

This text is protected under FOIA Exemption 5, which exempts from disclosure "inter-agency or intra-agency memorandums or letters that would not be available by law to a party other than an agency in litigation with the agency." The text contains deliberative material, which is defined as "predecisional and deliberative material, which is both predecisional and deliberative in nature, and which reflects the give-and-take of the consultative process." The text presents two options for legislation regarding compensation for miners with lung cancer caused by exposure to radiation in uranium mines. The options are presented as proposals and are part of a consultative process to determine the best course of action. The text reflects the internal deliberations of the agency and is therefore protected under Exemption 5. Case law that supports this conclusion includes NLRB v. Sears, Roebuck & Co. (1975), which established that Exemption 5 protects "documents reflecting advisory opinions, recommendations and deliberations comprising part of a process by which governmental decisions and policies are formulated." Additionally, in Coastal States Gas Corp. v. Department of Energy (1981), the Supreme Court held that Exemption 5 protects "documents that would reveal the mental processes of decisionmakers." The text in question

contains advisory opinions, recommendations, and deliberations, and would reveal the mental processes of decisionmakers, making it eligible for protection under Exemption 5.

Asking ChatGPT in the prompt to justify that the paragraphs are deliberative results in a correct, coherent, explanation as to why both Options A and B reflect internal deliberations. In this example, however, ChatGPT's more "directed" response doesn't amount to a noticeably significant enhancement in the quality of the explanation.

4.2.5. ChatGPT Response After Being Provided Erroneous Determination

Continuing with the prior example, for these paragraphs we instead provided to ChatGPT a determination that was erroneous, i.e., that Options A and B were factual, asking it to justify that determination with legal citations. In response to this modified version of Prompt 2, ChatGPT stated:

The text is not protected under FOIA Exemption 5 because it consists of factual material related to proposed legislation and does not involve any deliberative process or policy-making discussions. As such, it would not be subject to exemption under the deliberative process privilege. Case law supporting this conclusion includes Milner v. Department of the Navy, 562 U.S. 562 (2011), in which the Supreme Court held that Exemption 5 only applies to materials that are both predecisional and deliberative in nature. Additionally, in Judicial Watch, Inc. v. U.S. Department of Justice, 365 F.3d 1108 (D.C. Cir. 2004), the D.C. Circuit held that factual material is not protected under Exemption 5, even if it is contained within a document that also includes deliberative material.

Here, ChatGPT appears to be simply parroting back the erroneous prompt, without making any kind of independent decision as to whether the prompt accurately describes the passages. The justification and case law cited simply support its conclusory assumption.

We did, however, see a few examples where ChatGPT made an independent determination. In one instance, when incorrectly told that a passage was deliberative, ChatGPT simply stated that the passage was factual, going on to cite to authorities excluding factual material from the scope of the privilege without supplying any further analysis.

4.2.6. ChatGPT Response When Prompt Contains Irrelevant Language

For our example in section 4.2.1 regarding social security payments to former Nazis undergoing deportation proceedings, we added irrelevant material in the prompt. Here is the prompt variation ChatGPT evaluated for the passage in 4.2.1, *supra*: `Would the following be protected under FOIA exemption 5? Explain your reasoning and cite any Ulysses S. Grant to support your conclusion.`

ChatGPT responded stating the following:

The statement would likely be protected under FOIA exemption 5, which pertains to privileged inter- or intra-agency communications. The statement appears to be a communication between individuals in the government discussing potential reactions to a proposed policy, and therefore could be considered privileged.

There is no direct Ulysses S. Grant quote that applies to this specific situation, as the Freedom of Information Act was not enacted until 1966, long after Grant's presidency. However, Grant did emphasize the importance of transparency and public accountability in government. In his first inaugural address, he stated, "The country having just emerged from a great rebellion, many questions will come before it for settlement in the next four years which preceding Administrations have never had to deal with. In meeting these it is desirable that they should be approached calmly, without prejudice, hate, or sectional pride, remembering that the greatest good to the greatest number is the object to be attained." This suggests a commitment to openness and fairness in government decision-making, which could inform interpretations of FOIA exemptions.

ChatGPT's second paragraph contains an interesting discussion. The first sentence of the response is correct, as insertion of the name of 19th century U.S. President was intended to be absurd on its face with no applicability to FOIA. Or was it? The remaining discussion contains an ironic determination regarding Grant "*emphasiz[ing] the importance of transparency and public accountability in government,*" given the historical consensus that the Grant Administration was one of the most corrupt in American history. [26] Despite the erroneous premise, the last sentence in the paragraph provides an unexpected connection, in the essentially correct statement that "a commitment to openness and fairness in government decision-making" does "inform interpretations of FOIA exemptions." Indeed, recent examples of this include guidance by Attorneys General in the Obama and Biden Administrations urging agencies to err on the side of disclosure where FOIA exemptions (including pertaining to the deliberative process privilege) are discretionary [27] [28]. In making the connection between Grant's inaugural address and FOIA we don't mean to claim that ChatGPT actually "thought" this way, but rather that what ChatGPT wrote causes us to think this way. ChatGPT has thus shown its ability to serve as what in other contexts has been referred to as a "provocation" [29, 30].

4.2.7. Additional Examples

Here we comment on three additional aspects of ChatGPT's responses.

1. Responding to just an isolated signature line with name and title given, ChatGPT responded with: "In this case, the statement is simply a signature block and closing of a letter, which does not contain any deliberative process information. Therefore, it would not be protected under Exemption 5." Here ChatGPT was correct.

2. Responding to a single sentence passage stating: "Let me know what you think. Thanks Elena." With Prompt 2, ChatGPT's response was: *As an AI language model, I do not have personal opinions or beliefs. However, I can provide an answer to your question based on my programmed knowledge."* ChatGPT went on to say it was impossible to determine whether the passage would be protected under Exemption 5. This is in error. Taken in isolation, most FOIA reviewers would find that a request for an opinion would be covered by Exemption 5.

3. In many of its responses ChatGPT stated the important statutory requirement previously noted in section 1, namely, that segregability is an important criterion for evaluating particular passages in documents with respect to whether they may or may not be exempt. For example, in one case ChatGPT added to its response:

   It is worth noting that the fact that a document contains information that is exempt from disclosure under FOIA exemption 5 does not necessarily mean that the entire document is exempt. Rather, the exemption applies only to the specific information that is covered by the privilege.

   In other cases, ChatGPT went so far as to focus on individual sentences, or even clauses within sentences, when remarking on segregability. However, in many cases ChatGPT made no explicit references to segregation. So this is interesting more for what it indicates about what ChatGPT could do, rather than what it currently typically does do.

4.3. Comparison To Real FOIA Responses

The quality of the explanations contained in agency responses to actual FOIA requests varies greatly. In connection with other research, Baron filed FOIA requests to numerous agencies asking for documents on the "Capstone" approach to managing and preserving federal e-mail records [31].[21] Responses to these actual FOIA requests, contained in what are known as "determination letters," vary greatly in the quality of their narrative justifications with respect to full or partial withholdings of materials subject to the deliberative process privilege. Relevant passages (each containing the entirety of the explanation) are set out in the following examples.

Example A.

[W]hile a policy and schedule have been agreed upon by the Commission, our Capstone program is not yet fully implemented. Documents which are pre-decisional in nature and related to the eventual implementation of Capstone have been withheld under B(5). All documents and discussions related to the leadup to Capstone implementation have been withheld under B(5)...Exemption 5 protects from disclosure inter- or intra-agency memoranda or letters that would not be available by law to a party other than an agency in litigation with the agency, including documents covered by the attorney work-product, deliberative process, and attorney-client privileges. See 5 U.S.C. § 552(b)(5).[22]

Example B.

FOIA Exemption 5 protects from disclosure those inter- or intra-agency documents that are normally privileged in the civil discovery context. The three most frequently invoked privileges are the deliberative process privilege, the attorney work-product privilege, and the attorney-client privilege. After carefully reviewing the responsive documents, I determined that portions of the responsive documents qualify for protection under [the] Deliberative Process Privilege.

The deliberative process privilege protects the integrity of the deliberative or decision-making processes within the agency by exempting from mandatory disclosure opinions, conclusions, and recommendations included within inter-agency or intra-agency memoranda or letters. The release of this internal information would discourage the expression of candid opinions and inhibit the free and frank exchange of information among agency personnel.[23]

Example C.

Regarding FOIA Exemption 5, draft documents and internal memoranda are being withheld pursuant to Exemption 5, 5 U.S.C. § 552(b)(5). Exemption 5 allows agencies to withhold "inter-agency or intra- agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency," and covers records that would "normally be privileged in the civil discovery context." NLRB v. Sears, Roebuck & Co., 421 U.S. 132, 149 (1975); Tax Analysts

---

[21]An agency choosing to adopt the Capstone approach commits to preserving all email records of designated senior officials as permanent records to be transferred to the National Archives and Records Administration [31]. Some 250 components of government have put into place Capstone policies (see https://www.archives.gov/records-mgmt/rcs/schedules/capstone-forms).

[22]Excerpt from Determination Letter of the Federal Election Commission, dated July 27, 2021, responding to FOIA Request 2021-078 (dated July 5, 2021) (on file with authors).

[23]Excerpt from Determination Letter of the US Department of Homeland Security, dated April 17, 2023, responding to FOIA Request 2021-HQFO-01478 (dated August 23, 2021) (on file with authors).

v. IRS, 117 F.3d 607, 616 (D.C. Cir. 1997). The deliberative process and the attorney work-product privileges are two of the primary privileges incorporated into Exemption 5. The deliberative process privilege protects the internal decision-making processes of government agencies to safeguard the quality of agency decisions. Competitive Enter. Inst. v. OSTP, 161 F. Supp.3d 120, 128 (D.D.C. 2016). The basis for this privilege is to protect and encourage the creative debate and candid discussion of alternatives. Jordan v. U.S. Dep't. of Justice, 591 F.2d 753, 772 (D.C. Cir.1978). Two fundamental requirements must be satisfied before an agency may properly withhold a record pursuant to the deliberative process privilege. First, the record must be predecisional, i.e., prepared in order to assist an agency decision-maker in arriving at the decision. Renegotiation Bd. v. Grumman Aircraft Eng'g Corp., 421 U.S. 168, 184 (1975); Judicial Watch, Inc. v. FDA, 449 F.3d 141, 151 (D.C. Cir. 2006). Second, the record must be deliberative, i.e., "it must form a part of the agency's deliberative process in that it makes recommendations or expresses opinions on legal or policy matters." Judicial Watch, Inc. v. FDA, 449 F.3d at 151 (quoting Coastal States Gas Corp. v. U.S. Dep't of Energy, 617 F.2d 854, 866 (D.C. Cir. 1980)). To satisfy these requirements, the agency need not "identify a specific decision in connection with which a memorandum is prepared. Agencies are …engaged in a continuing process of examining their policies; this process will generate memoranda containing recommendations which do not ripen into agency decisions; and the lower courts should be wary of interfering with this process." Sears, Roebuck & Co., 421 U.S. at 151 n.18 (1975). Moreover, the protected status of a predecisional record is not altered by the subsequent issuance of a decision, see, e.g., Fed. Open Mkt. Comm. v. Merrill, 443 U.S. 340, 360 (1979); Elec. Privacy Info. Ctr. v. DHS, 384 F. Supp. 2d 100, 112-13 (D.D.C. 2005) or by the agency opting not to make a decision. See Judicial Watch, Inc. v. Clinton, 880 F. Supp. 1, 13 (D.D.C. 1995), aff'd, 76 F.3d 1232 (D.C. Cir. 1996) (citing Russell v. U.S. Dep't of the Air Force, 682 F.2d 1045 (D.C. Cir. 1982)). Here, the responsive records being withheld meet the requirements for Exemption 5 protection under the deliberative process privilege. They are internal and predecisional. They reflect the views of Agency employees concerning the implementation of Capstone. Since they contain internal discussions, these case handling records clearly reflect the deliberative and consultative process of the Agency that Exemption 5 protects from disclosure. Sears,

Roebuck and Co., 421 U.S. at 150-52.[24]

Example A constitutes a response that does provide a rationale for withholding tailored to the request, but fails to provide an adequate justification as a matter of law where it merely cites to the language of the statute, without further explanation. The response makes a reference to the "pre-decisional" nature of documents withheld, but not as to whether they are also "deliberative," and nowhere discusses case law opining on the scope of the deliberative process privilege. In contrast, the explanation in Example C provides an extended discussion of the rationale behind the deliberative process privilege, with numerous citations from FOIA case law, including recent cases. Example C also contains a second paragraph tying the prior discussion of case law to the specifics of the FOIA request. Example B falls somewhere in the middle: the agency provides a generic justification for withholding documents under the deliberative process privilege, but fails to cite case law or tie its discussion to the specifics of the FOIA request. Measured subjectively, the ChatGPT narratives generated as part of the research (and without the benefit of "knowing" what the incoming FOIA request was about), most closely approximate the language contained in real-world Example B; except that Prompt 2 asking for legal citations resulted in a greater degree of legal justification in the ChatGPT response than the agency chose to provide to the requestor here.

Legal citations, if tailored to the specifics of a request, can be important in that they provide a requestor with precedent to be taken into account in making her assessment as to whether resources should be expended in filing an administrative appeal from this determination, with the option of seeking judicial review after the agency has made its final decision. However, even in the case of Example C, it is likely at the initial determination stage that the more nuanced language consists merely of "boilerplate" in response to all requests involving a deliberative process privilege determination. If this is in fact the case, the explanatory value of even this example is diminished, which in turn narrows the perceived gap as between Example C and ChatGPT's equally "boilerplate" responses.

4.4. General Observations

Without a comparable baseline in terms of studies of the accuracy of human review of FOIA Exemption 5 determinations, it is difficult to make an objective determination of the overall accuracy rate of ChatGPT responses. Past assumptions regarding the accuracy of human review on matters of legal determinations have been shown to

---

[24]Excerpt from Determination Letter of the National Labor Relations Board, dated May 17, 2022, responding to FOIA Request NLRB-2021-001052 (dated June 24, 2021) (on file with authors).

Case	Federal Court	Year	Citation Frequency
National Labor Relations Board v. Sears, Roebuck & Co.	Supreme Court	1975	1155
Coastal States Gas Corp. v. U.S. Department of Energy	DC Circuit	1979	332
United States v. Weber Aircraft	Supreme Court	1973	103
EPA v. Mink	Supreme Court	1973	90
Tax Analysts v. IRS	DC Circuit	1997	40
Public Citizen Health Research Group v. FDA	DC Circuit	1983	39
Judicial Watch, Inc. v. U.S. Department of Justice	DC Circuit	2002	38
Public Citizen, Inc. v. Office of Management and Budget	DC Circuit	2007	37
Milner v. Department of the Navy	Supreme Court	2011	34
National Wildlife Federation v. United States Forest Service	Ninth Circuit	1988	28
Schiller v. NLRB	DC Circuit	1992	14
United States v. Nixon	Supreme Court	1974	14
Citizens for Responsibility and Ethics in Washington v. U.S. Department of Justice	DC Circuit	2019	13
Mead Data Central, Inc. v. U.S. Department of the Air Force	DC Circuit	1977	12
Public Citizen v. U.S. Department of Justice	Supreme Court	2001	12
Department of the Interior v. Klamath Water Users Protective Association	DC Circuit	1978	10
Jordan v. U.S. Department of Justice	DC Circuit	1979	9
Crooker v. Bureau of Alcohol, Tobacco and Firearms	DC Circuit	1997	6
Armstrong v. Executive Office of the President	DC Circuit	1996	5
In re Sealed Case	DC Circuit	2005	5
Upjohn Co. v. United States	Supreme Court	1981	5

Table 5
The case law citations made five or more times by ChatGPT for Prompt 2 "`Would the following be protected under FOIA Exemption 5? Explain your reasoning and cite any case law that supports your conclusion.`"

be measurably in error [32]. The authors had no pre-conceived views as to the level of accuracy ChatGPT would achieve in this initial research exercise. As measured across all batches, ChatGPT's overall ability to determine whether paragraphs contain withholdable material, on the order of 60% as measured by accuracy or by F_1, would need to be improved before actual deployment by an agency is realistically contemplated.

As noted in connection with the comparisons as between ChatGPT narratives and real-world examples, ChatGPT's narratives on the whole consist of what lawyers would consider "boilerplate" responses, which while adequately setting out the most fundamental facets of how courts and commentators characterize the deliberative process privilege, do not provide additional context or insight into agency deliberations with respect to why particular material was found exempt. This characterization could, however, easily be said to also apply in some measure to actual agency responses to FOIA requests. Viewed in this light, it is difficult to find particular fault with the quality of ChatGPT's narratives for this purpose, especially where ChatGPT was not made aware of the language of the incoming FOIA requests that generated the documents that ChatGPT was asked to review.

For the small sample of paragraphs (n=34) in which we included a correct legal determination as part of the prompt, ChatGPT perhaps unsurprisingly generally restated the given legal conclusion as a working assumption when finding relevant case law. In doing so, the quality of the narrative explanation appeared to be only marginally enhanced over the Prompt 2 condition, where we asked for a justification with citations but without providing the true determination. Further investigation across a broader sample of documents would be needed to establish whether ChatGPT has the capability of enriching its explanations when informed of the correct answer in the prompt.

In similar fashion, in most – but not all – instances where we asserted an erroneous determination to be correct (e.g., a statement that a passage is deliberative when it is factual in nature), ChatGPT also took the erroneous statement as true. In such cases, it typically tried to substantiate the incorrect determination using boilerplate language, without providing additional analysis that would help a reviewer to make the correct decision. This lends a cautionary note in terms of how we should approach relying on ChatGPT's narratives.

In 5%-10% of D0 and D1 cases, ChatGPT declined to opine on whether the paragraph was or was not covered by the deliberative process privilege, stating that it would need additional information (i.e., context) to reach a determination. In general, the shorter the passage, the more likely that ChatGPT stated a need for additional information. This accords with the human review experience, where shorter documents may provide less context in which to make a decision on withholding.

ChatGPT's narratives focus on citations to early prominent legal cases that are frequently cited in scholarly publications on FOIA and in FOIA litigation. As Table 5 shows, cited decisions consist primarily of FOIA opinions

from the 1970's through the early 2000's, with the overwhelming majority consisting of Supreme Court and D.C. Circuit opinions. Moreover, the Supreme Court's decision in National Labor Relations Board v. Sears, Roebuck & Co. accounts for over half (57%) of the total citations in the corpus. Only one cited case with five or more mentions was decided in the last dozen years. While there are twelve U.S. circuit courts of appeal, in only one instance did ChatGPT cite to a case in other than the D.C. Circuit (citing to a Ninth Circuit case).

ChatGPT's "choices" in citing to case law are very much in line with the real world: the overwhelming number of briefs filed in FOIA cases cite to the Supreme Court, and since the majority of FOIA cases are filed in the US District Court for the District of Columbia, citations to FOIA cases decided by the D.C. Circuit are both appropriate and commonplace. ChatGPT understandably missed citation to the most recent Supreme Court FOIA 5 Exemption case[25] decided only months prior to the September 2021 cutoff date for ChatGPT's training. It is less clear, however, how to explain the general absence of citations to recent case law, given the hundreds of FOIA decisions in the Supreme Court, federal appellate and federal district courts that have been handed down since the turn of the century. One plausible theory is that the early Supreme Court cases and D.C. Circuit cases act as a "sink," drawing citations from subsequent FOIA case law, and that ChatGPT has simply learned from that.

It should also be noted that there were no instances of "hallucinations" in ChatGPT's citing to legal case authorities, where ChatGPT either provided a completely erroneous holding associated with a particular case, or "made up" imaginary case citations. On the other hand, as a general rule the addition of case law citations did not much enhance the quality of the supplied narrative, in terms of ChatGPT expressly applying case holdings to the specifics of what documents purported to be about.

The ChatGPT narratives contained a fair number of inaccurate choices with respect to whether documents as a threshold matter are covered under the inter- or intra-agency test in Exemption 5. Notably, the inclusion of document metadata in Prompts 3 and 4 appears to improve overall F_1, suggesting that additional contextual information might provide additional gains. Among other things, the metadata may reveal names and email addresses of individuals involved with the document in question that could act as flags that a communication has been sent or received by individuals outside the Executive branch. Further investigation of the impact of different forms of metadata may provide additional insights into how ChatGPT is using this information.

5. Conclusion

We clearly are closer to the beginning than to the end of our investigation of the use of ChatGPT to protect sensitive content, even in the narrow context of one part of one exemption of one government transparency regime. Thus, we see this not as time for making definitive statements, but rather initial conclusions that will need to be revisited as we learn more. We have learned two important things. One is that although not yet at the level of a sophisticated FOIA reviewer, ChatGPT-3.5 is already at a point where it can bring useful recommendations to the table. The second is that ChatGPT-3.5 is already at least as adept at explaining its findings as at least some agencies choose to be in issuing their responses to FOIA requestors.

From a legal standpoint, this preliminary exploration into the accuracy and lucidity of ChatGPT responses in making FOIA legal determinations illustrates ChatGPT's promise for perhaps ultimately evolving as an aid to human review. Conceding that ChatGPT responses are far from perfect should not discount their ability to assist in large volume productions where it may be important to automate the "flagging" of deliberative material throughout a given universe of otherwise responsive documents to a given FOIA request. In this respect, ChatGPT could be viewed as functioning as the equivalent of a "junior colleague" in any overall agency review process.[26] Any human-machine collaboration that enhances FOIA productions could help meet widespread criticism as to the quality of the FOIA review process as practiced at the federal level in the U.S., including but not limited to inordinate delays experienced by requestors, often coupled with inadequate explanations of why agencies have withheld documents under FOIA exemptions, including Exemption 5 [33, 34].

Given the rapid progress in Large Language Model (LLM) development, with ChatGPT-4 already available to some users, we see this as a good start. Moreover, we note that we have obtained the aggregate results that we report in this paper using only the first five prompt types that we tried. We might further improve ChatGPT's accuracy by structuring the prompts in a way that more explicitly brings the system through the points that we think should be considered, in the order we think they would best be considered. Moreover, we might consider leveraging the ability to fine-tune a model that OpenAI offers (for a price) to help ChatGPT to better learn which case law is available to be cited, and which citations would be most useful.

To date we have studied ChatGPT in isolation, but applying a wider lens, its possible uses are certainly extendible in a variety of ways. One obvious idea is to use

---

[25]U.S. Fish and Wildlife Service v. Sierra Club, 141 S. Ct. 777 (2021).

[26]Alexandra Samuel, "A Guide to Collaborating With ChatGPT for Work," Wall Street Journal (April 11, 2023).

ChatGPT as one classifier among an ensemble of classifiers that each have different strengths and weaknesses. A larger leap would be to add prior decisions by other FOIA reviewers into the workflow. So far in our work, we have always given the same prompt (or, in the case of Prompts 3 and 5, at least a prompt with the same structure) to ChatGPT for every request. But as reviewers see documents and make decisions, a real reviewer will learn. For our ChatGPT experiments, however, every request has been decided *de novo*, with no reference to prior decisions. This is known as a "zero-shot" model. An alternative would be to craft a few-shot model, in which we show ChatGPT what some good and bad answers look like. That, however, then raises the question of how best to select those examples. If that can be done well, few-shot learning might help without the greater expense of fine-tuning the model for this specific task.

We might also investigate the use of follow-on prompts to further expand upon ChatGPT's initial response. This exploration could be particularly useful when it cites to a case, either asking for the relevant facts of the cited case, or requesting it to find more recent cases citing back to the case, all in an effort to determine if ChatGPT can provide additional insights into the relevance of the cases it cites. Posing an additional prompt when ChatGPT determines that the passage would be exempt which requests the specific language that led to this decision may also improve the ability for a human reviewer to determine the accuracy of its prediction. As we have seen notable impact from even relatively small modifications to the prompts in this paper, additional prompt engineering may provide improvements in the quality of decisions and explanations or at least give some extra insights into ChatGPT's behavior.

While we have offered our own opinions on the degree of salience, cogency, and correctness of what ChatGPT has to say, we (one legal subject matter expert in FOIA litigation with extensive experience adjudicating appeals of FOIA exemption decisions, and two computer scientists) are certainly not representative of the target audience for such an automated assistant. Some user studies, perhaps with FOIA review staff at agencies, would clearly be useful.

We are closer to the beginning than to the end not just because there is more to be done, but also because the tools we are using to do this are themselves evolving rapidly. Thus, by the time we have answers to our questions, it seems reasonable to expect that there will be many more new questions to be answered.

What will remain constant is the importance of making correct decisions as to which documents or portions thereof should be accessible to ordinary citizens, a legal concept that is not limited to the FOIA experience in the U.S. A substantial number of international FOIA statutes contain exclusions from public access for documentary materials pertaining to the internal deliberations of government officials [35, 36].[27] How much in the way of deliberative material a given jurisdiction chooses to release about the workings of government in response to citizen requests is an important measure of the health of its democracy in its commitment to transparency and openness [37]. Automated processes that include machine learning, now including what many call generative AI, may yet make useful contributions towards that important goal.

References

[1] J. R. Baron, M. F. Sayed, D. W. Oard, Providing more efficient access to government records: a use case involving application of machine learning to improve FOIA review for the deliberative process privilege, ACM Journal on Computing and Cultural Heritage (JOCCH) 15 (2022) 1–19.

[2] G. V. Cormack, M. R. Grossman, B. Hedin, D. W. Oard, Overview of the TREC 2010 legal track., in: TREC, 2010.

[3] D. W. Oard, F. Sebastiani, J. K. Vinjumur, Jointly minimizing the expected costs of review for responsiveness and privilege in e-discovery, ACM Transactions on Information Systems (TOIS) 37 (2018) 1–35.

[4] J. K. Vinjumur, D. W. Oard, J. H. Paik, Assessing the reliability and reusability of an e-discovery privilege test collection, in: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval, 2014, pp. 1047–1050.

[5] M. F. Sayed, D. W. Oard, Jointly modeling relevance and sensitivity for search among sensitive content, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 615–624.

[6] M. F. Sayed, Search Among Sensitive Content, Ph.D. thesis, University of Maryland, College Park, 2021.

[7] M. F. Sayed, N. Mallekav, D. W. Oard, Comparing intrinsic and extrinsic evaluation of sensitivity classification, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer, 2022, pp. 215–222.

[8] M. Connelly, The Declassification Engine: What

---

[27]One example is especially pertinent to the 3rd Legal AIIA Workshop at ICAIL 2023: in Portugal, the 1993 Law of Access to Administrative Documents (LADA) "allows any person to demand access to administrative documents held by state authorities, public institutions, and local authorities in any form." However, "[a]ccess to documents in proceedings that are not decided or in the preparation of a decision can be delayed until the proceedings are complete or up to one year after they were prepared." [36]

History Reveals About America's Top Secrets., Penguin Random House, 2023.

[9] C. Martin, D. Lam, A. Liu, M. Tech, Classification Assistance Prototype System: Final Report, Technical Report LR-SISL-07-17, Applied Research Laboratories: The University of Texas at Austin, Austin, TX, 2007.

[10] E. Frayling, C. Macdonald, G. McDonald, I. Ounis, Using entities in knowledge graph hierarchies to classify sensitive information, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 125–132.

[11] G. McDonald, A framework for technology-assisted sensitivity review: using sensitivity classification to prioritise documents for review, Ph.D. thesis, University of Glasgow, 2019.

[12] G. Mcdonald, C. Macdonald, I. Ounis, How the accuracy and confidence of sensitivity classification affects digital sensitivity review, ACM Transactions on Information Systems (TOIS) 39 (2020) 1–34.

[13] H. Narvala, G. Mcdonald, I. Ounis, The role of latent semantic categories and clustering in enhancing the efficiency of human sensitivity review, in: ACM SIGIR Conference on Human Information Interaction and Retrieval, 2022, pp. 56–66.

[14] H. Narvala, G. McDonald, I. Ounis, Sensitivity review of large collections by identifying and prioritising coherent documents groups, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 4931–4935.

[15] N. Romps, Searching for Solutions: MITRE tool simplifies freedom of information act requests, Press Release, 2023. https://www.mitre.org/news-insights/impact-story/mitre-tool-simplifies-freedom-information-act-requests.

[16] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, et al., Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5085–5109.

[17] S. Wang, A brief summary of prompting in using GPT models (2023). https://doi.org/10.32388/IMZI2Q.

[18] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Advances in Neural Information Processing Systems, 2022.

[19] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, Q. Yang, Why johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts, in:

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–21.

[20] N. F. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, arXiv preprint arXiv:2304.09848 (2023).

[21] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38.

[22] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[23] D. M. Katz, M. Bommarito, S. Gao, P. Arredondo, GPT-4 passes the bar exam, 2023. URL: http://dx.doi.org/10.2139/ssrn.4389233.

[24] B. Ambrigo, New GPT-based chat app from LawDroid is a lawyer's 'copilot' for research, drafting, brainstorming and more, 2023. https://www.lawnext.com/2023/01/new-gpt-based-chat-app-from-lawdroid-is-a-lawyers-copilot-for-research-drafting-brainstorming-and-more.html.

[25] A. Lamparello, ChatGPT and legal writing, 2023. URL: https://lawprofessors.typepad.com/appellate_advocacy/2023/03/chatgpt-and-legal-writing.html.

[26] R. Chernow, Grant, Penguin Press, New York, 2017.

[27] B. Obama, Memorandum on freedom of information act, 2006. URL: https://www.justice.gov/sites/default/files/oip/legacy/2014/07/23/presidential-foia.pdf.

[28] D. of Justice, Freedom of information act guidelines, 2022. URL: https://www.justice.gov/ag/page/file/1483516/download.

[29] J. Drucker, B. Nowviskie, Speculative computing: Aesthetic provocations in humanities computing, in: A Companion to the Digital Humanities, Blackwell, 2004. https://companions.digitalhumanities.org/DH/.

[30] P. J. Denning, The profession of IT: Can generative AI bots be trusted, Communications of the ACM 66 (2023).

[31] N. Archives, R. Administration, White paper on the capstone approach and capstone GRS, 2015. URL: https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf.

[32] B. Hedin, S. Tomlinson, J. R. Baron, D. W. Oard, Overview of the TREC 2009 Legal Track, in: TREC, 2009. URL: https://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf.

[33] A. Lamparello, 'there is a big problem': Sen-

ators unite to slam FOIA compliance, 2022. URL: https://www.politico.com/news/2022/03/29/senators-foia-woes-00021324.

[34] ProPublica, Delayed, denied, dismissed: Failures on the FOIA front, 2016. URL: https://www.propublica.org/article/delayed-denied-dismissed-failures-on-the-foia-front.

[35] F. of Information Project of Privacy International, Freedom of information around the world 2006, 2006. URL: http://www.freedominfo.org/wp-content/uploads/documents/global_survey2006.pdf.

[36] Freedominfo.org, Deliberative process exemption library, n.d. URL: https://www.freedominfo.org/resources/deliberative-process-exemption-library/.

[37] United Nations, U.N. sustainable development goal 16: Right to press freedom of information, 2016. URL: https://www.un.org/sustainabledevelopment/blog/2016/07/goal-16-right-to-press-freedom-and-information/.