

Investigando a Integração Estrutural de LLMs em Ecossistemas de Software: Um Estudo com Modelagem SSN

Francisco Victor da S. Pinheiro
Universidade Federal do Ceará
Quixadá, Ceará, Brazil
victorpinheiro@ufc.br

Emanuel F. Coutinho
Universidade Federal do Ceará
Quixadá, Ceará, Brazil
emanuel.coutinho@ufc.br

Marcelo M. da Silva
Universidade Federal do Ceará
Quixadá, Ceará, Brazil
martins2016eng@gmail.com

Sidarta A. Lobo de Carvalho
Universidade Federal do Ceará
Quixadá, Ceará, Brazil
sidarta@ufc.br

Rossana M. C. Andrade
Universidade Federal do Ceará
Fortaleza, Ceará, Brazil
rossana@ufc.br

RESUMO

Ecossistemas de software (ECOS) são compostos por atores interconectados que colaboram na produção de software. A notação *Software Supply Network* (SSN) permite modelar estruturalmente esses ecossistemas, destacando as relações entre plataformas, ferramentas e agentes envolvidos. Com a crescente adoção de Modelos de Linguagem em Larga Escala (LLMs), como GitHub, Copilot e ChatGPT, novas dinâmicas emergem no desenvolvimento de código aberto, ainda pouco compreendidas de uma perspectiva estrutural. Este trabalho propõe uma análise preliminar do papel dos LLMs como agentes em quatro ecossistemas distintos — LangChain, VSCode, Transformers e Pandas — por meio da modelagem SSN e análise automatizada de Pull Requests. Os resultados revelam padrões contrastantes de centralidade dos LLMs, destacando seu potencial como elementos de reestruturação em redes de produção de software. A abordagem oferece uma nova lente para entender a inserção da IA em ECOS e levanta hipóteses relevantes para investigações futuras.

PALAVRAS-CHAVE

Ecossistemas de Software, Modelos de Linguagem de Grande Escala, Rede de Fornecimento de Software.

1 Introdução

Com as constantes mudanças na indústria e a criação de novos paradigmas, a área de Engenharia de Software (ES) tem progredido em pesquisa teórica e aplicada, buscando lidar com uma gama de novos conhecimentos do mercado atual [18]. Dentre esses novos conhecimentos está o conceito de Ecossistema de Software (ECOS) que têm crescido rapidamente, tanto em número de casos reais como em volume de publicações, o que tem tornado o tema mais maduro nos últimos anos [9].

Tradicionalmente, um ECOS é uma metáfora de ES que foi aplicada para a compreensão da dinâmica da rede de fornecimento de software centrada em plataformas de software [5, 6, 10]. Boucharras et al. (2009) propuseram o uso da notação SSN - Rede de Produção de Software, uma notação já utilizada pela comunidade de ECOS para a modelagem. O SSN é uma série de software, hardware e organizações de serviços interligados que cooperam para atender às demandas do mercado. A notação foi estendida por Costa et al. (2013), adicionando mais componentes e fornecendo uma maior

explicação sobre cada um deles, como também para o seu uso de modo geral.

Para a compreensão do comportamento do ECOS é importante perceber a estruturação e os papéis dos atores e seus relacionamentos, assim como a plataforma central onde os mesmos atuam [4, 17]. Contudo, tanto a definição de ECOS quanto suas aplicações em diversos domínios são emergentes nos campos da pesquisa, ensino e indústria. Vários problemas e limitações são encontrados na literatura relacionados a modelagem e modelos SSN de ECOS [1, 14, 15]

Nos últimos anos, Modelos de Linguagem de Grande Escala (LLM), como o GPT e o *Copilot*, têm emergido como atores relevantes nos processos de desenvolvimento de software [8]. Esses modelos não apenas auxiliam na geração de código, mas também influenciam arquiteturas, fluxos de trabalho e dinâmicas de colaboração em diferentes ecossistemas de software *open-source* [12, 19]. Apesar do crescente interesse em avaliar o impacto técnico e ético dos LLMs, poucos estudos têm investigado seu papel como parte integrante das redes de produção de software [7, 11].

Dessa forma, este trabalho busca preencher essa lacuna, aplicando a notação Software SSN para modelar e compreender como os LLMs interagem e influenciam diferentes ecossistemas de software, sejam eles centrados na integração de LLM, no desenvolvimento de modelos, ou em projetos tradicionais. O objetivo deste trabalho é analisar o impacto dos LLMs no fluxo de contribuições em diferentes ecossistemas de software *open-source*, buscando compreender de que forma esses modelos atuam como agentes ativos ou ferramentas auxiliares no processo de desenvolvimento.

Para isso, foram investigados quatro ECOS com características distintas: *LangChain*, voltado para a integração de LLM em aplicações; *VSCode*, uma IDE que incorpora assistentes de código como o *Copilot*; *Transformers*, da Hugging Face, dedicado ao desenvolvimento de LLM; e *Pandas*, um projeto tradicional de ciência de dados. A diversidade desses projetos permite uma análise comparativa sobre o papel dos LLMs em diferentes contextos e estágios de maturidade tecnológica dentro do ecossistema *open-source*.

Este trabalho é guiado por três questões de pesquisa principais: QP1: De que formas os LLMs são referenciados, utilizados ou integrados em diferentes ecossistemas de software *open-source*?; QP2: Qual o papel estrutural e o grau de centralidade dos LLMs nas redes de produção dos diferentes ecossistemas estudados, segundo a modelagem SSN?; e QP3: Quais padrões e diferenças emergem na

comparação entre ecossistemas centrados em LLM, ecossistemas com integração parcial e projetos tradicionais quanto ao impacto dos LLMs no fluxo de contribuições e na dinâmica colaborativa?. Essas questões orientam toda a análise apresentada neste estudo, permitindo compreender não apenas o uso dos LLMs, mas também sua influência estrutural e comparativa nos diferentes contextos analisados.

Este trabalho está dividido nas seguintes seções além desta introdução: na Seção 2 alguns trabalhos relacionados são discutidos; a Seção 3 apresenta a metodologia empregada no trabalho; a Seção 4 apresenta os resultados encontrados nos repositórios dos projetos; a Seção 5 apresenta uma modelagem SSN preliminar para cada ECOS analisado; a Seção 6 apresenta algumas discussões e análises dos resultados assim como desafios e oportunidades de pesquisa; e por fim, as conclusões e trabalhos futuros são apresentados na Seção 7.

2 Trabalhos Relacionados

Na literatura, diversas iniciativas exploram o uso de LLM no contexto do desenvolvimento de software *open-source*, abordando desde aspectos relacionados à produtividade de desenvolvedores até o impacto nos fluxos de colaboração. No entanto, ainda existe uma lacuna significativa quanto à compreensão do papel dos LLMs como atores nos ECOS, especialmente em relação à modelagem de suas interações e influências nas redes de produção de software. Diante disso, a seguir são apresentados alguns trabalhos relacionados que investigam o impacto dos LLMs em diferentes dimensões da engenharia de software e que contribuem para o embasamento deste estudo.

Yeverechyahu et al. (2024) analisaram o impacto do *GitHub Copilot* em projetos *open-source*, observando que seu uso aumentou principalmente as contribuições iterativas, relacionadas à manutenção de código, enquanto as contribuições de origem (novas funcionalidades) tiveram crescimento menor. Esse resultado evidencia que LLMs favorecem tarefas de refinamento, mas não impulsionam igualmente a inovação, o que se alinha à proposta deste estudo ao reforçar que os papéis dos LLMs variam nos fluxos de contribuição. No entanto, enquanto o trabalho de Yeverechyahu et al. (2024) foca em um único projeto e tipo de tarefa, este estudo amplia a análise para diferentes ecossistemas, modelando como os LLMs interagem em cada contexto.

Song et al. (2024) investigaram o impacto do *GitHub Copilot* na produtividade e participação em projetos *open-source*, identificando um aumento de 6,5% na produtividade em nível de projeto e 5,5% na produtividade individual, além de 5,4% na participação. No entanto, observaram também um aumento de 41,6% no tempo de integração, sugerindo maiores custos de coordenação. O estudo diferencia os efeitos entre desenvolvedores centrais e periféricos, destacando que os centrais se beneficiam mais do *Copilot*. Em contraste, o presente trabalho busca modelar o papel dos LLMs em diferentes ecossistemas, indo além de métricas quantitativas para compreender como os LLMs influenciam estruturalmente os fluxos de contribuição, usando a notação SSN.

Zheng et al. (2025) realizam uma revisão sistemática sobre a integração e eficácia dos LLMs em tarefas de engenharia de software, categorizando 123 artigos desde 2022 em sete tipos principais de tarefas, como geração de código, documentação e testes. Os autores

buscam responder se LLMs são eficazes nessas tarefas e quais são suas principais aplicações e limitações. Embora forneça uma visão ampla do estado da arte, o foco está em avaliar capacidades técnicas dos LLMs em tarefas específicas. Já o presente estudo complementa essa perspectiva, analisando como essas capacidades impactam os fluxos de contribuição em ecossistemas reais, modelando essas interações com a notação SSN para entender o papel sistêmico dos LLMs.

De modo geral, os trabalhos relacionados exploram diferentes dimensões do impacto dos LLMs na engenharia de software, com foco predominante em métricas quantitativas (como produtividade, participação e tipos de contribuições) ou em revisões sistemáticas sobre as aplicações desses modelos em tarefas específicas, como geração de código ou testes. No entanto, nenhum dos estudos investiga o papel dos LLMs como atores sistêmicos nas redes de produção de software, limitando-se a análises locais ou a perspectivas centradas em ferramentas específicas, como o *Copilot*. O estudo proposto se diferencia por modelar essas interações no nível estrutural, utilizando a notação SSN para compreender como os LLMs influenciam o fluxo de contribuições em diferentes ecossistemas *open-source*. A Tabela 1 apresenta uma visão geral comparativa entre os trabalhos relacionados e o proposto.

3 Metodologia

A metodologia deste trabalho consiste em uma abordagem empírica e estruturada para investigar o papel dos Modelos de LLMs em diferentes ecossistemas de software *open-source*. Inicialmente, foram selecionados quatro ecossistemas com perfis distintos quanto à integração de LLMs. Em seguida, foi realizada a coleta automatizada de *Pull Requests* utilizando a API do *GitHub*, seguida da classificação dos PRs em categorias de uso dos LLMs. Posteriormente, os dados coletados foram analisados quantitativamente por meio de gráficos e tabelas, e os papéis dos LLMs foram modelados utilizando a notação SSN. Por fim, os resultados foram compilados e analisados de forma comparativa entre os diferentes ecossistemas. A Figura 1 ilustra o fluxo metodológico seguido neste estudo.

3.1 Seleção dos ecossistemas de software

O primeiro passo consistiu na escolha de quatro ecossistemas de software *open-source* com perfis contrastantes em relação à presença e integração de Modelos de LLMs. Foram selecionados o *LangChain*, por ser um projeto intrinsecamente centrado em LLMs e integração de agentes inteligentes; o *VSCode*, por incorporar de forma relevante assistentes baseados em LLMs como o *GitHub Copilot*, mas manter uma proposta geral de ambiente de desenvolvimento extensível; o *Transformers*, da Hugging Face, como exemplo de repositório dedicado à pesquisa e implementação de LLMs em nível de infraestrutura e modelo; e o *Pandas*, um projeto tradicional de ciência de dados, sem centralidade em LLMs, escolhido para fins de contraste e controle metodológico.

3.2 Coleta automatizada de pull requests

O segundo passo envolveu o desenvolvimento e execução de *scripts* para a coleta automatizada de *Pull Requests* dos quatro repositórios selecionados, utilizando a API pública do *GitHub*. Para maximizar a abrangência da busca, os *scripts* filtraram os PRs utilizando um

Tabela 1: Comparação entre os trabalhos relacionados e o estudo proposto

Trabalho	Objetivo	Abordagem Analítica	Escopo do Ecossistema	Contribuições Únicas do Estudo Proposto
Yevechayahu et al. (2024)	Avaliar o impacto do <i>Copilot</i> nas contribuições (manutenção vs. origem)	Quantitativa (experimento natural)	Projeto único (<i>GitHub open-source</i>)	Modelagem estrutural do papel dos <i>LLMs</i> em diferentes ecossistemas
Song et al. (2024)	Avaliar o efeito do <i>Copilot</i> na produtividade e participação	Quantitativa (controle sintético)	Projeto único (<i>GitHub open-source</i>)	Modelagem comparativa do papel dos <i>LLMs</i> em diferentes ECOSs
Zheng et al. (2025)	Revisão sistemática das aplicações de <i>LLMs</i> na engenharia de software	Revisão de literatura (123 artigos)	Diversas tarefas (código, testes, documentação)	Modelagem estrutural (SSN) dos <i>LLMs</i> como atores sistêmicos
Estudo Proposto	Modelar o impacto dos <i>LLMs</i> no fluxo de contribuições em ecossistemas <i>open-source</i>	Qualitativa + estrutural (modelagem SSN)	Múltiplos ecossistemas (<i>LangChain, VSCode, Transformers, Pandas</i>)	Modelagem dos <i>LLMs</i> como atores sistêmicos, comparando papéis em diferentes tipos de ECOSs

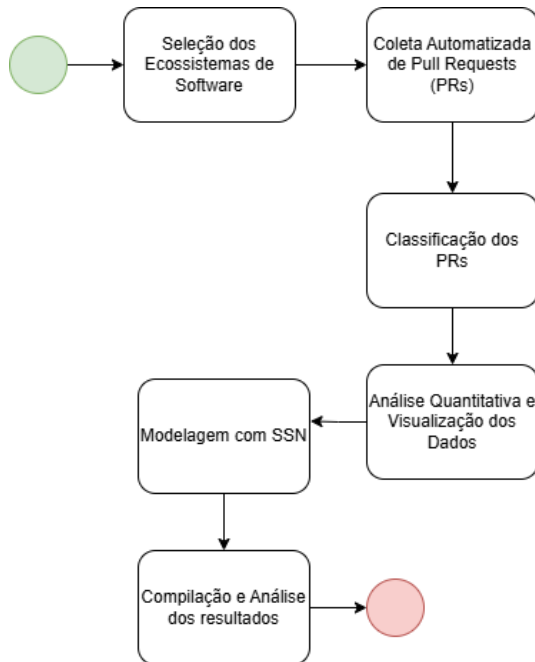


Figura 1: Fluxo metodológico realizado neste trabalho.

conjunto amplo de termos relacionados a *LLMs*, incluindo “*Copilot*”, “*gpt*”, “*chatgpt*”, “*llm*”, “*OpenAI*”, “*language model*”, entre outros. Foram coletadas informações como título, corpo do PR, data de criação, estado (aberto ou fechado) e URL.

3.3 Classificação dos *pull requests*

Após a coleta dos dados, os PRs foram submetidos a um processo automatizado de classificação, empregando critérios baseados em

palavras-chave e padrões textuais presentes nos títulos e descrições dos PRs. Os PRs foram classificados em três categorias principais: “*LLM diretamente usado*” (quando o PR envolvia implementação, integração ou uso explícito de *LLMs* no código), “*LLM citado/conceitual*” (quando o PR mencionava *LLMs* em discussões, documentação, ou como referência de contexto), e “*Não relacionado*” (quando não havia relação perceptível com *LLMs*). Esse processo visou identificar tanto o uso prático e efetivo dos modelos quanto sua presença na discussão e planejamento técnico dos projetos.

3.4 Análise quantitativa e visualização dos dados

Com a classificação realizada, procedeu-se uma análise quantitativa dos PRs, envolvendo cálculos de frequências, distribuição temporal, percentual de cada categoria e cruzamento com atributos como estado do PR e ano de submissão. Foram desenvolvidos gráficos de barras, linhas e tabelas para representar visualmente a intensidade e o padrão de menção e uso de *LLMs* em cada projeto. Essa etapa permitiu identificar tendências, variações entre projetos e potenciais correlações entre o contexto do ecossistema e o grau de integração dos modelos de linguagem, subsidiando análises comparativas mais aprofundadas.

3.5 Modelagem com SSN

O quinta passo consistiu na modelagem dos quatro ecossistemas selecionados utilizando a notação SSN na ferramenta *ECOS Modeling* [13, 15]. Foram construídos diagramas conceituais representando os principais atores, suas relações e fluxos de contribuição, com ênfase nos pontos em que *LLMs* atuam como elementos estruturais, agentes periféricos ou ferramentas auxiliares.

3.6 Compilação e análise dos resultados

Por fim, os dados quantitativos, as classificações e os modelos SSN foram integrados em uma análise comparativa e interpretativa. Esta etapa envolveu a síntese dos achados, destacando padrões recorrentes e diferenças significativas entre os ecossistemas estudados. A análise crítica foi guiada pelas questões de pesquisa propostas e

buscou apontar limitações, oportunidades e caminhos para investigações futuras na área.

4 Resultados

A investigação conduzida sobre os quatro ecossistemas de software analisados — *LangChain*, *VSCoDe*, *Transformers* e *Pandas* — evidencia dinâmicas contrastantes quanto à inserção e ao papel central dos Modelos *LLMs* no desenvolvimento *open-source*. Os resultados a seguir detalham como esses ecossistemas têm assimilado os *LLMs* em suas estruturas, destacando diferenças significativas em termos de adoção, relevância e impacto desses agentes inteligentes nos fluxos colaborativos e produtivos de cada projeto. A Figura 2 apresenta um panorama geral dos resultados encontrados para cada projeto.

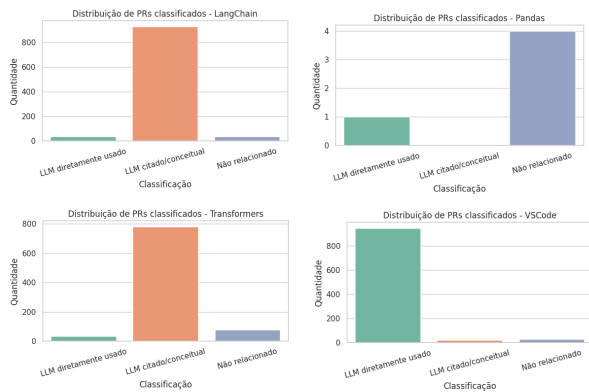


Figura 2: Distribuição de PRs por projeto.

No ecossistema *LangChain*, há um predomínio expressivo de PRs relacionados a *LLMs*, representando cerca de 96% do total analisado. Desses, a imensa maioria refere-se à categoria “LLM citado/conceitual”, demonstrando um ambiente no qual discussões sobre integração, uso e referência a tecnologias *LLM* permeiam o cotidiano do desenvolvimento. PRs classificados como “LLM diretamente usado” correspondem a uma fração menor, refletindo integrações práticas e experimentais. O crescimento temporal das contribuições relacionadas a *LLMs* acompanha a ascensão do projeto como referência em aplicações e infraestrutura para IA generativa. Na modelagem SSN, os *LLMs* aparecem como elementos centrais tanto técnica quanto conceitualmente, aproximando a inovação acadêmica e a adoção produtiva.

O cenário no *VSCoDe* apresenta um perfil ainda mais acentuado: mais de 97% dos PRs analisados possuem relação direta com o uso de *LLMs*, com amplo destaque para a categoria “LLM diretamente usado”. Esse padrão evidencia uma estratégia deliberada de integrar modelos de linguagem como funcionalidades nativas do editor, principalmente via *Copilot* e recursos correlatos. A evolução temporal mostra saltos marcantes em anos-chave, indicando ondas de adoção guiadas por grandes lançamentos ou iniciativas institucionais. No modelo SSN, os *LLMs* ocupam a posição de provedor central, articulando-se com *plugins*, extensões e outras ferramentas do ecossistema, e promovendo a consolidação de *VSCoDe* como ambiente de inovação em IA aplicada ao desenvolvimento.

No *Transformers*, observa-se também uma forte incidência de PRs relacionados a *LLMs* (cerca de 91%). Aqui, o predomínio é da categoria “LLM citado/conceitual”, evidenciando o papel do projeto como catalisador de discussões, experimentações e avanços em modelos de linguagem. Os PRs classificados como “LLM diretamente usado” são menos frequentes, mas têm aumentado nos anos recentes, acompanhando a rápida evolução e diversificação dos modelos suportados. A modelagem SSN destaca o *Transformers* como um *hub* de referência e conexão entre diversas iniciativas, aproximando comunidades acadêmicas, empresariais e de desenvolvedores.

Em contraponto, o *Pandas* exibe uma baixa incidência de PRs relacionados a *LLMs*: apenas cerca de 20% dos PRs analisados possuem alguma menção, sendo a grande maioria classificada como “não relacionado”. Isso evidencia que, até o momento, o projeto permanece focado em seu objetivo principal — manipulação e análise de dados —, adotando tecnologias *LLM* apenas de maneira pontual ou experimental. Na modelagem SSN, os *LLMs* aparecem como atores periféricos, sem papel estrutural relevante no ecossistema do *Pandas*. Com base na análise comparativa dos quatro ecossistemas estudados, observa-se uma clara divisão entre projetos nos quais os Modelos *LLMs* desempenham um papel central e projetos onde sua presença é periférica ou residual. A Figura 3 ilustra os resultados encontrados para cada um dos repositórios em relação aos anos de 2019-2025.

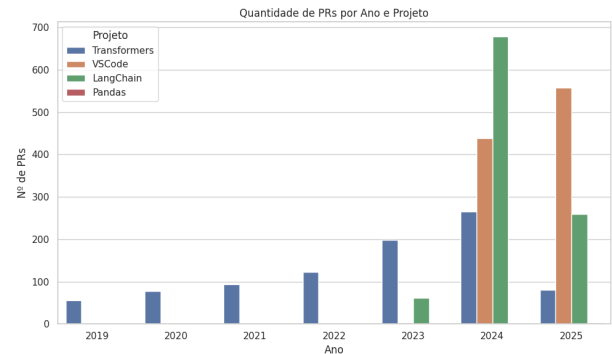


Figura 3: Quantidade de PRs por projeto nos anos de 2019-2025.

LangChain, *VSCoDe* e *Transformers* demonstram alto grau de integração e centralidade dos *LLMs*, embora cada um apresente nuances distintas. Enquanto o *VSCoDe* prioriza a implementação prática e funcional, refletida pelo predomínio de *Pull Requests* com uso direto de *LLMs*, *LangChain* e *Transformers* destacam-se pela forte ênfase em discussões técnicas, arquiteturas e conceituais sobre a aplicação e evolução desses modelos. Isso sugere que, nesses projetos, os *LLMs* atuam tanto como insumos técnicos fundamentais quanto como catalisadores de inovação e debate, permeando as decisões estratégicas e operacionais dos ecossistemas.

A Tabela 2 apresenta um comparativo dos resultados encontrados.

Por outro lado, o *Pandas* apresenta um cenário contrastante, com raríssimas menções ou integrações de *LLMs*, evidenciando uma resistência ou cautela na adoção dessas tecnologias disruptivas em

Tabela 2: Tabela comparativa da classificação dos PRs analisados nos projetos

Projeto	LLM citado/conceitual (%)	LLM diretamente usado (%)	Não relacionado (%)	Total de PRs analisados
<i>LangChain</i>	92.70	3.60	3.70	1000
<i>Pandas</i>	0.00	20.00	80.00	5
<i>Transformers</i>	87.56	3.92	8.52	892
<i>VSCode</i>	2.10	94.90	3.00	1000

projetos de natureza tradicional. Essa disparidade ressalta que o impacto estrutural dos *LLMs* não é homogêneo no universo *open-source*: sua inserção depende fortemente da missão, do público-alvo e do estágio de maturidade de cada ecossistema. Assim, a análise revela que, enquanto determinados projetos se consolidam como protagonistas da inovação baseada em *LLMs*, outros mantêm foco em sua vocação original, mostrando que a transformação impulsivada por IA é seletiva, incremental e fortemente condicionada ao contexto e à trajetória do software analisado.

5 Modelagem SSN

Para compreender o papel estrutural dos *LLMs* em diferentes ecossistemas de software, propomos uma modelagem SSN específica para cada projeto investigado. O modelo destaca as relações entre desenvolvedores humanos, plataformas, *LLMs* e organizações envolvidas, evidenciando fluxos de fornecimento, integração e consumo de tecnologia. A aplicação da notação SSN permitiu mapear, de forma comparativa, o grau de centralidade e o tipo de inserção dos *LLMs* em diferentes ecossistemas de software.

Em projetos como *VSCode* e *LangChain*, observa-se o *LLM* como ator central e estruturante, redefinindo relações e fluxos produtivos. Em *Transformers*, os *LLMs* funcionam como insumos críticos de inovação, enquanto em *Pandas* sua presença é marginal. Essa modelagem contribui para uma visão sistêmica dos impactos das IAs generativas no cenário *open-source*, indicando possíveis tendências de evolução das redes de produção de software. A Figura 4 apresenta os modelos preliminares para cada um dos ECOS analisados elencando o papel dos *LLMs* nos ECOSs.

A modelagem SSN dos quatro ecossistemas evidencia diferentes graus de centralidade dos *LLMs* e suas funções dentro das redes de produção de software. Em *VSCode*, os *LLMs*, exemplificados pelo *GitHub Copilot*, atuam como fornecedores diretos de sugestões de código, sendo integrados ao fluxo produtivo via plataforma e avaliados posteriormente por revisores humanos. O papel das grandes organizações, como *Microsoft* e *OpenAI*, destaca a influência institucional no avanço da Inteligência Artificial (IA) aplicada ao desenvolvimento. Já em *LangChain*, a centralidade está na própria plataforma, que integra múltiplos *LLMs* de diferentes provedores (*OpenAI*, *Anthropic*, *Gemini*) e os disponibiliza a desenvolvedores, atuando como intermediário e facilitador da experimentação e inovação baseada em IA.

Por outro lado, no ecossistema *Transformers*, os *LLMs* funcionam principalmente como acessórios de produção e insumos tecnológicos, desenvolvidos de maneira colaborativa por uma rede aberta de indivíduos e organizações. A biblioteca *Transformers* age como elo entre esses modelos e os usuários finais, reforçando o papel da comunidade e dos parceiros institucionais como agregadores e disseminadores. Em contraste, o *Pandas* mantém uma estrutura

clássica de produção, centrada nos desenvolvedores humanos e em ferramentas tradicionais de apoio, sem participação expressiva dos *LLMs* em sua dinâmica, o que reforça sua posição periférica frente às tendências de integração de IA generativa no ecossistema *open-source*.

6 Discussão e Análise dos Resultados

Nesta seção, apresentamos uma discussão aprofundada dos resultados obtidos a partir da análise dos quatro ecossistemas de software selecionados, com foco na presença, centralidade e impacto dos Modelos de *LLMs* nos fluxos de desenvolvimento *open-source*. A discussão é guiada pelas três questões de pesquisa formuladas, abordando desde as diferentes formas de integração e referência aos *LLMs*, até o papel estrutural desses modelos segundo a modelagem SSN e os efeitos observados na dinâmica colaborativa dos projetos. Ao responder cada uma das questões propostas, buscamos evidenciar padrões, contrastes e implicações relevantes, oferecendo uma visão abrangente sobre o fenômeno da adoção de *LLMs* e seu potencial de transformação nos ecossistemas de software analisados.

QP1 – De que formas os *LLMs* são referenciados, utilizados ou integrados em diferentes ecossistemas de software *open-source*? A análise dos PRs evidencia que a presença dos *LLMs* nos ecossistemas *open-source* ocorre em diferentes graus e formatos. Nos ecossistemas *LangChain*, *VSCode* e *Transformers*, observa-se que a vasta maioria dos PRs está relacionada a *LLMs*, seja por meio de integração direta, implementação de funcionalidades baseadas em *LLMs* ou discussão conceitual sobre modelos e suas aplicações. Destaca-se que no *VSCode* prevalecem PRs de uso direto – indicando aplicação prática de *LLMs* como *Copilot*, enquanto *LangChain* e *Transformers* apresentam também elevado número de PRs de discussão conceitual, sinalizando ambientes de intensa inovação e debate técnico. Em contraste, no ecossistema *Pandas*, a presença dos *LLMs* é residual, aparecendo apenas em menções pontuais, sem reflexo expressivo no fluxo de desenvolvimento, o que ilustra uma adoção muito mais cautelosa.

QP2 – Qual o grau de centralidade e o papel estrutural dos *LLMs* nas redes de produção dos diferentes ecossistemas estudados, segundo a modelagem SSN? A modelagem SSN dos ecossistemas permite identificar que, em *VSCode*, os *LLMs* desempenham papel estrutural central, sendo frequentemente posicionados como provedores diretos de funcionalidades para usuários finais, intermediários e agregadores (ex: plugins, extensões, assistentes). *LangChain* e *Transformers* também posicionam os *LLMs* de maneira central, mas com destaque para sua atuação como insumo técnico e referência para discussões e experimentações arquiteturais. Nessas redes, os *LLMs* catalisam a emergência de novos fluxos de dependência e de colaboração, influenciando diretamente a dinâmica dos demais atores. Por outro lado, em *Pandas*, os *LLMs* não ocupam

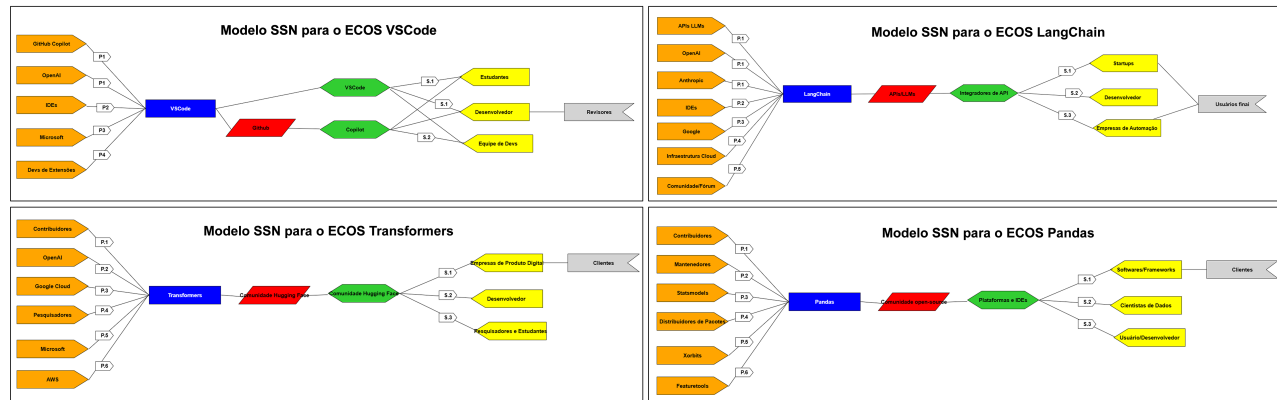


Figura 4: Modelos SSN preliminares para os ECOSs *VSCode*, *LangChain*, *Transformers* e *Pandas*.

posição central, nem mesmo periférica relevante, mantendo-se à margem da rede produtiva, o que indica que seu papel estrutural ainda é mínimo ou inexistente neste ecossistema.

QP3 – Como a presença e o tipo de uso dos LLMs influenciam o fluxo de contribuições (PRs) e a dinâmica colaborativa nos projetos *open-source* analisados? Os resultados sugerem que, em ecossistemas nos quais os LLMs são centrais – como *LangChain*, *VSCode* e *Transformers* –, há uma clara intensificação do fluxo de PRs relacionados à integração, aprimoramento e discussão sobre LLMs. No *VSCode*, a predominância do uso direto resulta em ciclos rápidos de adoção e lançamento de novas funcionalidades, além de incentivar a colaboração em torno de melhorias práticas. Em *LangChain* e *Transformers*, a natureza das contribuições reflete não apenas implementação, mas também o avanço conceitual e experimental dos LLMs, promovendo debates contínuos e colaboração em alto nível técnico. Já no *Pandas*, a ausência ou baixa incidência de PRs ligados a LLMs sugere que a dinâmica colaborativa permanece orientada a aspectos tradicionais da biblioteca, como manutenção e evolução incremental, sem influência substancial dos modelos de linguagem. Assim, o impacto dos LLMs sobre o fluxo colaborativo é diretamente proporcional ao seu grau de centralidade e adoção em cada ecossistema.

6.1 Desafios e oportunidades de pesquisa

Apesar dos avanços evidenciados na integração de LLMs em diferentes ecossistemas de software, diversos desafios persistem e abrem espaço para novas investigações. Entre os principais desafios identificados está a dificuldade de caracterizar, com precisão, os limites entre uso conceitual e uso direto dos LLMs, especialmente em projetos com documentação fragmentada ou contribuições heterogêneas. Além disso, a rápida evolução dos modelos e APIs exige abordagens analíticas continuamente atualizadas para acompanhar mudanças nos fluxos de trabalho e na arquitetura dos ecossistemas.

O fenômeno da centralização de decisões técnicas em torno de poucos atores-chave, frequentemente ligados a grandes organizações, também levanta questões sobre governança, diversidade e sustentabilidade dos projetos. Oportunidades de pesquisa incluem a análise longitudinal do impacto dos LLMs em métricas de qualidade, colaboração e inovação, o estudo da difusão de práticas

orientadas por IA em comunidades menores, e o desenvolvimento de frameworks de modelagem capazes de capturar nuances emergentes na interação entre humanos e agentes de IA.

7 Conclusão e Trabalhos Futuros

Este trabalho investigou o papel dos LLMs em quatro ecossistemas de software *open-source*, integrando análise automatizada de *Pull Requests* e modelagem SSN para mapear sua centralidade e impacto estrutural nas redes de produção de software. Os resultados evidenciam que a presença e o grau de integração dos LLMs variam substancialmente entre os projetos analisados, refletindo diferentes estratégias de adoção, maturidade tecnológica e vocação dos ecossistemas. Enquanto alguns apresentam os LLMs como elementos centrais e estruturantes, outros mantêm uma relação mais periférica e cautelosa. Diferente dos trabalhos anteriores, que focam em métricas quantitativas ou estudos de caso isolados, este estudo avança ao usar a modelagem SSN para revelar o papel sistêmico dos LLMs nas redes de software, oferecendo uma visão inédita sobre sua posição e influência na dinâmica colaborativa dos ecossistemas.

O estudo possui limitações, como a possibilidade de não identificar todos os usos de LLMs devido à busca por palavras-chave e potenciais erros na classificação automática. A escolha de apenas quatro ecossistemas restringe a generalização dos resultados e a modelagem SSN pode envolver interpretações subjetivas dos pesquisadores. Reconhecer essas limitações é importante para interpretar os resultados corretamente.

Como trabalhos futuros, propomos ampliar o estudo para mais projetos, incluir métricas qualitativas sobre o impacto para os desenvolvedores e aprofundar a modelagem das redes de influência entre humanos e IA. Também é importante investigar como políticas de governança e documentação influenciam a adoção responsável dos LLMs.

DISPONIBILIDADE DE ARTEFATO

Os dados coletados e usados neste trabalho estão disponíveis publicamente em: <https://doi.org/10.5281/zenodo.16687001>

REFERÊNCIAS

- [1] L. M. Ambrósio, P. Marques, J. M. N. David, R. Braga, M. A. Ribeiro Dantas, V. Ströele, and F. Campos. 2019. An Approach to Support Data Integration in a Scientific Software Ecosystem Platform. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 39–44.
- [2] Vasilis Boucharas, Slinger Jansen, and Sjaak Brinkkemper. 2009. Formalizing Software Ecosystem Modeling (*IWOCE '09*).
- [3] Gabriella Costa, Felyppe Silva, Rodrigo Santos, Cláudia Werner, and Toacy Oliveira. 2013. From Applications to a Software Ecosystem Platform: An Exploratory Study. In *Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems*.
- [4] Carla Coutinho, Emanuel e Bezerra. 2020. A study on dynamic aspects variability in the SOLAR educational software ecosystem. *Journal of the Brazilian Computer Society* 26, 1 (2020), 1–19.
- [5] Emanuel F. Coutinho, Italo Santos, Leonardo O. Moreira, and Carla I. M. Bezerra. 2019. A Report on the Teaching of Software Ecosystems in Software Engineering Discipline. In *Proceedings of the XXXIII Brazilian Symposium on Software Engineering (SBES 2019)*.
- [6] Emanuel Ferreira Coutinho, Davi Viana, and Rodrigo Pereira dos Santos. 2017. An Exploratory Study on the Need for Modeling Software Ecosystems: The Case of SOLAR SECO. In *9th International Workshop on Modelling in Software Engineering (MISE) (MISE '17)*.
- [7] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence. [arXiv:2401.14196 \[cs.SE\]](https://arxiv.org/abs/2401.14196) <https://arxiv.org/abs/2401.14196>
- [8] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2024. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* 33, 8 (2024), 1–79.
- [9] Slinger Jansen. 2020. A focus area maturity model for software ecosystem governance. *Information and Software Technology* 118 (2020), 106219.
- [10] Slinger Jansen, Eko Handoyo, and Carina Alves. 2015. Scientists' needs in modelling software ecosystems. In *Proceedings of the 2015 European Conference on Software Architecture Workshops*. 1–6.
- [11] Yacine Majdoub and Eya Ben Charrada. 2024. Debugging with Open-Source Large Language Models: An Evaluation. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24)*. ACM, 510–516. <https://doi.org/10.1145/3674805.3690758>
- [12] Ipek Ozkaya. 2023. Application of large language models to software engineering tasks: Opportunities, risks, and implications. *IEEE Software* 40, 3 (2023), 4–8.
- [13] Francisco Pinheiro, Emanuel Coutinho, Ronier Lima, Maria Silva, Carla Bezerra, and Rossana Andrade. 2025. ECOS Modeling: A Modeling Tool, Repository for Models and Evolution Analysis of Software Ecosystems. In *Anais do XXI Simpósio Brasileiro de Sistemas de Informação (Recife/PE)*. SBC, Porto Alegre, RS, Brasil, 369–378. <https://doi.org/10.5753/sbsi.2025.246501>
- [14] Francisco Victor Pinheiro, Emanuel Coutinho, Maria Eriane Silva, and Carla Bezerra. 2024. A Systematic Mapping of Health, Quality, Evolution, Simulation and Modeling in Software Ecosystems. In *Proceedings of the 20th Brazilian Symposium on Information Systems (Juiz de Fora, Brazil) (SBSI '24)*. Association for Computing Machinery, New York, NY, USA, Article 26, 10 pages. <https://doi.org/10.1145/3658271.3658297>
- [15] Francisco Victor da S. Pinheiro, Emanuel Ferreira Coutinho, Italo Santos, and Carla I. M. Bezerra. 2022. A Tool for Supporting the Teaching and Modeling of Software Ecosystems Using SSN Notation. *Journal on Interactive Systems* 13, 1 (Sep. 2022), 192–204. <https://doi.org/10.5753/jis.2022.2602>
- [16] Fangchen Song, Ashish Agarwal, and Wen Wen. 2024. The impact of generative AI on collaborative open-source software development: Evidence from GitHub Copilot. *arXiv preprint arXiv:2410.02091* (2024).
- [17] Ivo van den Berk, Slinger Jansen, and Lützen Luinenburg. 2010. Software Ecosystems: A Software Ecosystem Strategy Assessment Model. In *Proceedings of the Fourth European Conference on Software Architecture: Companion Volume (Copenhagen, Denmark) (ECSA '10)*. Association for Computing Machinery, New York, NY, USA, 127–134. <https://doi.org/10.1145/1842752.1842781>
- [18] Jelle Wouters, JR Ritmeester, AW Carlsen, Slinger Jansen, and Krzysztof Wnuk. 2019. A SECO meta-model. In *International Conference on Software Business*. Springer, 31–45.
- [19] Doron Yeverechyahu, Raveesh Mayya, and Gal Oestreicher-Singer. 2024. The impact of large language models on open-source innovation: Evidence from GitHub Copilot. *arXiv preprint arXiv:2409.08379* (2024).
- [20] Zibin Zheng, Kaiwen Ning, Qingyuan Zhong, Jiachi Chen, Wenqing Chen, Lianghai Guo, Weicheng Wang, and Yanlin Wang. 2025. Towards an understanding of large language models in software engineering tasks. *Empirical Software Engineering* 30, 2 (2025), 50.