INTELIGÊNCIA ARTIFICIAL · UNIFOR · 11 DE ABRIL DE 2026

Evaluating Frontier LLMs in Answering Game-Design Questions: A Preliminary Study

Por Adams Castro Filho, Maria Andréia Formico Rodriguês e Nabor C. Mendonça

Fonte: OpenAlex · Unifor · Publicação original

INOVAÇÃO	APLICABILIDADE	POTENCIAL ECONÔMICO	MATURIDADE
7/10	9/10	8/10	Média

Estudo da UNIFOR que avalia o desempenho de LLMs de ponta (GPT-5, Gemini 2.5 Pro, DeepSeek) na resolução de problemas de design de software para jogos. A pesquisa utiliza um protocolo 'LLM-as-a-judge' sobre 40 questões reais, concluindo que as IAs são precisas e úteis, mas falham em abranger restrições completas (completude) e oferecem soluções genéricas (baixa originalidade).

NÚCLEO DA ANÁLISE

Ideia de startup ou produto

Desenvolvimento de um 'Árbitro de Código IA' especializado em GameDev: uma SaaS que utiliza o protocolo LLM-as-a-judge para validar arquiteturas de jogos, garantindo que restrições de performance e casos extremos sejam atendidos antes do deploy.

Aplicações práticas

Assistentes de IA para programação de jogos, ferramentas de revisão automatizada de código focadas em arquitetura, treinamento de desenvolvedores juniores e geração de documentação técnica.

Potencial de mercado

Alto. A indústria de games busca avidamente ferramentas de IA para aumentar a produtividade. Há uma lacuna de mercado para ferramentas que garantam a 'Completude' técnica que LLMs genéricas ignoram.

FUNDAMENTO CIENTÍFICO

Problema abordado

A eficácia de Modelos de Linguagem de Grande Escala (LLMs) em tarefas complexas de arquitetura de software e design de jogos, especificamente na identificação de trade-offs, padrões de design e restrições de integração.

Metodologia

Avaliação comparativa utilizando 40 perguntas e respostas validadas pela comunidade (Q&A). Aplicação de um protocolo 'LLM-as-a-judge' onde modelos avaliam respostas anonimizadas baseadas em critérios de Acurácia, Completude, Utilidade e Originalidade.

Principais descobertas

Os modelos demonstram alta Acurácia e Utilidade, alinhando-se às respostas da comunidade. No entanto, apresentam baixa confiabilidade em Completude (omitindo restrições e casos extremos) e Originalidade (limitando-se a padrões familiares sem inovação).

LEITURA PARA GESTÃO PÚBLICA

Criação de editais de fomento para P&D em IA aplicada ao Entretenimento Digital e formação de hubs de inovação tecnológica focados em automação de desenvolvimento de software.

PARCERIA UNIVERSIDADE × EMPRESA

Parceria entre o grupo de pesquisa da UNIFOR e estúdios de jogos do Ceará (ex: Green.Game) para criar datasets específicos de design de jogos e treinar modelos ajustados (fine-tuning) que superem as limitações de completude identificadas.

OPORTUNIDADES DERIVADAS

3 direções estratégicas identificadas

PRODUTO CORPORATIVO · IMPACTO ALTO · SOFTWARE

Ferramenta de Validação de Arquitetura para Games

Plugin ou ferramenta SaaS para motores de jogo (Unity/Unreal) que audita automaticamente decisões de design contra o descritivo do projeto, prevendo falhas de completude identificadas no estudo.

STARTUP · IMPACTO MÉDIO · CIÊNCIA DE DADOS

QA de Código via LLM-as-a-Judge

Startup focada em serviços de auditoria de qualidade de software para indústria criativa, utilizando o protocolo de avaliação automatizada escalável desenvolvido pelos pesquisadores.

PARCERIA · IMPACTO ALTO · INTELIGÊNCIA ARTIFICIAL

Cooperação UNIFOR-Indústria de Games

Projeto de cooperação tecnológica para refinar modelos de IA com dados proprietários de estúdios locais, resolvendo o problema da falta de originalidade e insights profundos.

ABSTRACT ORIGINAL

We evaluate how three widely used frontier LLMs (GPT-5, Gemini 2.5 Pro, and DeepSeek) answer questions from game-development practice that are relevant to software design (e.g., design patterns, component boundaries, UI/interface structure, and interface/integration trade-offs). Using 40 posts sourced from a public Q&A forum with community-vetted baseline answers, we score each model's response on four criteria: Accuracy, Completeness, Usefulness, and Originality. To scale the assessment, we apply an LLM-as-a-judge protocol in which models rate anonymized answers and a rating is accepted only when at least two evaluators agree. Overall, the models are consistently strong on Accuracy and Usefulness, often aligning with the community baselines, but less reliable on Completeness, omitting constraints, edge cases, or long-term consequences. Originality is also weaker, where answers tend to restate familiar patterns rather than offer genuinely new insights. These results suggest that frontier LLMs can support game design reasoning and trade-off articulation, but they should be used with care when decisions demand full constraint coverage and robust validation. Our full prompt set, rubric, and analysis procedure, along with replication materials, are available in a companion repository at https://github.com/adamscastro/llm-game-designing.

MATÉRIA PARA LEIGOS

Para leigos: O potencial e os limites de IAs de ponta no design de jogos

O cenário atual

O desenvolvimento de jogos eletrônicos exige decisões complexas de software, envolvendo padrões de design, estruturas de interface e escolhas entre diferentes abordagens técnicas. Com o avanço da inteligência artificial, modelos de linguagem de grande escala (LLMs) estão sendo considerados para auxiliar nessas tarefas, mas sua capacidade real de responder a perguntas práticas dessa área ainda precisa ser melhor compreendida.

O que os pesquisadores fizeram

Pesquisadores da UNIFOR realizaram um estudo preliminar para avaliar como três modelos de IA de ponta (GPT-5, Gemini 2.5 Pro e DeepSeek) respondem a questões reais do dia a dia do desenvolvimento de jogos. A equipe selecionou 40 perguntas de um fórum público de perguntas e respostas que já possuíam soluções validadas pela comunidade. Eles então analisaram as respostas das IAs com base em quatro critérios: Acurácia (precisão), Completude, Utilidade e Originalidade.

Como funciona na prática

Para conseguir avaliar as respostas em escala, os pesquisadores utilizaram um protocolo onde uma IA atua como juiz. Nesse sistema, as respostas das máquinas eram enviadas de forma anônima para avaliação. Uma nota só era aceita se houvesse concordância entre pelo menos dois avaliadores virtuais, garantindo uma validação mais rigorosa das respostas geradas pelas IAs testadas.

Resultados e evidência

O estudo mostrou que os modelos são consistentemente fortes em Acurácia e Utilidade, muitas vezes alinhando-se bem com as respostas consideradas corretas pela comunidade de desenvolvedores. Por outro lado, as IAs demonstraram ser menos confiáveis na Completude. Com frequência, elas omitiram restrições importantes, casos extremos (edge cases) ou consequências a longo prazo das decisões de design. A Originalidade também foi um ponto fraco, com as tendendo a repetir padrões familiares em vez de oferecer insights genuinamente novos.

Implicações práticas

Os resultados sugerem que as IAs de ponta podem funcionar como ferramentas de apoio para o raciocínio sobre design de jogos e para explicar as vantagens e desvantagens (trade-offs) de diferentes opções. No entanto, a recomendação é que sejam usadas com cautela. Quando as decisões exigem uma cobertura total de restrições e uma validação robusta, a intervenção humana ainda é indispensável para garantir que nada importante tenha sido esquecido.

Limitações e próximos passos

O estudo é classificado como preliminar e identificou limitações claras na capacidade das IAs de fornecerem respostas completas e originais. O paper não detalha planos futuros específicos dos autores, mas garante que todos os materiais, como os conjuntos de comandos (prompts), critérios de avaliação e procedimentos de análise, estão disponíveis em um repositório público para que outros pesquisadores possam replicar o estudo.

QUEM SÃO OS PESQUISADORES

Quem são os pesquisadores

A pesquisa foi conduzida por Adams Castro Filho, Maria Andréia Formico Rodriguês e Nabor C. Mendonça, todos afiliados à UNIFOR (Universidade de Fortaleza). O paper não detalha a formação acadêmica específica, trajetória profissional ou outros vínculos anteriores dos autores além de sua participação neste estudo sobre a avaliação de modelos de linguagem no contexto de design de jogos.

PALAVRAS-CHAVE

Frontier · Replication (statistics) · Constraint (computer-aided design) · Protocol (science) · Component (thermodynamics) · Scale (ratio) · Baseline (sea) · Research design