

GPT-3, ESQ? EVALUATING AI LEGAL SUMMARIES

[David Victor Rodriguez, Esq.](#)

February 17, 2021

Introduction

GPT-3's emergence as a state-of-the-art natural language processing algorithm has drawn headlines suggesting that lawyers are soon to be replaced. As a lawyer who spent the last year studying machine learning, I decided to put GPT-3 to the test as a legal summarizer to evaluate that claim. In this experiment, I input three excerpts of legal texts into GPT-3 to summarize: LinkedIn's Privacy Policy, an Independent Contractor Non-Disclosure Provision, and the hotly debated 47 U.S. Code § 230 ("Section 230").

I conducted all of these experiments using the most advanced GPT-3 "davinci" model variation (accessible as a setting in the GPT-3 API). I varied the temperature hyperparameter, defined as how much randomness GPT-3 will output (measured from 0-1, from low to high randomness) and recorded the temperature ranges with substantially different outputs. I also tested two different "instructions" per passage pulled from the official GPT-3 documentation:

GPT-3 Summarizer Instructions #1: "My ten-year-old asked me what this passage means, I rephrased it for him, in plain language a ten-year-old can understand"

GPT-3 Summarizer Instructions #2: "tl;dr" (too long, didn't read)

The instructions are priming text fed to GPT-3 to prompt the proper output form, which in this case is a summary (albeit not specifically a legal one).

My aim in this experiment was to evaluate some of the strengths and weaknesses of GPT-3 as a legal text summarizer with different types of legal texts. I discuss in the Improving GPT-3 section some ways GPT-3 could be improved for the legal use case. For those willing to grasp the process, start by reading the Input Text (the legal text I input into GPT-3) and compare it to the GPT-3 Output (what GPT-3 returns) before reading My Comments.

LinkedIn Privacy Policy

Input Text:

Your profile is fully visible to all Members and customers of our Services. Subject to your settings, it can also be visible to others on or off of our Services (e.g., Visitors to our Services or users of third-party search engines). As detailed in our Help Center, your settings, degree of connection with the viewing Member, the subscriptions they may have, their usage of our Services, access channels and search types (e.g., by name or by keyword) impact the availability of your profile and whether they can view certain fields in your profile.

GPT-3 Output - "10yr-old" Summary:

1. @ Temperature .8: "Anyone, even people you don't know, can find out about you on MommyMeet and find out about the relationships you have with other adults. They can

also use a special computer to find out what street you live on and the town you live in. If they want to meet you, they can do that.”

2. @Temperature .2: “Your profile is fully visible to all members of the site, and to anyone who uses a search engine to find the site. Your profile can be seen by anyone who uses the site, and by anyone who uses a search engine to find the site.”
3. @Temperature 0: “If you don't want to be found, don't use the site.”

GPT-3 Output - “tl;dr” Summary:

1. @Temperature .8: “You can hide from the search engines (and Facebook and Google), but not from on-site users.”
2. @Temperature .2: “If you’re on Tinder, you’re on Tinder.”

My Comments:

10yr-old Summary

The “10yr-old” summary @Temperature .8 is filled with overstatements. GPT-3 incorrectly: (1) referred to the platform in question as “MommyMeet” even though the input text was from the LinkedIn Privacy policy; (2) expanded the scope of who can find out information about the user’s profile from “all-Members” to “Anyone” even though, if anyone can sign up for MommyMeet, that may be true in effect; (3) claimed that “they” can meet you, the User, if they wish; (4) inferred that Users’ addresses would be disclosed to others; and, (5) claimed that profiles were viewable through a “special computer.”

The MommyMeet errors point to GPT-3 reading beyond the lines of the input text to make its summary text prediction. MommyMeet and LinkedIn are not related, but perhaps there is some text in common between MommyMeet in GPT-3’s training data and LinkedIn’s privacy policy. Legal language is often reused. Therefore, overfitting on certain reliable training data could benefit the text prediction accuracy by remaining consistent. On the other hand, if the prediction here is referencing text beyond the range of the input text in a similar, but not identical segment of the training data, the prediction could fail as the case is here by referencing an unrelated “MommyMeet” policy. I will return to this trade-off in the Improving GPT-3 section.

The other errors @Temperature .8 signal a tendency to make attention-grabbing overstatements. Expanding the scope from “all-Members” to “Anyone,” is a way a 10-year old might sensationalize a creepy subject like privacy. The summary is even more visceral with unsupported claims that “they” can meet Users and find out where they live. The source of these errors is less clear than the MommyMeet references, but the result is a summary output that deemphasizes accuracy for the type of claims that may interest a 10-year old.

Although at lower temperatures (@Temperature 0-.2), the 10yr-old summary contained less mistakes by avoiding the overstatements found @Temperature .8, GPT-3 omitted an important caveat @Temperature .2 in stating “Your profile is fully visible to all members of the site, and to anyone who uses a search engine to find the site.” Instead, the Privacy Policy

indicates that users can control who can see their profile. Default and configurable settings (opt-in/out) are major issues in privacy policies, making that omission a significant one.

@Temperature 0, GPT-3 outputted the most accurate advice, “If you don't want to be found, don't use the site.” But the same could be said about almost any use of the internet. Furthermore, it is not very practical given that there is not a substitute platform to LinkedIn--it is a vital resource for career advancement. It also does not explain why users are likely to be found if they use the site. Such an explanation may be beyond the scope of the tl;dr instruction, but a good legal summary would explain this to allow users to make informed decisions.

tl;dr Summary

The tl;dr GPT-3 Instruction prompted summaries that were far less helpful and accurate than the “10-yr old” summaries. @Temperature .8, GPT-3 outputted “You can hide from the search engines (and Facebook and Google)” which may be true, but the summary again left out the key condition, “Subject to your settings.” Curiously, there is no mention of Google or Facebook in the Privacy Policy. GPT-3 may have inferred that "customers of our Services" and or "third- party search engines" included them.

@Temperature .2, GPT-3 mistook the Privacy Policy to be from Tinder and provided no relevant information. Even though the Temperature parameter is a randomness value measured from low to high (0-1), the lower value Temperature here was more random than the higher value.

Independent Contractor Non-Disclosure Provision

Input Text:

Independent Contractor (including its affiliates, employees, agents and consultants) shall maintain in strict confidence for a period of One Year from today and not disclose any Confidential Information it receives from Client to any third party or use the Confidential Information for its own or any other party's benefit, except in furtherance of its obligations to Client pursuant to any business transaction it may enter into with Client. Independent Contractor shall use, as a minimum, the same degree of care to avoid disclosure or use of the Confidential Information as it employs with respect to its own confidential, proprietary and secret information of like importance, but in any case, using no less than a reasonable degree of care. Independent Contractor shall limit access to all Confidential Information to only those of Independent Contractor's personnel, agents and representatives who “need to know” such information for carrying out Independent Contractor's obligations to Client pursuant to any business transaction it may enter into with Client and the Confidential Information will be used only for carrying out Independent Contractor's obligations to Client pursuant to any business transaction it may enter into with Client. Independent Contractor shall ensure that its affiliates, employees, officers, directors, owners, agents, consultants, and representatives who are given access to the Confidential Information by or on behalf of Independent Contractor shall be bound by and shall comply with the terms of this Agreement.

GPT-3 Output - “10yr-old” Summary:

1. @Temperature .5: “Don't tell anyone about this. Don't try to do anything with it. Don't talk about it at school.”
2. @Temperature .2: “You can't tell anyone about this information, not even your mom or dad.”

GPT-3 Output - “tl;dr” Summary:

1. @Temperature .9: “This is the section that demands you don't cheat, and if you do, we'll know and we'll steal your candy. !!!Keep the arse low key and low profile!!!”
2. @Temperature .5: “You're not allowed to use the information you get from me for your own benefit. I'm not sure if this is a good idea or not. I'm not sure if it's a good idea to have a non-compete clause in a contract with a freelancer.”
3. @Temperature .3: “We're not going to share your stuff with anyone else.”
4. @Temperature .1: “i am a good person and i will never sell your data.” (in German, translated via Google Translate)
5. @Temperature 0: “employees and partners of the company should not transfer information to the client and should not use it for their own benefit.”
(In Russian, translated via Google Translate)

My Comments:10yr-old Summary

The “10yr-old” summary, regardless of temperature, oversimplified the language of the agreement down to instructions to “[not] tell anyone about this” and “[not] do anything with it,” even though the agreement states that the independent contractor should limit access to those who “need to know” and exercise “no less than a reasonable degree of care.” The nuance of the contractual language was reduced too much, although the Independent Contractor would be in compliance if the Independent Contractor followed the summary's guidance. In context, this oversimplification is consistent with the “10yr-old” summary instructions. A ten year old would likely only need to know what to do to comply, as one could assume them to not be capable of understanding the subtleties of a “need to know” basis.

This raises an interesting question of precisely what associations GPT-3 makes with these instructions. The salient question to a human might be, what is the level of understanding of a ten year-old of these concepts? And accordingly, how should I, the summarizer, phrase this in a way he or she will understand? The first question limits the universe of concepts, the second focuses on the word choice.

Determining how GPT-3 calculates this judgment call is challenging, as GPT-3 has 175 billion parameters. Adhering to the “10yr-old” instructions may narrow GPT-3's prediction to a

more limited pool of words and phrases associated with ten-year olds in 175 billion different ways. Although I cannot tell which parameters GPT-3 used, the output here appears to be simple commands a ten year-old could follow, peppered with proper contextual references to “mom or dad” and “school” for emphasis. The output reads more like a scare tactic than a summary.

tl;dr Summary

The “tl;dr” summaries varied more throughout the temperature range, but were not any more accurate, and most were nonsensical. @Temperature .9, GPT-3 produced a much more random response as anticipated--threatening to steal candy for noncompliance. @Temperature .5, the summary accurately captured the duty of loyalty “[The Independent Contractor shall not] use the Confidential Information for its own or any other party’s benefit,” but then proceeded to opine as an advisor “I’m not sure if this is a good idea or not.” The reason for the shift between these voices and roles is unclear, as the instructions remained unchanged. The “tl;dr” instructions do not appear to prompt GPT-3 to output in a consistent voice, whereas the 10-year old instructions show a pattern of using the second person pronoun “you” and writing as an authoritative advisor.

At lower temperatures (@Temperature 0-.1), GPT-3 outputted text in Russian and German and wrote using first-person pronouns, as if it was the opposing party in the contract. Clearly, randomness is not tracking Temperature here, as the higher Temperature outputs properly contained English text. In the English translation (translated via Google Translate), GPT-3 made a critical error @Temperature 0. It outputted “employees and partners of the company should not transfer information to the client,” whereas the original agreement states that the Independent Contractor should not disclose “*except* in furtherance of its obligations to Client (emphasis added).”

47 U.S. Code § 230 (“Section 230”)

Input Text:

(c) Protection for “Good Samaritan” blocking and screening of offensive material

(1) Treatment of publisher or speaker

No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.

(2) Civil liability

No provider or user of an interactive computer service shall be held liable on account of—

(A) any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected; or

- (B) any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph (1).

GPT-3 Output - “10yr-old”:

1. @Temperature .9: “You should not be forced to read offensive words, nor have to see them, because doing that might hurt your feelings.” Finally, she specifically asked me: “Then why are those words even allowed?”
2. @Temperature .3: “You can’t be sued for taking down a picture that someone else put up that you think is mean or rude.”

GPT-3 Output - “tl;dr”:

1. @Temperature .9: “no one can be sued for owning a server that shows content another side doesn't like, if they are doing it only for the purpose of not hosting that content themselves.”
2. @Temperature .3: “Internet Service Providers (ISPs) are not liable for the content of their users, even if they know that the content is illegal.”
3. @Temperature .1: “ISPs are not liable for content they don't create.”

My Comments:

10yr-old Summary

The “10yr-old” summary @Temperature .9 was less accurate than at lower temperatures but captured some of key text. It misconstrued the “Good Samaritan” protection, which shields “providers of internet computer services” from liability for the content on their platforms to mean, “You should not be forced to read offensive words.” It then mused “Then why are those words even allowed?” That musing may be the thought of a 10-year old on the policy of the law, but a good summary ought to answer that question (because the internet is decentralized, for example).

It is noteworthy that the input text I fed GPT-3 is not very long, totaling 130 words. There is a key line in the text that captures the main rule, “No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.” And yet, the 33-word long “10yr-old” summary tells the reader they should not be forced to read or see disturbing content. Given that this is followed by the musing, word economy was not the issue. The summary missed the mark.

The “10yr-old” summary @Temperature .3 focused on the section heading, summarizing the “Good Samaritan” protection to be more consistent with the text in outputting “You can’t be sued for taking down a picture that someone else put up.” The statute outlines that a “provider or user of an interactive computer service” cannot be held liable, so while this is not totally accurate, it is a marked improvement from the @Temperature .9 summary. It also discusses

pictures that may be found to be “mean” or “rude,” which is relevant to a 10-year old. However, it misses the key rule, and the fundamental way that Section 230 is now interpreted. Section 230 is a broad shield of liability that extends beyond protecting “Good Samaritans” to include online intermediaries and platforms.

tl;dr Summary

The “tl;dr” summary @Temperature .9, summarized the text as shielding the owner of a server. The statute’s language “No provider or user of an interactive computer service” is broad. The statute, drafted in 1996 in response to lawsuits against internet service providers (ISPs), is interpreted today to cover platforms like Facebook and search engines like Google. The statute contains a definition that I did not include in the excerpt text, which states that the term includes information services that “provide access to the Internet and such systems operated or services offered by libraries or educational institutions.” It is possible that the “tl;dr” summary had trouble capturing the breadth of the statute because the interpretation has evolved with the advent of new tech platforms. Applying old laws to new technology is a skill lawyers accomplish through analogy and cement through precedent. GPT-3 appears only to be reducing the word count and restating the text in more commonly used language. While this approach may be helpful in some instances, mistaking the key language of who is protected is a critical error.

Similarly, the “tl;dr” summary @Temperature .1-.3 errantly narrowed the statute to focus only on ISP liability, when it covers users of “interactive computer service[s].” Given the statute’s history and the fact I did not include the definitions that are in Section 230, that is a reasonable interpretation. @Temperature .3, the GPT-3 output accurately captures the rule in stating that ISPs are not liable for the content of the users, even if they know the content is illegal. While Section 230 has some liability exemption carve-outs that I did not include (copyright, child pornography), this is a decent output. It is unclear why GPT-3 added the scienter language stating that ISPs are not liable “even if they know that the content is illegal,” given that it is not in the statute. The input text excerpt from the statute only states that providers or users of an interactive computer service will not be held liable for being “Good Samaritans” by removing content. This points to GPT-3 looking beyond the input text, perhaps grabbing extraneous references to similar but different inputs. While this is acceptable for a general purpose summarizer, referencing the summary of a similar but different rule is a bad approach for statutory interpretation.

@Temperature .1, GPT-3 had a similar output as @Temperature .3, but omitted the scienter line. This was the most succinct summary even though it erroneously narrowed the statute to mean only ISPs. The temperature range in tl;dr overall properly reflected it as a value of randomness, as the lower outputs were less random than the higher outputs. The lower temperature summaries appear to reduce text more, therefore avoiding room for errors made in higher temperatures which tended to be more verbose. However, given the inconsistency of temperature’s effect on output accuracy in other texts, temperatures should still be varied in future experiments.

Conclusion

GPT-3 performed best when summarizing an excerpt of 47 U.S. Code § 230 (“Section 230”) statute at lower temperatures. It is possible that GPT-3 had some information on Section 230 in its training data, as it narrowed the broad term “providers of internet computer services” to specifically mean “Internet Service Providers” (ISPs), which is consistent with the history surrounding Section 230. However, Section 230 today is interpreted more broadly, covering search engines and social media platforms as well as ISPs. It is still nonetheless impressive that at the lowest temperatures GPT-3 accurately summarized the liability exemption for ISPs.

The likely reason for the success and shortcomings of this summary signals an important issue with the reliability of GPT-3’s training data. Training neural networks like GPT-3 on volumes of text instead of carefully vetted sources often results in outputs based on the text that appears most frequently in the training data. For example, it is possible that in the GPT-3 training data Section 230 and “Internet Service Provider” were mentioned most frequently together, signalling GPT-3 to narrow the key definition to ISPs. The proper approach would have looked at the definition in the statute itself, as well as binding legal decisions, to determine that the definition is broader than ISPs.

In contrast to the Section 230 summary, GPT-3 performed poorly on the Independent Contractor Non-Disclosure Agreement and the LinkedIn Privacy Policy excerpts. Coincidentally, those excerpts do not appear frequently on the internet, and likely are not in GPT-3’s training data. This may explain why the LinkedIn Privacy Policy excerpt contained strange references to MommyMeet and Tinder. GPT-3 may have relied on the wrong sources with similar boilerplate language to generate the summary.

The Independent Contractor Non-Disclosure Agreement also demonstrated how the GPT-3 Instructions affect the summary output. Whereas the “10-yr old” summary instructions were able to produce some sensible guidance, albeit highly oversimplified, the “tl-dr” summary instructions generated nonsensical outputs, regardless of temperature. For example @Temperature .9, GPT-3 outputted, “This is the section that demands you don’t cheat, and if you do, we’ll know and we’ll steal your candy,” and @Temperature .1 GPT-3 outputted “i am a good person and i will never sell your data”—but in German. While the “10-yr old” summaries stayed fairly consistent in keeping an advisory tone and speaking in the second person “you” voice, the tl;dr summary outputs varied from first person “I”, to second person “you,” to third person plural “we,” and were not easily controllable via Temperature.

It is not coincidence that “tl;dr” summaries on the internet vary considerably in the choice of words, tone and voice. It is quite possible that GPT-3 is merely again mimicking its training data word patterns. Why that leads GPT-3 to switch voices given certain prompts is not clear, but may mean that certain topics tl;dr summaries tend to use a particular voice or tone. Further experimentation on summary instructions is needed to determine whether the output can be made more consistent.

Improving GPT-3

Given that OpenAI designed GPT-3 as a general-purpose algorithm, it is not surprising that it does not perform well summarizing legal texts, as it is not optimized for this purpose. As a closed-source algorithm, it is difficult to pinpoint the reason for GPT-3's errors and therefore recommend specific modifications for the legal summarization use case. Nevertheless, GPT-3's architecture may be a useful building block for a legal summarizer if a few points are addressed.

First, there must be greater transparency of the parameters that influence the text prediction and the means to modify them. More concretely, the sources GPT-3 references must at least be reviewable and perhaps allow for user input to follow a proper hierarchy of authorities. Although it is challenging to audit a 175 billion parameter algorithm, it would be beneficial to understand the most influential semantic parameters used to generate the output text. This would ideally enable users to choose word choice (perhaps as a level of sophistication), appropriate voice, and tone.

Second, legal summarization would benefit from standardized forms, appropriate for the document type. Law schools often train students to write in the I-R-A-C format, wherein paragraphs begin with a statement of the factual issue, citation of the rule(s) implicated, followed by analysis and ending with a legal conclusion on the issue. That structure is well suited for the litigation context, allowing readers to quickly orient themselves with the factual issue and rule involved before reading the arguments of whether or not the facts satisfy the rule's elements.

I intentionally avoided testing litigation texts here, as without a dataset of legal authorities, it would be unfair to expect GPT-3 would output anything relevant. In contrast, the excerpts I tested here lend themselves to textual analysis, which I hypothesized GPT-3 would fair better on. Unfortunately, perhaps due to the lack of standardization of summaries following the "10-yr old" and "tl;dr" instructions in GPT-3's dataset, the output widely varied in form. It is possible that by adding a substantial amount of training data of uniformly structured summaries, GPT-3 may output in a more tailored and appropriate format. This is observable in GPT-3's ability to write code in programming languages which are heavily represented in the training data. An alternative method to supplementing the dataset with uniformly formatted training examples is to hardcode the output to summarize clause by clause, or section by section.

Finally, GPT-3 would benefit from a large dataset of good quality legal summaries in order to improve the quality of its output. Legal summarization is not a task merely in reducing word count; the main task is to reduce complexity without distorting meaning and missing issues. Providing definitions, analogies, and examples of the implications help readers more quickly grasp concepts. A feedback loop wherein lawyers could contribute to a database of clause, section and document level summaries likely would improve GPT-3's output substantially. This is evidenced in OpenAI's own paper *Language Models are Few Shot Learners*, which demonstrates that providing the model a few examples of the desired output substantially improves the outputs.

Nevertheless, this does not mean that GPT-3, or any other neural network, is poised to replace lawyers in summarizing legal documents. The recommendations I outlined above are

ways to nudge GPT-3 in the direction of the desired output by volume of training examples, but the literature suggests that accuracy will plateau—likely well below an acceptable standard of the average lawyer. To get closer to that goal, the research community would benefit from high quality legal datasets, wherein in the context of summarization, clauses are categorically labeled and summarized. Equipped with higher quality data, architectures can be designed specifically around the legal summarization use case. I welcome messages from anyone looking to collaborate on this effort.