**MINISTÉRIO PÚBLICO FEDERAL**
**PROCURADORIA-GERAL DA REPÚBLICA**

Ofício nº 488/2023 - ASSEXP/PGR

Brasília, *data da assinatura digital*.

A Sua Excelência o Senhor
Conselheiro **RODRIGO BADARO ALMEIDA DE CASTRO**
Conselho Nacional do Ministério Público
Protocolo ELO / Tel: (61) 3366-9100
SAFS, Quadra 2, Lote 3 - Edifício Adail Belmonte
70070-600 - Brasília - DF

Assunto: **Ofício nº 5/2023/GAB/RB, de 12/05/2023. Pedido de Providências nº 1.00085/2023-10.**

Senhor Conselheiro,

Cumprimentando-o cordialmente, ao tempo em que acuso o recebimento do ofício em epígrafe, encaminho a Vossa Excelência o Relatório nº 35/2023 (PGR-00180414/2023), oriundo da Secretaria de Tecnologia da Informação e Comunicação/SG, que tem por escopo a análise do *Chat GPT* e outras tecnologias generativas, em especial, a possibilidade de uso de tais tecnologias no âmbito do Ministério Público Federal, respectivas vantagens para a Instituição, bem como riscos associados e formas de mitigá-los.

Na oportunidade, renovo a Vossa Excelência manifestação de apreço e consideração.

***Augusto Aras***
Procurador-Geral da República
*Assinado digitalmente*

**MINISTÉRIO PÚBLICO FEDERAL**

**PROCURADORIA GERAL DA REPÚBLICA**

**SECRETARIA GERAL**

**SECRETARIA DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO/SG**

RELATÓRIO N° 35/2023/SECRETARIA DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO/SG

| | |
|---|---|
| REFERÊNCIA | OFÍCIO 5/2023 - PGR-00174959/2023 |
| SOLICITANTE | CNMP - Conselheiro Rodrigo Badaró |
| EMENTA | Análise do Chat GPT e Tecnologias Generativas - Possibilidade de uso no âmbito do Ministério Público Federal - Vantagens - Riscos Associados e formas de mitigá-los. |

## 1 INTRODUÇÃO

1. O presente relatório tem por escopo a análise do Chat GPT e outras tecnologias generativas, em especial, a possibilidade de uso de tais tecnologias no âmbito do Ministério Público Federal, respectivas vantagens para a Instituição bem como riscos associados e formas de mitigá-los.

2. Inicialmente, será realizada sucinta análise técnica das tecnologias citadas, com foco no GPT ("Generative Pre-trained Transformer"), nas versões 3.5 e 4.

3. Serão, então, abordadas as diversas possibilidades de uso de forma institucional.

4. Por fim, serão elencados os principais riscos e respectivas sugestões de ações para mitigá-los ou dirimi-los.

## 2 CHAT GPT E TECNOLOGIAS GENERATIVAS - O QUE SÃO E COMO FUNCIONAM

5. Inicialmente, é importante esclarecer o que é a tecnologia GPT e como funciona.

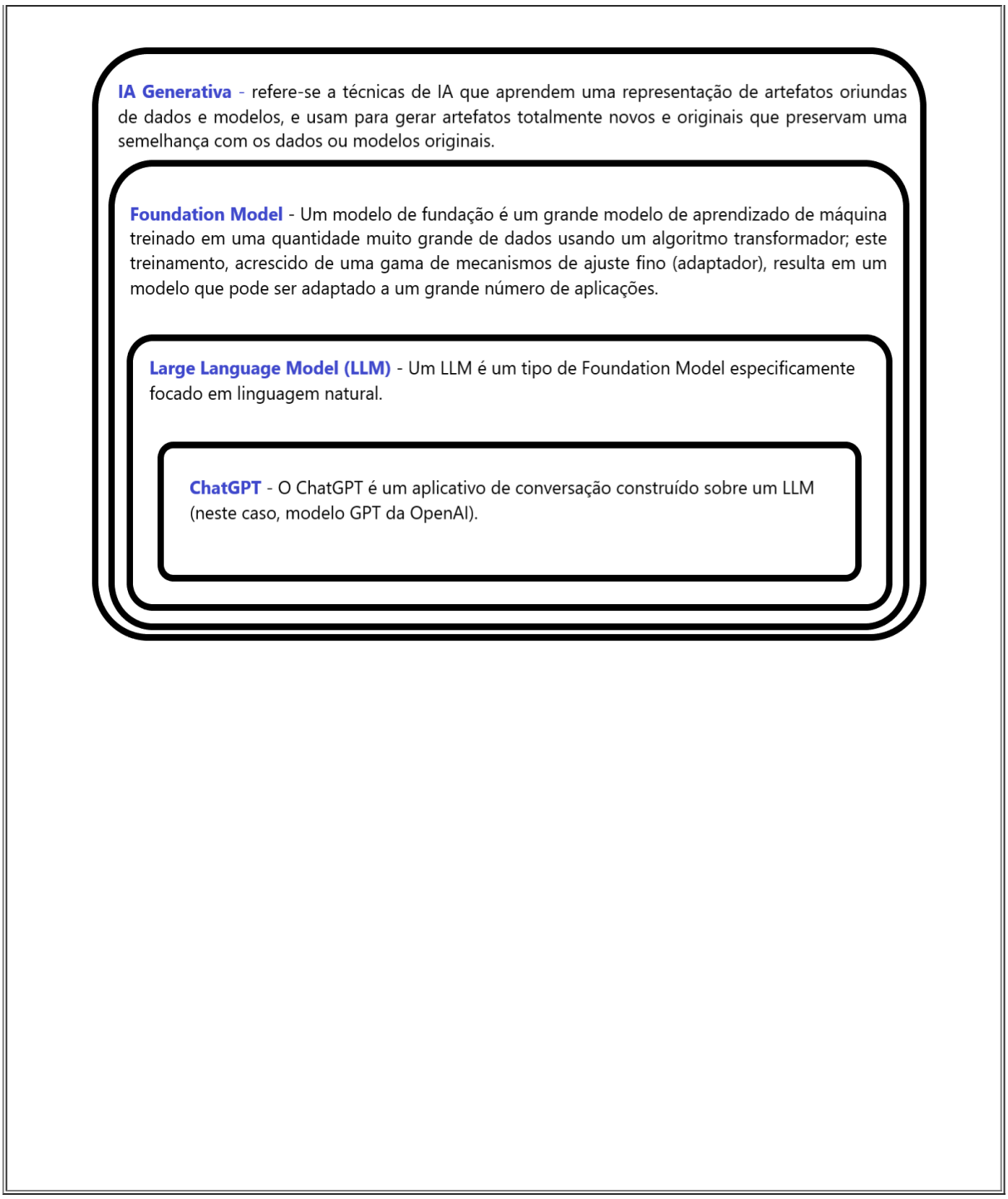
6. O **GPT**, ou "**Generative Pre-trained Transformer** ", é uma tecnologia de IA Generativa que utiliza grandes modelos de linguagem natural **LLM** (do inglês "**Large Language Models**") treinados para gerar respostas coerentes em diálogos, com base em estímulos pré-determinados.

7. Esses modelos funcionam através da análise de vastos conjuntos de dados, o que lhes permite aprender com base em padrões observados e gerar respostas semelhantes às de um ser humano.

8. Dentre os modelos LLM, os **modelos fundacionais** (do inglês "**Foundation model**") são as opções mais utilizadas, uma vez que são mais generalizados e permitem que o usuário treine o diálogo em diferentes contextos.

9. O ChatGPT, por sua vez, é um serviço da OpenAI. Ele incorpora um chatbot de conversação alimentado por um transformador generativo pré-treinado (especificamente, GPT-3.5) que foi treinado em 2021, lançado em novembro de 2022 e rapidamente se tornou viral. Possui recursos impressionantes de compreensão e geração de linguagem natural e pode gerar uma ampla gama de formatos e estilos de escrita.

10. O ChatGPT foi treinado em um corpus de qualidade, veracidade e viés variados, o que pode ser refletido nos resultados. Ele também pode entregar respostas plausíveis, mas incorretas. Independentemente desse aviso, as pessoas o estão aplicando a muitos casos de uso em que são necessários resultados objetivos e inofensivos.




11. O ChatGPT é apenas um exemplo de uma tendência muito maior de usar modelos de base e outras técnicas para habilitar a IA generativa em um amplo número de casos de uso.

## 2.1 GPT-4

12. O GPT-4 é o modelo de linguagem grande mais recente e avançado da OpenAI, lançado em março de 2023. É um avanço do GPT-3.0 e usa uma arquitetura de modelo de transformador para prever o próximo token (uma parte de uma palavra).

13. O GPT-4 foi treinado em grandes quantidades de dados públicos e alguns privados, e o modelo foi ajustado usando aprendizado por reforço com feedback humano.

14. Os principais recursos novos ou aprimorados no GPT-4 incluem a capacidade de oferecer suporte a entradas de imagem e texto, bem como suporte para formas mais longas de entrada e saída.

15. As melhorias de qualidade do GPT-4 estão na precisão, factualidade e desempenho em inglês e em outros idiomas; melhor desempenho em tarefas quantitativas e condução de conversas; e filtros de proteção para evitar resultados nocivos e indesejáveis. Estes filtros são baseados em testes da OpenAI e ainda precisam ser totalmente comprovados no mundo real.

16. Como seus predecessores, o GPT-4 está sujeito a alucinações e vieses. A OpenAI afirma ter melhor desempenho nesses fatores, o que ainda precisa ser comprovado com o uso ao longo do tempo.

## 2.2 PRINCIPAIS DIFERENÇAS ENTRE O GPT-4 e o GPT-3.5

### 2.2.1 Capacidade de entrada multimodal

17. GPT-4 é multimodal, o que significa que pode aceitar entrada de imagem e usá-la para geração de texto, enquanto as versões anteriores eram restritas à entrada de texto. Ele pode não apenas extrair o conteúdo da imagem, mas também determinar o contexto e o relacionamento das partes. Esse recurso permite a mistura de entrada visual com prompts de texto para gerar várias respostas textuais. Este é apenas um primeiro passo em direção à multimodalidade, já que o GPT-4 atualmente não permite outros modos de entrada (por exemplo, áudio ) e saída (por exemplo, saídas visuais).

### 2.2.2 Aceitação de entrada de texto mais longo

18. Na versão padrão, o GPT-4 aumenta o tamanho do texto de entrada que pode ser enviado para aproximadamente 3.000 palavras (para a versão básica) e 21.000 palavras (para a versão estendida), um aumento significativo do GPT- 3.5. Isso permite documentos, manuais e outras literaturas mais longas que podem ser usados com o modelo para fornecer

conteúdo e contexto adicionais que podem ser usados nas respostas.

## 2.3 MELHORIAS DO GPT-4, DE ACORDO COM OpenAI

### 2.3.1 Melhor desempenho em outros idiomas

19. De acordo com o OpenAI, as respostas do GPT-4 na maioria dos idiomas testados (24 de 26) foram melhores do que o desempenho do GPT-3.5 em inglês para um determinado benchmark.

### 2.3.2 Precisão e factualidade aprimoradas

20. A OpenAI afirma que o GPT-4 reduz significativamente a alucinação, marcando 19 pontos percentuais a mais em média em nove diferentes medidas de precisão projetadas internamente. A conexão com um serviço que pode acessar a internet, como o Bing Chat, permite citações factuais de fontes atuais, suportando a veracidade e precisão da saída, aproximando-se dos casos de uso em que o suporte factual é necessário.

### 2.3.3 Recursos generativos aprimorados:

21. O modelo supera os modelos anteriores de última geração em relação a vários benchmarks humanos, incluindo exames acadêmicos e profissionais. Isso não significa necessariamente que o GPT-4 terá um bom desempenho nessas áreas além do desempenho do teste.

### 2.3.4 Alinhamento intencional e dirigibilidade:

22. O modelo tem desempenho melhor do que o GPT-3.5 em termos de capacidade de seguir a intenção do usuário. Em um teste OpenAI, as respostas GPT-4 foram preferidas em relação a GPT-3.5 por humanos 70% das vezes. Da mesma forma, a API GPT-4 inclui uma “mensagem do sistema” específica que pode ser usada para orientar o comportamento do chatbot durante a conversa.

### 2.3.5 Aprimoramentos criativos

23. Recursos adicionados ao GPT-4 são a capacidade de manter o contexto e a coerência das histórias dentro de um corpo maior de saída gerada e aplicar estilos criativos para estilos narrativos aplicados a um resultado (por exemplo, um poema de Shakespeare sobre o aquecimento global).

### 2.3.6 Raciocínio complexo

24. A capacidade de compreender, analisar e resolver problemas científicos complexos em áreas como matemática (cálculo, geometria), física, química e biologia é bastante aprimorada com o GPT-4.

### 2.3.7 Melhorias de segurança

25. A mitigação de riscos, proteções e testes adicionais melhoraram a segurança, reduzindo o comportamento incorreto em prompts confidenciais e não permitidos, minimizando a geração de conteúdo tóxico.

## 2.4 ONDE O GPT-4 ESTÁ DISPONÍVEL

26. A seguir, elencamos alguns produtos e serviços que oferecem GPT-4 ou disponibilizam GPT-4 para recursos novos ou melhores.

### 2.4.1 Produtos e serviços da OpenAI

27. **ChatGPT Plus:** O plano de assinatura do ChatGPT tem acesso antecipado ao GPT-4 (o GPT-4 não está disponível atualmente na versão gratuita do ChatGPT).

28. **API OpenAI:** GPT-4 está disponível por meio de uma API para desenvolvedores (atualmente existe uma lista de espera).

### 2.4.2 Parceria da Microsoft

29. **Serviço Azure OpenAI:** A Microsoft anunciou que o GPT-4 está disponível por meio de seu serviço Azure OpenAI.

30. **Pesquisa do Bing:** O novo Bing é executado em GPT-4 personalizado para pesquisa.

31. **Microsoft 365 Copilot:** oferece a integração de diferentes LLMs, incluindo GPT-4, com seus dados de negócios nos aplicativos Microsoft Graph e Microsoft 365 (como Word, Excel, PowerPoint, Outlook e Teams). Um novo aplicativo - Business Chat - está incluído, que pode realizar trabalhos no LLM, nos aplicativos da Microsoft e nos dados do usuário (como calendários, e-mails, bate-papos, documentos, reuniões e contatos).

32. **Microsoft Dynamics 365 Copilot:** Isso inclui a adição do serviço Azure OpenAI e recursos baseados em GPT em aplicativos de negócios da Microsoft, incluindo vendas, serviços, marketing e cadeia de suprimentos.

33. **Copilot do GitHub:** O novo GitHub Copilot integra recursos GPT-4 para ir além dos recursos de preenchimento automático de geração de código, adicionando uma nova interface de bate-papo e a capacidade de dar comandos verbais.

34. **Copilot em Power Platform:** um copiloto de IA está disponível por meio do Microsoft Power Apps, Power Automate e Power Virtual Agents que permite o desenvolvimento usando linguagem natural.

35. **Por meio de aplicativos e fornecedores que usam o GPT-4:** existem alguns exemplos iniciais de organizações que usam ou avaliam o GPT-4, incluindo Duolingo, Stripe e Khan Academy. A Salesforce também incorporará o GPT-4 como um assistente digital em seus serviços em nuvem. Esses serviços estão sendo implementados gradualmente. À medida que mais desenvolvedores usam a API GPT-4, o modelo será implantado amplamente para diversos casos de uso e aplicativos.

## 2.5 CONCORRENTES DO GPT

36. Análises extensas de modelos concorrentes estão além do escopo deste documento que, para fins didáticos, detalha com mais cuidado o ChatGPT, possibilidades de usos e riscos associados, bem como tecnologias relacionadas.

37. Entretanto, faz-se necessário elencar algumas tecnologias que realizam ações semelhantes às passíveis de serem realizadas pelo GPT-4. São elas:

- BERT, PaLM e LaMDA do Google
- LLaMA da Meta
- ERNIE da Baidu
- Claude (assistente de IA baseado em um LLM) da Anthropic Models
- NeMo da NVIDIA
- Modelos de código aberto, como o BLOOM da Big Science Initiative

## 3. POSSIBILIDADES DE USO DE FORMA INSTITUCIONAL

38. São inúmeros os artefatos que podem ser gerados com a IA Generativa. Abaixo alguns exemplos de uso:

**AUDIOVISUAL:**

Rostos e corpos em 2D e 3D, desenhos, fotografias, arte, personagens (voz e vídeos), música.

**PROGRAMAÇÃO E ATIVOS DE DADOS:**

Treinamento de dados, por exemplo (eventos, comportamento, etc.); teste e dados sintéticos; código de softwares; converter código de uma linguagem de programação para outra, corrigir código errado, explicar código; criar dados de contexto, por exemplo, condições de mercado, etc.

**LINGUAGEM NATURAL:**

Perguntas e respostas; resumos; estórias, narrativas; documentos; listagem

de tarefas.

**DESIGN:**

Projetos de produtos; materiais; modelos 3D; mundos de jogos; fluxos de processo; esquemas e especificações, por exemplo, drogas, infraestrutura e edifícios, edições de vídeo, etc.

39. No Ministério Público Federal poder-se-ia utilizar ferramentas como o ChatGPT de inúmeras formas, trazendo enorme vantagem para a Instituição. Citamos abaixo, apenas a título de exemplo, alguns possíveis usos.

## 3.1 CRIAÇÃO E MELHORA NA CRIAÇÃO DE CONTEÚDO DOCUMENTAL.

40. Ferramentas como o GPT-4 podem produzir "rascunhos" de texto, que são então revisados pelo usuário. Podem produzir o tamanho e o estilo de texto desejados. Podem produzir um texto modificado para suavizar a linguagem ou profissionalizar a escrita, por exemplo, pegando documento formal e fazendo um texto conversacional para e-mail ou chat.

41. Além disso, oferece a possibilidade de fazer resumos de conversas, artigos, e-mails e páginas web para o tamanho especificado. Pode converter textos em listas com tópicos; criar títulos, notas, e extrair conteúdos-chave de outros conteúdos; pode sumarizar, revisar conteúdos classificando-os em áreas definidas.

## 3.2 MELHORIA NA GESTÃO E NA TOMADA DE DECISÃO A PARTIR DE DADOS ESTRUTURADOS

O GPT-4 possui uma grande capacidade de extração de dados estruturados a partir de documentos textuais (peças judiciais, representações, documentos administrativos, etc), o que pode ser usado para aprimorar a qualidade e aumentar a quantidade de dados disponíveis para melhoria na gestão e na tomada de decisão pelo MPF. Pode, assim, auxiliar membros e servidores a tomarem decisões mais informadas. Os chatbots podem analisar informações e gerar relatórios baseados em dados, ajudando a fornecer insights importantes que podem ser usados na tomada de decisões.

## 3.3 POSSIBILIDADE DE USO NA IDENTIFICAÇÃO DE FRAUDES E ILEGALIDADES

42. O ChatGPT e análogos podem ser treinados para procurar por palavras-chave e outros indicadores de possíveis fraudes, corrupção e outras ilegalidades, levando a uma maior eficácia no combate a essas práticas.

### 3.4 AGILIDADE NA RESPOSTA A PROCESSOS

43. Graças à sua capacidade de processar grandes quantidades de dados em pouco tempo, essa tecnologia pode ser usada para ajudar os membros do MPF a lidar com o grande volume de informações contido nos processos judiciais.

### 3.5 SUPORTE TÉCNICO

44. Outra forma de uso do chat GPT no MPF é oferecer suporte técnico aos usuários da instituição. Os chatbots podem ser programados para responder perguntas frequentes sobre o funcionamento de sistemas, programas e outros recursos tecnológicos, o que agiliza e descomplica o suporte técnico.

### 3.6 DESEMPENHO AUMENTADO NA ÁREA ADMINISTRATIVA

45. A capacidade de análise de grandes blocos de informações e posterior criação de artefatos poderia ser muito útil nas áreas de gestão de pessoas e contratações do MPF. Em especial, nesta última, cujo processo é moroso tendo em vista o volume e a complexidade dos inúmeros processos e informações que devem ser analisados, atualmente, por servidores. O uso de tal tecnologia poderia acelerar muito, além de trazer grande grau de assertividade.

46. Em resumo, a IA generativa pode ser uma ferramenta valiosa para o MPF, oferecendo uma variedade de benefícios para a instituição. Desde a prestação de atendimento ao público interno e, eventualmente o externo, até a identificação de ilícitos e tomada de decisões mais informadas, a tecnologia pode ajudar a melhorar o desempenho e a eficácia da instituição, levando a uma maior eficiência e satisfação no trabalho.

## 4. ANÁLISE DOS RISCOS ASSOCIADOS E FATORES DE MITIGAÇÃO

47. Não obstante todas as vantagens apresentadas, o uso do GPT apresenta riscos que lhe são inerentes. O texto gerado pelo ChatGPT depende do prompt, tarefa ou domínio e da qualidade e quantidade dos dados de treinamento levando a riscos de modelo e risco de uso indevido.

48. Abaixo, listamos alguns riscos associados ao uso do GPT-4, de acordo com o GPT-4 Technical Report, da OpenAI e de diversas pesquisas realizadas pela empresa de consultoria em gestão de TI, Gartner, todos referenciados ao final deste texto.

## 4.1 QUAIS SÃO OS RISCOS DO USO DO GPT-4?

### 4.1.1 imprecisões e Alucinações

49. Apesar das melhorias, o GPT-4 pode "alucinar" fatos e cometer erros de raciocínio. Isso limita sua aplicação a casos de uso em que não é necessária alta precisão.

50. Esse risco se estende ao código gerado com GPT-4, que pode apresentar vulnerabilidades. De acordo com a OpenAI, o desempenho do GPT-4 é 19% superior ao do GPT-3.5 em avaliações factuais. Ainda assim, em uma variedade de categorias, o desempenho factual varia aproximadamente de 70% a 80%.

51. ***Registre-se que não está a se tratar de um buscador ou mecanismo de pesquisa. O que o GPT faz é gerar/prever possível palavra, texto ou outro artefato baseado em todo o seu estudo pregresso. É dizer que ele gera/cria informação, não busca em uma base a informação já existente. Daí o risco de ser gerada informação imprecisa que sempre necessitará da revisão humana.***

### 4.1.2 Viés

52. O GPT-4 pode gerar uma saída que representa preconceitos sociais e visões de mundo derivadas de seus dados de treinamento. Uma análise da OpenAI descobriu que "o modelo tem o potencial de reforçar e reproduzir preconceitos e visões de mundo específicos, incluindo associações estereotipadas e degradantes prejudiciais para certos grupos marginalizados".

### 4.1.3 Solicitação contraditória

53. Apesar das melhorias, ainda é possível gerar solicitações contraditórias para GPT-4 ("jailbreaking" do modelo), o que pode levar à geração de conteúdo prejudicial. Os modelos de fundação reduzem o custo de criação de conteúdo, o que significa que fica mais fácil criar deepfakes (que se assemelham ao original) e até mesmo hiperpersonalizá-lo.

### 4.1.4 Privacidade

54. As garantias de privacidade variam para as versões API e não API do OpenAI do GPT-4.

55. Por padrão, os dados enviados para versões não API (como ChatGPT Plus) podem ser usados para melhorar os modelos (existe uma maneira de desativar).

56. Por outro lado, **as versões da API têm mais garantias de privacidade**, mas ainda é importante ler os termos de serviço para avaliar o risco de privacidade para seus casos de uso específicos.

57. No âmbito do MPF, por exemplo, se o GPT-4 for utilizado tanto para a área finalística quanto para a área-meio, deve se ter o máximo cuidado em relação a dados sensíveis e dados pessoais. Para isso, sugere-se que seja desenvolvida uma maneira de desidentificar expedientes que contenham tais dados para não se correr o risco de eventual exposição de tais dados.

### 4.1.5 Risco na Proteção da Propriedade Intelectual e direitos do criador

58. Este é um padrão que vem de versões anteriores. Provavelmente estará sujeito a um exame mais minucioso relativo à propriedade intelectual quando a produção for feita com envolvimento generativo e humano híbrido, bem como usando direitos autorais protegidos para geração de nova produção intelectual.

### 4.1.6 Riscos de uso indevido

59. **Deepfakes** — Essas saídas geradas pelo ChatGPT podem parecer realistas, mas na verdade são conteúdos falsos. O Ministério Público deve estar atento para identificar notícias falsas, desinformação, falsificações ou tentativas de manipular a opinião pública.

60. **Fraude e abuso** — Pessoas mal-intencionadas já estão explorando o ChatGPT escrevendo avaliações falsas, spam e phishing. Como qualquer fraude, o uso do ChatGPT para fins maliciosos será contínuo. É importante que especialistas apropriados da instituição sejam informados sobre os riscos do ChatGPT e estejam envolvidos em atividades defensivas.

## 4.2 SEGURANÇA E GESTÃO DE RISCOS

61. Face ao já exposto, o uso do GPT, além de ser enormemente vantajoso, traz uma série de riscos inerentes ao seu uso. Tais riscos, entretanto, podem ser mitigados ou dirimidos. Apresentamos abaixo algumas medidas que entendemos como capazes de mitigar riscos e, em alguns casos, até mesmo eliminá-los.

### 4.2.1 Orientar servidores e membros sobre os riscos de segurança e parâmetros adequados de uso

62. Todos os servidores e membros que utilizam ou utilizarão o OpenAI ChatGPT ou tecnologia semelhante devem ser instruídos a tratar as informações que publicam como se estivessem postando em um site público (por exemplo, uma rede social ou um blog público).

63. Eles não devem postar informações de identificação pessoal, informações do MPF ou de qualquer cidadão que não estejam geralmente disponíveis ao público. Atualmente não há garantias claras de privacidade ou confidencialidade, especialmente, nos modelos gratuitos.

64. Além disso, as informações postadas podem ser usadas para treinar ainda mais o modelo. Em relação especificamente ao ChatGPT, já se sabe que a Microsoft apresentará garantias de privacidade para seu serviço Azure OpenAI ChatGPT, assim como faz para seus outros serviços de software.

### 4.2.2 Orientar usuários a não confiar apenas nos filtros de conteúdo do ChatGPT

65. O ChatGPT tem um filtro tóxico para entradas e saídas, mas esta é uma tarefa complexa de realizar e requer uma boa modelagem contextual e um alto nível de nuances. Portanto, os usuários não devem confiar em sua saída do ponto de vista de conformidade ou risco e devem garantir que os humanos estejam no circuito para avaliar as entradas e saídas.

### 4.2.3 Orientar usuários de que suas conversas com ChatGPT podem, eventualmente, ser vistas por outras pessoas

66. Os provedores de serviços do ChatGPT (OpenAI e Microsoft) podem revisar as conversas para melhorar seus sistemas e garantir que o conteúdo esteja em conformidade com suas políticas e requisitos de segurança.

67. Não há garantias em relação a funcionários, contratados ou parceiros que possam visualizar as informações que você publica. Espera-se que a versão do serviço do Azure siga os demais serviços existentes do Azure OpenAI a esse respeito, em que se confere maior grau de privacidade.

### 4.2.4 Orientar usuários sobre o uso dos dados

68. As conversas com o ChatGPT serão usadas para treinamento e podem ser revistos pelos formadores. Não é possível excluir prompts específicos, portanto, cuidado com o que você compartilha. É possível excluir uma conta, mas isso não excluirá os dados de treinamento.

### 4.2.5 Orientar sobre o risco dos vieses

69. O ChatGPT é ajustado ou aprimorado para se alinhar às preferências dos treinadores (classificação do rotulador), em vez de fatos verificados, o que significa que sua saída é plausível, mas não confiável para muitos casos de uso.

70. Além disso, o viés pode estar presente nos grandes conjuntos de dados usados para treinar o modelo GPT-3 subjacente. Apesar dos esforços da OpenAI para minimizar o viés e a discriminação no ChatGPT, já existem casos conhecidos desses vieses.

### 4.2.6 Criar uma política institucional do MPF em torno do ChatGPT e ferramentas análogas

71. Mais do que provável, já é sabido que muitos servidores e membros já estão utilizando o ChatGPT seja para experimentação ou até mesmo para apoiar suas tarefas relacionadas ao trabalho.

72. O bloqueio total pode levar ao uso "sombra" do ChatGPT, ao mesmo tempo em que fornece às organizações uma falsa sensação de conformidade.

73. Assim, todos os que utilizam o ChatGPT devem ser instruídos a tratar a informação que publicam como se a estivessem a publicar em um site público, como uma rede social ou blog público.

74. Uma abordagem sensata seria monitorar o uso, incentivar a inovação, mas garantir que seja usado apenas como inteligência aumentada dos servidores e membros e não seja usado sem filtros e sem a devida supervisão humana com cidadãos e demais interlocutores do MPF.

### 4.2.7 Implementar a revisão manual do modelo para detectar resultados incorretos ou mal informados

75. Revisar manualmente todos os resultados do modelo e use-os apenas como uma primeira ferramenta de rascunho. Algumas saídas podem levar a responsabilidade comercial, perda de reputação ou danos.

### 4.2.8 Proibir os usuários, por meio de políticas claramente compreendidas e amplamente divulgadas, de fazer perguntas ao ChatGPT que exponham dados organizacionais ou pessoais confidenciais

76. Exemplos, tais políticas podem incluir:

> **a) E**xigir a revisão humana de todos os resultados do ChatGPT usados ¿¿nas interações com o cliente para garantir a qualidade e a precisão das informações;
>
> **b)** Proibir qualquer recortar e colar de conteúdo corporativo, como e-mails, relatórios, logs de bate-papo, nos prompts do ChatGPT;
>
> **c )** Não permitir quaisquer entradas no ChatGPT que incluam dados de cidadãos ou pessoal interno que já não estejam publicados.

**4.2.9 Monitorar usos não sancionados do ChatGPT com controles e painéis de segurança existentes para detectar violações de políticas (por exemplo, firewalls podem bloquear o acesso de usuários corporativos, informações de segurança e sistemas de gerenciamento de eventos podem monitorar logs de eventos em busca de violações e gateways da Web seguros podem monitorar chamadas de API não permitidas).**

**4.2.10 Aplicar as medidas supracitadas por meio de Engenharia de Prompt.**

Podem ser criados prompts projetados especificamente para uso direcionado a determinadas atividades de forma que seu uso seja feito com a privacidade de dados sensíveis e com segurança. Além disso, eles podem representar um corpus de prompts bem ajustados (fine-tuned) e altamente desenvolvidos que podem ser mais facilmente reutilizados, e compartilhados.

77. Entendemos que **o ChatGPT deve ser usado em uma configuração controlada onde o texto gerado pode ser avaliado e ajustado adequadamente. Além disso, é de todo recomendável que seja criada uma forma de desidentificar previamente documentos que eventualmente venham a ser submetidos em sua totalidade ao ChatGPT**.

78. Além disso, **é recomendável o uso ChatGPT do Azure OpenAI Service em vez do ChatGPT do OpenAI** para uso em sua corporações como o MPF, pois **a Microsoft oferece segurança corporativa e controles de conformidade já associados a outros produtos da Microsoft.** Em contraste, o ChatGPT da OpenAI não oferece nenhuma segurança de nível empresarial em sua "visualização de pesquisa" nem em sua edição "Plus".

79. Se adotada a ferramenta, é importante que se monitore as práticas de segurança do Azure ChatGPT depois de implementadas para garantir que os dados estejam seguros, conforme anunciado.

80. Para uso imediato do ChatGPT (do Azure ou do OpenAI), é imperioso que os humanos revisem cada saída para detectar resultados incorretos, mal informados ou tendenciosos.

81. É altamente recomendável a implementação de uma política que proíba os usuários de fazer perguntas ao ChatGPT da OpenAI que revelem dados institucionais confidenciais.

## 5. CONCLUSÃO

82. Diante das possibilidades de uso do ChatGPT pelo Ministério Público Federal, é importante ressaltar que, apesar dos riscos que essa ferramenta possa apresentar, seu uso é preferível à sua proibição. Isso se deve ao fato de que ela apresenta inúmeras vantagens, que podem contribuir significativamente para a atuação do MPF.

83. Ao mesmo tempo, é importante lembrar que, como toda ferramenta tecnológica, ela apresenta riscos, como o vazamento de informações sigilosas e a manipulação de dados. Por isso, é fundamental que o MPF adote medidas de segurança para garantir a integridade das informações trocadas por meio do chat.

84. No entanto, é importante ressaltar que os riscos apresentados não deveriam ser vistos como razões para sua proibição. Pelo contrário, é necessário que o MPF esteja atento a esses riscos e adote medidas adequadas para minimizá-los, como a capacitação dos usuários internos e a definição de uma forma segura e balizada de uso.

## 6. REFERÊNCIAS


- Gartner Addresses Frequently Asked Questions on ChatGPT
- Gartner - Quick Answer: How Can You Manage Trust, Risk and Security for ChatGPT Usage in Your Enterprise?
- ChatGPT Research Highlights
- Data, privacy, and security for Azure OpenAI Service
- Gartner - GPT-4 Technical Report
- Gartner - GPT-4 - OpenAI
- Our approach to AI safety - Ensuring that AI systems are built, deployed, and used safely is critical to our mission
- AI-written critiques help humans notice flaws - OpenAI
- Gartner Webinar: Beyond the Hype: Enterprise Impact of ChatGPT and Generative AI
- Gartner - Quick Answer: What Is ChatGPT?
- Gartner - Quick Answer: What Is GPT-4?
- Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk
- Introducing Microsoft 365 Copilot - Your copilot for work
- GitHub Copilot X: The AI-powered developer experience
- Introducing GPT-4 in Azure OpenAI Service
- MARCH 14 2023 Confirmed: the new Bing runs on OpenAI's GPT-4, Microsoft Bing Blogs

- Product, OpenAI
- Inteligência artificial é aprovada no concurso da diplomacia
- Google: Primeiros passos com o algoritmo BERT integrado
- Google: Ai Across Google - PaLM2
- Google: LaMDA: our breakthrough conversation technology
- Meta AI: Introducing LLaMA: A foundational, 65-billion-parameter large language model
- Baidu - ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology
- Anthropic Models - Introducing Claude
- NVIDIA NeMo
- BigScience Initiative - Introducing The World's Largest Open Multilingual Language Model: BLOOM

É o Relatório.

Brasília, 16 de maio de 2023.

**PAULO HENRIQUE CORREA DE MOURA**

SECRETARIO ADJUNTO DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO