

Architecture Decisions in AI-based Systems Development: An Empirical Study

Beiqi Zhang^{1,2}, Tianyang Liu³, Peng Liang^{1,2*}, Chong Wang^{1,2*}, Mojtaba Shahin⁴, Jiaxin Yu^{1,2}

¹ School of Computer Science, Wuhan University, Wuhan, China

² Hubei LuoJia Laboratory, Wuhan, China

³ Department of Computer Science and Engineering, University of California San Diego, La Jolla, USA

⁴ School of Computing Technologies, RMIT University, Melbourne, Australia
{[liangp](#), [cwang](#)}@whu.edu.cn

背景

- 人工知能(AI)技術の急速な発展
 - AIベースのシステムは、様々な応用領域で広く利用されるようになった。
- ソフトウェアアーキテクチャ
 - 本質的な要件に対応した設計上の意思決定の組み合わせ。
 - 過去20年にわたり広く研究されてきた。
 - アーキテクチャの決定は、大規模で動的なソフトウェアシステムの設計開発、および進化をより良く制御することにつながる。
- AIベースのシステム開発におけるアーキテクチャの決定に関する研究はほとんど行われていない。

先行研究

- Franchらの研究[8]

- AIベースのシステム開発における文献調査
- アーキテクチャの決定に関する研究と実践の現状をさらに調査する必要性を強調

- AIベースのシステムは従来のソフトウェア集約型システムと異なり非常に複雑で、アーキテクチャを困難にしている。

- アーキテクチャの決定を行うことで課題の解決を図る。

目的とアプローチ

●目的

- AIベースシステム開発のアーキテクチャ決定を分析し，実務者の視点からアーキテクチャ決定を行う際の限界と課題に関して特徴付ける．

●アプローチ

- Stack OverflowとGitHubをデータソースとして，アーキテクチャの決定に関するデータを収集．
- それらをあらかじめ定義された分類と定数比較法(Constant Comparison method)[27]の両方を用い分析．

RQ

- RQ1：AIを活用したシステム開発において、開発者はアーキテクチャをどのように表現しているか
- RQ2：AIを用いたシステム開発では、どのようなアーキテクチャの決定が行われるか
- RQ3：AIを用いたシステムで、アーキテクチャの決定が行われる領域は何か
- RQ4：AIを用いたシステム開発において、開発者がアーキテクチャを決定する際に考慮される品質属性は何か
- RQ5：AIを活用したシステム開発において、アーキテクチャを決定することの制限と課題は何か

実験概要

● 下図の手順で実行

1. システム開発におけるアーキテクチャの決定に関するSO投稿・Issueを探し出すため、検索用語を特定
2. 候補から上位のものを選定
3. 一連の包括・除外基準に従いフィルタリング
4. 得られたSO投稿・Issueからデータを抽出
5. あらかじめ定義された分類と定数比較法(Constant Comparison method)[27]の両方を用い分析

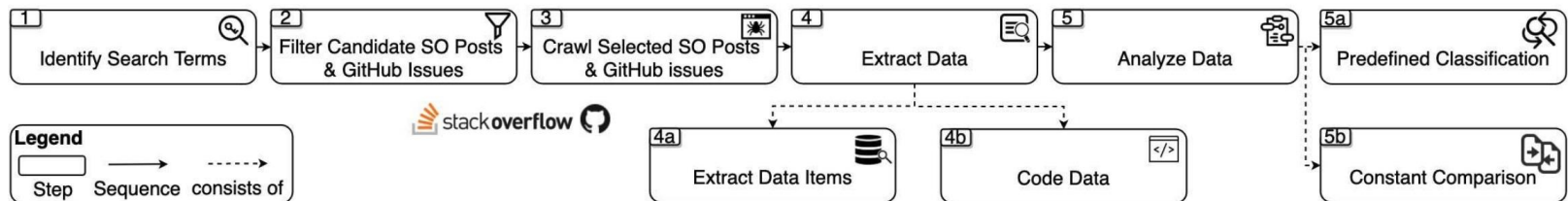


図1.研究プロセスの概要

実験概要

●分析は手順で実行

1. 第一著者と第二著者は、フィルタリングされた投稿と問題の内容にデータ項目をラベル付けした。
2. 第一著者は第二著者のラベリング結果を再チェックし、抽出されたデータが正しくラベリングされていることを確認。
3. 第一著者はすべてのコードを上位概念に結合し、カテゴリーに変換。その後、第三著者が分析結果を検討し、第一著者との議論を通じて意見の相違を解消。

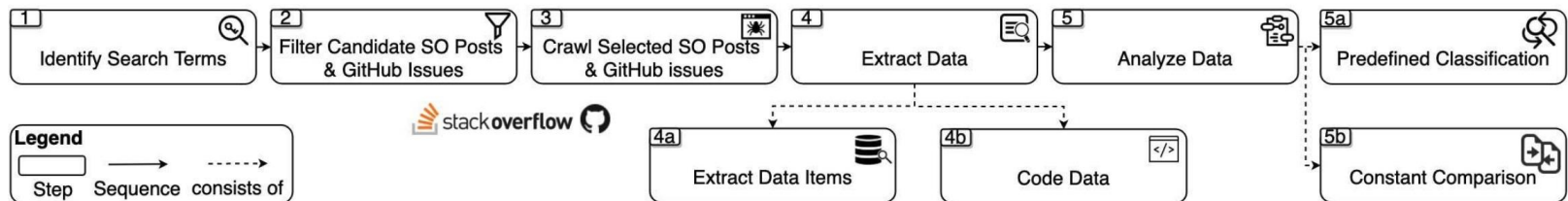


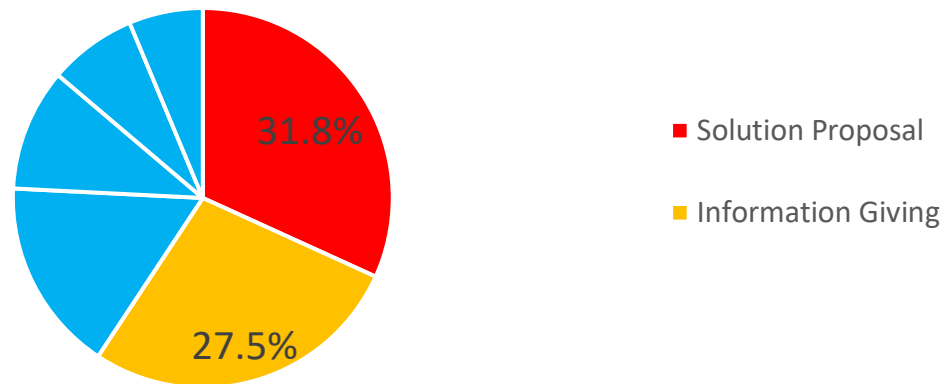
図1.研究プロセスの概要

RQ1

AIを活用したシステム開発において、開発者はアーキテクチャをどのように表現しているか

- あらかじめ定義された分類（Sorboら[26]が提案した言語パターン[25]）を用い6つに分類
- ソリューション提案・情報提供が最も多い
 - ソリューション提案（Solution Proposal）
 - アーキテクチャの決定として可能な解決策を提案
 - 情報提供（Information Giving）
 - 意思決定に関連する情報を他人に提供する際に使用

AIを用いたシステム開発におけるアーキテクチャ決定の言語的パターン

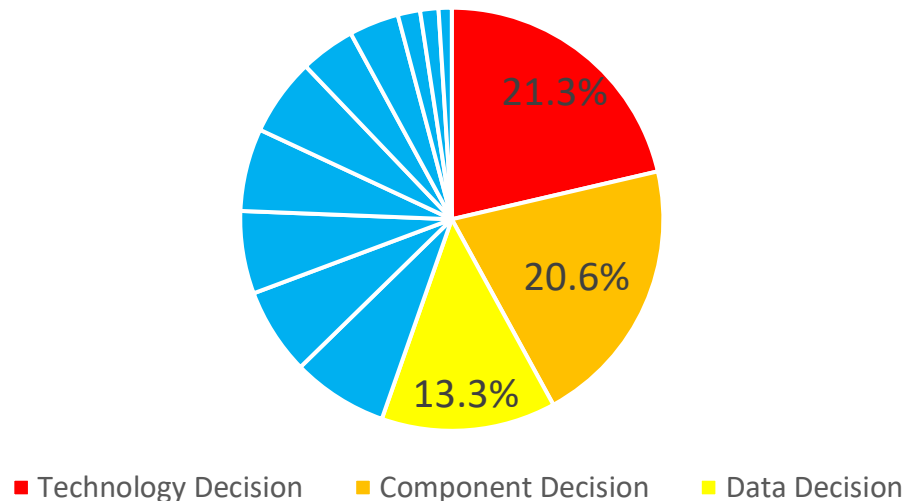


RQ2

AIを用いたシステム開発では、どのようなアーキテクチャの決定が行われるか

- 13種類のアーキテクチャの決定がある
- 上位3件は以下の通りである
 - 技術の意思決定 (Technology Decision)
 - 構成要素の意思決定 (Component Decision)
 - データの意思決定 (Data Decision)

AIを用いたシステム開発におけるアーキテクチャ決定の種類



RQ3

AIを用いたシステムで、アーキテクチャの決定が行われる領域は何か

- 145インスタンス（SO：106，GitHub：39）を獲得
- 18のアプリケーションドメインを獲得，上位4件は下記の通りである
 - ゲーム（Game）
 - データ処理（Data Processing）
 - 画像処理（Image Processing）
 - モバイルアプリケーション（Mobile Application）

アーキテクチャの決定が行われるドメイン



RQ4

AIを用いたシステム開発において、開発者がアーキテクチャを決定する際に考慮される品質属性は何か

- 137インスタンスを抽出，12種に分類
- 開発者が最もよく考慮しているのは「Performance」

TABLE VIII
THE QUALITY ATTRIBUTES CONSIDERED WHEN MAKING ARCHITECTURE DECISIONS IN AI-BASED SYSTEMS DEVELOPMENT

Quality Attribute	Example	Count	%
Performance	<i>However, a multi-threaded application would be more performant, as it would allow parallel usage without incurring a large memory requirement per "thread". (AllenNLP issue #443)</i>	51	37.2%
Maintainability	<i>It would hopefully cause it to be more maintainable AND have the added side effect of implementing your scripting interface for you. (SO post #609076)</i>	19	13.9%
Ease of Implementation	<i>Why can't OpenCV.org develop a simple module of OpenPose so that it can be easily imported, and effectively call some functions like find Hands location or draw the skeleton etc. (OpenCV issue #17271)</i>	13	9.5%
Cost and Effort	<i>Deploying Sagemaker to the same account that holds the rest of the business infrastructure is critical to avoid infrastructure overhead. (MLflow issue #4931)</i>	9	6.6%
Scalability	<i>AWS is a great fit for the application you describe. AWS Fargate/RDS will host your Django application. You have the option of using AWS Batch to handle your processing. One huge advantage is the ability to scale according to the needs of your application. (SO post #64390919)</i>	8	5.8%
Flexibility	<i>TensorLayer provides the access to the native APIs of TensorFlow and therefore help users to achieve a flexible control within the engine. (TensorLayer issue #185)</i>	8	5.8%
Complexity	<i>It all depends on the complexity your are ready to put in your project. (SO post #64892374)</i>	8	5.8%
Security	<i>If your model is non-generic and you want to protect it from being copied, this also has added security benefits since users won't have access to the model itself. (SO post #55065392)</i>	6	4.4%
Reliability	<i>Generally, people opt for a message bus like Kafka to handle back-pressure from slow downstream consumers and also to provide reliability (by persisting to disk) to prevent data loss. (SO post #39253328)</i>	6	4.4%
Usability	<i>We will learn more about the usability and improve in the future. (Kubeflow issue #5509)</i>	6	4.4%
Robustness	<i>It also has a very compelling convention over configuration mechanism (...). This is more powerful than any other mechanism I've seen. (SO post #10799101)</i>	2	1.5%
Persistence	<i>However even if location entities are nicely isolated there another problem - persistence. (SO post #1697820)</i>	1	0.7%

RQ5

AIを活用したシステム開発において、アーキテクチャを決定することの制限と課題は何か

- 49の事例を収集・分析し7種の制限と課題を特定
- AIベースのシステムに関する「Design Issues」とデータ保存・処理に関する「Data Issues」で90%以上を占める

TABLE IX
LIMITATIONS AND CHALLENGES OF MAKING ARCHITECTURE DECISIONS IN AI-BASED SYSTEMS DEVELOPMENT

Limitation & Challenge	Example	Count	%
Design Issue	<i>You made a valiant effort in the past which unfortunately stranded on architectural design. I don't see the use case for plotting the mean of several channels. (MNE-Python issue #3393)</i>	22	59.2%
Data Issue	<i>Then how should I architect my application in such a way that it should take care of scale-ability in term of supporting thousands of user logging huge amount of data every mill second. (SO post #52494424)</i>	16	32.7%
Integration Issue	<i>Unfortunately, it was not clear how to integrate DL4J within what we currently have or if it simply replicates the stack I have now. (SO post #47803989)</i>	1	2.0%
Programming Language Issue	<i>I could not find it spelled out in the documentation: Do AI Platform custom containers only support Python ML frameworks, or are other languages also supported? (SO post #61041666)</i>	1	2.0%
Protocol Issue	<i>Perhaps the AIs run as separate processes, and communicate with a central "board" process via sockets? I don't think there is a common protocol. (SO post #37369415)</i>	1	2.0%
Deployment Issue	<i>If I want to deploy my project including launch of multiple deep models built with different frameworks (Pytorch, tensorflow) then what's good option for that? (SO post #65785528)</i>	1	2.0%

まとめ

- 実務者がアーキテクチャの決定を表現するために使用する主要な言語パターンを特定
- アーキテクチャ決定について、決定の種類、アプリケーションドメイン、品質属性など様々な側面から検討
- アーキテクチャを決定する際に最も頻繁に遭遇する制限と課題を特定し、研究者が探求すべき潜在的な方向性を提示
- AIベースのシステム開発においてアーキテクチャを決定する際に、特定された制限や課題を解決することを目指す

データ項目について（資料）

#	Data Item (データ項目)	説明	データ分析方法	RQ
D1	Decision description 決定事項	AIを活用したシステム開発におけるアーキテクチャの決定について、実務家の理解に基づいた記述	Predefined classification[26] あらかじめ定義された分類	RQ1
D2	Decision type 判定タイプ	AIを用いたシステム開発におけるアーキテクチャ決定の分類	Constant Comparison method 定数比較法[27]	RQ2
D3	Application domain アプリケーションドメイン	アーキテクチャの決定が行われるAIベースのシステムの応用領域	Constant Comparison method 定数比較法[27]	RQ3
D4	Quality attribute considered 考慮される品質属性	AIベースのシステム開発において、実務者がアーキテクチャを決定する際に考慮される品質属性	Constant Comparison method 定数比較法[27]	RQ4
D5	Limitation and challenge 限界と課題	AIを活用したシステム開発において、実務者がアーキテクチャを決定する際の制約と難しさ	Constant Comparison method 定数比較法[27]	RQ5

RQ1 (資料)

TABLE V
LINGUISTIC PATTERNS OF ARCHITECTURE DECISIONS IN AI-BASED SYSTEMS DEVELOPMENT

Linguistic Pattern	Example	Count	%
Solution Proposal	<i>The other solution that we thought about was to send the output folder from the trainer to the ML Engine job, so that when our Module A asks for the status of that ML Engine Job, it gets the output folder.</i> (SO post #45411311)	769	31.8%
Information Giving	<i>I've attempted to reduce the number of modules as I don't require all modules, I have also removed slices from the framework to just leave arm64.</i> (OpenCV issue #13439)	664	27.5%
Information Seeking	<i>What is the architecture of spaCy (modules/layers and so on)?</i> (spaCy issue #5408)	398	16.5%
Feature Request	<i>This is not a bug rather a feature request. As TensorFlow library provides TFLite models to run on Android, iOS platform, can we a build tfjs wrapper to allow tfjs to directly load TFLite model in browser. This will allow same model being used across multiple platforms.</i> (TensorFlow.js issue #991)	251	10.4%
Problem Discovery	<i>The main question when using so many technologies is to scale well the infrastructure and the resources given to each one.</i> (SO post #27730628)	182	7.5%
Opinion Asking	<i>Does an object need to know how to update and draw itself, even having dependencies from higher levels that somehow are supplied to them?</i> (SO post #5458760)	153	6.3%

RQ2 (資料)

TABLE VI
TYPES OF ARCHITECTURE DECISIONS IN AI-BASED SYSTEMS DEVELOPMENT

Type	Example	Count	%
Technology Decision	<i>The general question is - what can be a resonable solution/technology/architecture of such system? (SO post #39253328)</i>	61	21.3%
Component Decision	<i>What you are asking to do is a big project that consists of too many components and no one can really show you how to do everything like that. (AirSim issue #1744)</i>	59	20.6%
Data Decision	<i>For video streaming, how should this c/s architecture program transmit data, and where should the decoder be placed? (Triton issue #4487)</i>	38	13.3%
Service Decision	<i>I have a problem with this service design. I was thinking about creating the main service on Kubernetes that will pull messages from PubSub then this main service would create pods (or rather jobs) to run the actual ML work. (SO post #66225804)</i>	21	7.3%
Implementation Decision	<i>Implement an efficient encoding of a knowledge base + all APIs / interfaces, to integrate with the current processing pipeline. (spaCy issue #3339)</i>	19	6.6%
Deployment Decision	<i>We have deployed a non-managed Kubernetes cluster using an EC2 instance and we connect the Worker Nodes to the master using a VPN. (SO post #55886095)</i>	18	6.3%
Tool Decision	<i>Looking for any references architectures, or tooling that can be stiched together so we can incrementally improve our setup. (SO post #11587135)</i>	18	6.3%
Concurrency Decision	<i>I'm developing a multiplayer game application with C++ and currently in the process of choosing an appropriate multithreading architecture for it. (SO post #1697820)</i>	17	5.9%
Integration Decision	<i>We decided to build a flask server acting as a relay for incoming requests to Triton, meaning we have integrated the Triton Python HTTP client inside a Flask app. (Triton issue #2744)</i>	12	4.2%
Architecture Pattern Decision	<i>I am also using a flask driven microservice architecture. (AllenNLP issue #1313)</i>	11	3.8%
Design Pattern Decision	<i>While looking at this architecture, I am immediately thinking about a Chain of Responsibility Design Pattern. (SO post #63788496)</i>	5	1.7%
Containerization Decision	<i>When NM launches a container, move job container into k8s pod. (OpenPAI issue #2195)</i>	4	1.4%
Decomposition Decision	<i>I'm guessing this will be addressed by a re-engineering of the demo architecture, separating different models? (AllenNLP issue #443)</i>	3	1.0%

RQ3 (資料)

TABLE VII
THE DOMAINS IN WHICH ARCHITECTURE DECISIONS ARE MADE

Domain	Count	%	Domain	Count	%
Game	45	31.0%	E-Business	3	2.1%
Data Processing	24	16.6%	Chatbot	3	2.1%
Image Processing	17	11.7%	Social Network	2	1.4%
Mobile Application	14	9.7%	Food Industry	1	0.7%
Unmanned System	9	6.2%	Business Process Management	1	0.7%
Real-time System	7	4.8%	Expert System	1	0.7%
Internet of Things (IoT)	6	4.1%	Embedded System	1	0.7%
Biology and Chemistry	5	3.4%	Mechatronics	1	0.7%
Audio Processing	4	2.8%	Robotic System	1	0.7%