AVANTGARDE
ASSET MANAGEMENT

Avantgarde Asset Management Gestão de Recursos Ltda.
Avenida das Nações Unidas, 12399 – 7º andar
Brooklin – São Paulo – SP
CEP: 04578-000
avantgardeam.com.br
+55 11 3400-7061

Março, 2023

Performance do ChatGPT no exame CFA

> "ChatGPT took one giant step for artificial intelligence, and one massive leap for humanity"
>
> ChatGPT parafraseando Neil Armstrong

Introdução

O ChatGPT, um *large language model (LLM)* lançado pela OpenAI no final de 2022, impressiona por sua capacidade quase humana de responder, com bastante sofisticação e fluência, à perguntas complexas em diferentes campos do conhecimento. Como consequência, a plataforma tem atraído bastante atenção e gerado debates (e controvérsias) sobre seu impacto nas diversas esferas da sociedade, como saúde, educação, pesquisa científica e mercado de trabalho.

A repercussão inicial do ChatGPT foi notável, com mais de um milhão de usuários em apenas cinco dias. Essa taxa de adoção supera, por muito, a velocidade de outras plataformas tecnológicas populares, como o Instagram e Netflix, o que mostra a magnitude do seu impacto. Como resultado, os pesquisadores têm se dedicado a explorar como essa nova tecnologia pode afetar suas áreas de pesquisa e a sociedade em geral. Dentre esses novas campos de estudo, uma tendência recente é a avaliação do desempenho da ferramenta em diversos exames de admissão e proficiência.

Um trabalho, conduzido por Winter (2023), demonstrou que o ChatGPT alcançaria uma nota comparável à média dos estudantes no exame de compreensão de leitura em inglês na Holanda. Ademais, Choi, Hickman, et al. (2023) apresentam resultados que indicam que a ferramenta teria uma nota média de C+ em quatro cursos de direito da Universidade de Minnesota. Embora essa nota seja baixa, ela seria suficiente para aprovação nas disciplinas. Por fim, Tiago Tavares, pesquisador do Insper, mostra que a IA teria uma nota de 614 na prova do ENEM, o que poderia ser suficiente para aprovação em alguns cursos pelo Sisu.

Desse modo, nesta Carta, seguiremos essa linha de pesquisa e analisaremos o desempenho dos diferentes modelos desenvolvidos e disponibilizados pela OpenAI no exame de certificação *Chartered Financial Analyst (CFA)*. O programa CFA é composto por três níveis de exame que abrangem tópicos em investimentos, ética e padrões profissionais. Esse exame é conhecido por ser rigoroso e exige uma média de 300 horas de estudo por nível. A taxa de aprovação é, geralmente, inferior a 50%, tornando essa certificação uma das mais importantes e respeitadas no mercado financeiro.

Metodologia

Antes de apresentarmos os resultados, é necessário esclarecer dois pontos: como iremos simular o exame CFA e quais modelos utilizaremos. Dado que o *CFA Institute* não disponibiliza provas anteriores, iremos utilizar os simulados gratuitos do IFT - um instituto autorizado pelo *CFA Institute* para oferecer cursos preparatórios para os três níveis do exame. Os simulados têm uma

Avantgarde Asset Management Gestão de Recursos Ltda.
Avenida das Nações Unidas, 12399 – 7º andar
Brooklin – São Paulo – SP
CEP: 04578-000
avantgardeam.com.br
+55 11 3400-7061

estrutura bastante semelhante, com o simulado do nível um tendo 90 perguntas, enquanto que os simulados dos níveis dois e três têm 44 questões cada. Em todos os casos, as questões são compostas por uma pergunta seguida de três alternativas de múltipla escolha (A, B e C).

Em relação aos modelos, vamos considerar quatro: "davinci-001" e "davinci-002", ambos da família GPT-3 e lançados em maio de 2020 e janeiro de 2022, respectivamente; "davinci-003" e o "gpt-3.5-turbo", pertencentes à família GPT-3.5 e lançados em novembro de 2022 e março de 2023, respectivamente[1]. Ao analisarmos essas diferentes IAs, seremos capazes de avaliar como a evolução dessa tecnologia afeta o desempenho no exame.

Para garantir uma resposta mais precisa, o *prompt* começa com uma instrução de que a pergunta é parte de um teste que contém A, B e C como alternativas. Em seguida, fornecemos a pergunta e as opções de resposta disponíveis. Visando diminuir a aleatoriedade, ajustamos o parâmetro de temperatura do modelo para 0,2.

Por fim, em alguns casos, a resposta fornecida pelo *LLM* não é adequada para propósitos de avaliação. Isso acontece, por exemplo, quando a ferramenta indica que a alternativa correta é D, mesmo essa opção de resposta não existindo. Nesses casos, consideramos a resposta como errada. Para corrigir esse problema, uma possibilidade seria repetir a pergunta até que o modelo retorne uma alternativa válida.

Resultados

Table 1: Desempenho no Exame CFA por Modelo e Nível

Modelo	**Respostas Corretas**	**Número de Questões**	**% Corretas**
Painel A: CFA Nível 1			
davinci-001	40	90	44
davinci-002	45	90	50
davinci-003	53	90	59
gpt-3.5-turbo	54	90	60
Painel B: CFA Nível 2			
davinci-001	16	44	36
davinci-002	26	44	59
davinci-003	16	44	36
gpt-3.5-turbo	21	44	48
Painel C: CFA Nível 3			
davinci-001	17	44	39
davinci-002	21	44	48
davinci-003	23	44	52
gpt-3.5-turbo	27	44	61

[1] O site lifearchitect.ai fornece um *timeline* do desenvolvimento desses modelos e uma visão geral da família GPT-3.

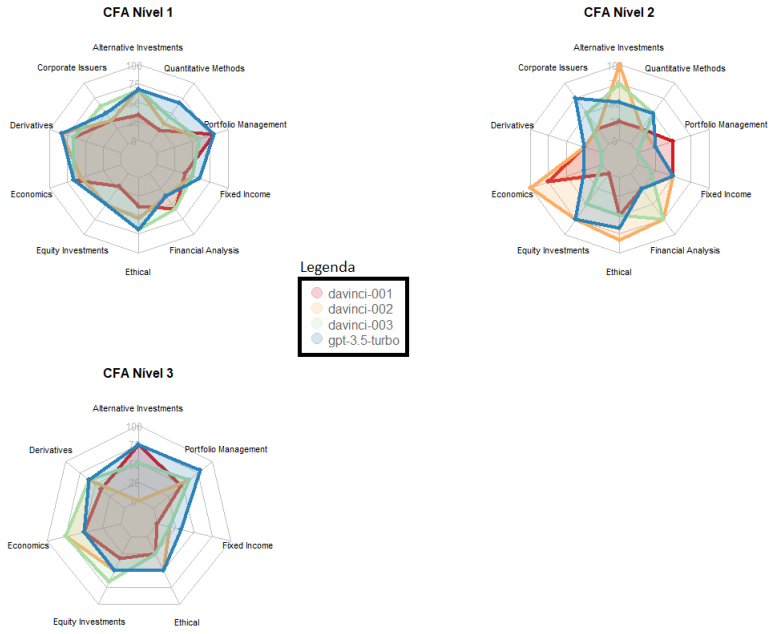
AVANTGARDE
ASSET MANAGEMENT

Avantgarde Asset Management Gestão de Recursos Ltda.
Avenida das Nações Unidas, 12399 – 7º andar
Brooklin – São Paulo – SP
CEP: 04578-000
avantgardeam.com.br
+55 11 3400-7061

Analisando a Tabela 1, podemos observar uma relação quase perfeitamente monotônica entre a porcentagem de acertos e a data de lançamento dos *LLMs*, exceto pelo desempenho do "davinci-002" no nível 2. Assim, é perceptível que as IAs mais antigas apresentaram um desempenho inferior em relação aos modelos mais recentes. Mesmo essa diferença já sendo esperada, a magnitude das discrepâncias impressiona, demonstrando que houve uma melhora substancial na performance em menos de três anos.

Passando para uma visão um pouco mais granular, percebemos que o "davinci-001" - primeiro modelo da família GPT-3 - tem um desempenho não muito superior a um "chute". Em contrapartida, o "gpt-3.5-turbo" teve uma performance bem melhor, quase que duplicando a probabilidade base de acerto (33%) para os níveis um e três. Mesmo com essa melhora expressiva, os modelos, muito provavelmente, não seriam capazes de conseguir a aprovação em todos os níveis da certificação. Desse modo, podemos afirmar que os *LLMs* atuais ainda não atingem o grau de conhecimento esperado de um *CFA Charteholder*[2].

Figure 1: Desempenho no Exame CFA por Modelo, Nível e Área do Conhecimento



Em relação à Figura 1, podemos analisar o desempenho de cada modelo por área do conhecimento presente na prova. Uma observação interessante é que embora uma IA possa ter um desempenho superior no agregado, ele não necessariamente domina os outros modelos em todas as áreas. O "davinci-001", por exemplo, mesmo sendo mais defasado, apresenta um desempenho superior em relação ao "gpt-3.5-turbo" nas áreas de Economia e Análise Financeira. Essa constatação

[2]Isso está em consonância com o artigo do Financial Times.

Avantgarde Asset Management Gestão de Recursos Ltda.
Avenida das Nações Unidas, 12399 – 7º andar
Brooklin – São Paulo – SP
CEP: 04578-000
avantgardeam.com.br
+55 11 3400-7061

é uma motivação valiosa para testar um algoritmo que agregue as respostas de diferentes modelos (*ensemble*) a fim de gerar uma resposta final mais precisa e abrangente.

Além disso, é importante destacar que a performance dos modelos parece deteriorar em áreas que exigem conhecimentos matemáticos mais avançados. Isso é evidenciado pelo desempenho inferior nas áreas de derivativos, métodos quantitativos e renda fixa. Essa é uma limitação já conhecida na literatura (Frieder, Pinchetti, et al. (2023)). No entanto, de acordo com Shakarian e Koyyalamudi (2023), pode haver uma melhoria na acurácia das respostas se solicitarmos que os modelos forneçam uma resposta passo a passo, detalhando o processo de raciocínio utilizado.

Conclusão

Com base no exposto anteriormente, fomos capazes de evidenciar uma evolução significativa, e rápida, na qualidade dos modelos desenvolvidos pela OpenAI. Em menos de três anos, a taxa de acerto nos diferentes níveis do exame CFA aumentou, em média, mais de 40%. Contudo, apesar desse aumento de qualidade, é improvável que esses modelos alcancem o patamar necessário para aprovação nos três níveis do exame. Portanto, eles não seriam capazes de adquirir o selo *CFA Charteholder* na condição atual. Apesar disso, é provável que a soberania humana nessa área não dure muito tempo, uma vez que já existem candidatos promissores para essa tarefa, como o GPT-4 e o Bloomberg GPT.

Por fim, é válido ressaltar que o estudo acima não exaure essa discussão. Pesquisadores interessados podem utilizar outros simulados, como o do próprio *CFA Institute* ou da Kaplan, por exemplo. Ademais, outros *LLMs* podem ser considerados, como o GPT-4 e o Bard, do Google. Mesmo que uma pesquisa extensiva acerca dessa classe de modelos já venha sendo realizada há anos, o universo de possíveis projetos ainda é bastante vasto e representa uma ótima oportunidade para pessoas interessadas e curiosas.

Atenciosamente,

Pedro Teles

AVANTGARDE
ASSET MANAGEMENT

Avantgarde Asset Management Gestão de Recursos Ltda.
Avenida das Nações Unidas, 12399 – 7º andar
Brooklin – São Paulo – SP
CEP: 04578-000
avantgardeam.com.br
+55 11 3400-7061

Avantgarde Multifatores FIA

Em março, o fundo teve retorno de -3,19% contra -3,07% do IBX. Em 2023, o retorno acumulado do fundo é de -7,02% contra -7,27% do *benchmark*. Desde o início, o fundo tem 51,25% de retorno total contra 2,81% do IBX. Na perspectiva de risco, a volatilidade - desvio padrão anualizado - das cotas diárias do fundo, desde o início, está em 22% contra 29% do IBX. O beta, por sua vez, é de 0,65 (modelo CAPM). O alfa de Jensen resultante é de 8,6% ao ano.

As posições que mais contribuíram para o retorno do fundo foram: Irani (RANI3), Mills (MILS3), BP P.L.C (B1PP34), Uber (U1BE34), General Eletric (GEOO34). Por outro lado, as posições que menos contribuíram foram: Metalúrgica Gerdau (GOAU4), Gerdau (GGBR4), Vale (VALE3), Petro Rio (PRIO3) e Bradespar (BRAP4).

O Avantgarde Multifatores FIA entra em abril totalmente alocado em ativos de risco, com aproximadamente 90% do patrimônio alocado em ações, 9% em recibos de empresas estrangeiras negociados no Brasil (BDR) e 1% de caixa.

Avantgarde Icatu Prev 70 FIC FIM

Em março, o fundo teve retorno de -0,89% contra 2,66% do IMA-B. Em 2023, o retorno acumulado do fundo é de -2,16% contra 3,98% do *benchmark*. Desde o início, o fundo tem 1,23% de retorno total contra 10,94% do IMA-B. Na perspectiva de risco, a volatilidade - desvio padrão anualizado - das cotas diárias do fundo, desde o início, está em 10,26% contra 5,42% do IMA-B. O fundo tem volatilidade alvo de 10% ao ano, bem próximo da volatilidade observada.

O fundo entra em abril com 69% de alocação na estratégia Multifatores de renda variável e com 31% em renda fixa.

Fatores de Risco

Os principais fatores de risco *long-only* tiveram os seguintes retornos no mês:

- Carry: -2,45%
- Low Vol: -3,24%
- Momentum: -1,71%
- Quality: 0,1%
- Value: -3,25%