

Study Sapuri

現場で実践！RAG活用術 Lunch LT — 運用して
分かった"つらみ"とその対策

RAGの簡易評価によるフィードバックサイクル実践

Takeshi Kondo (@chaspy)



Takeshi Kondo (@chaspy)

Director of Engineering
StudySapuri K12 at Recruit Co., Ltd.

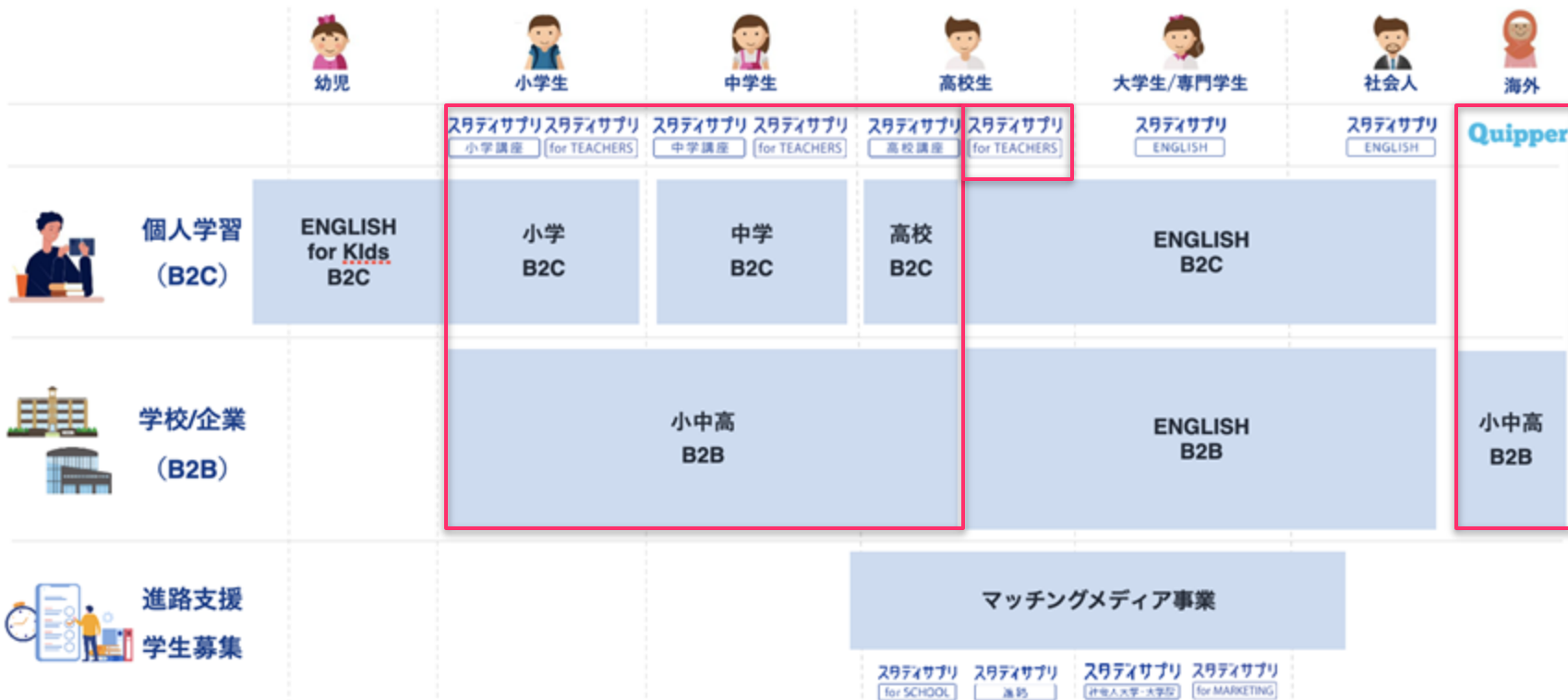
観葉植物 🌿 クラフトビール 🍺 が好き
今日は友達増やしにきました！

🔄 chaspy
✕ chaspy_
🏠 <https://chaspy.me>



スタディサプリプロダクトについて

国内小中高と
海外を担当しています



Agenda | 01 背景

02 RAG のフィードバックを得る方法

03 課題 (つらみ)



こういうのを作りました



chaspy 🍷 2 minutes ago

@zeus DarkLaunch ってなんですか？

社内の Feature
Toggles 基盤のこと

Also sent to the channel



zeus APP 2 minutes ago

DarkLaunchとは、新機能のリリースを特定のユーザーに対してのみ行うことで、全体に公開する前にその機能のテストや評価を行う手法です。DarkLaunchV2は、スタディサプリのWebアプリやネイティブアプリが自由にアクセスできる新機能のリリースなどの粒度を司るシステムであり、公開APIと管理画面から成り立っています[doc1]。

このシステムを用いることで、リクルート社員または提携業者によるスタディサプリの開発をより安全かつ高速に行えるようにすることが目的です[doc1]。特定のユーザーに対して特定の新機能がリリースされているかどうかをtrue/falseで返すAPIを提供し、ユーザーの同定はUserIDやorganizationIDなどで行います[doc1]。

また、DarkLaunchV2はインターネットフェイシングなAPIであり、誰でもアクセスが可能ですが、実際に新機能のリリース状況を把握するためにはユーザーの同定情報が必要です[doc1]。管理画面では、特定の機能を有効化する条件を簡単に設定することができ、例えば全ユーザーに対して有効、特定のユーザーIDにのみ有効などの設定が可能です[doc1]。

回答生成に利用されたドキュメントはこちら:

- doc1
- doc2
- doc3

ソースドキュメント
へのリンク

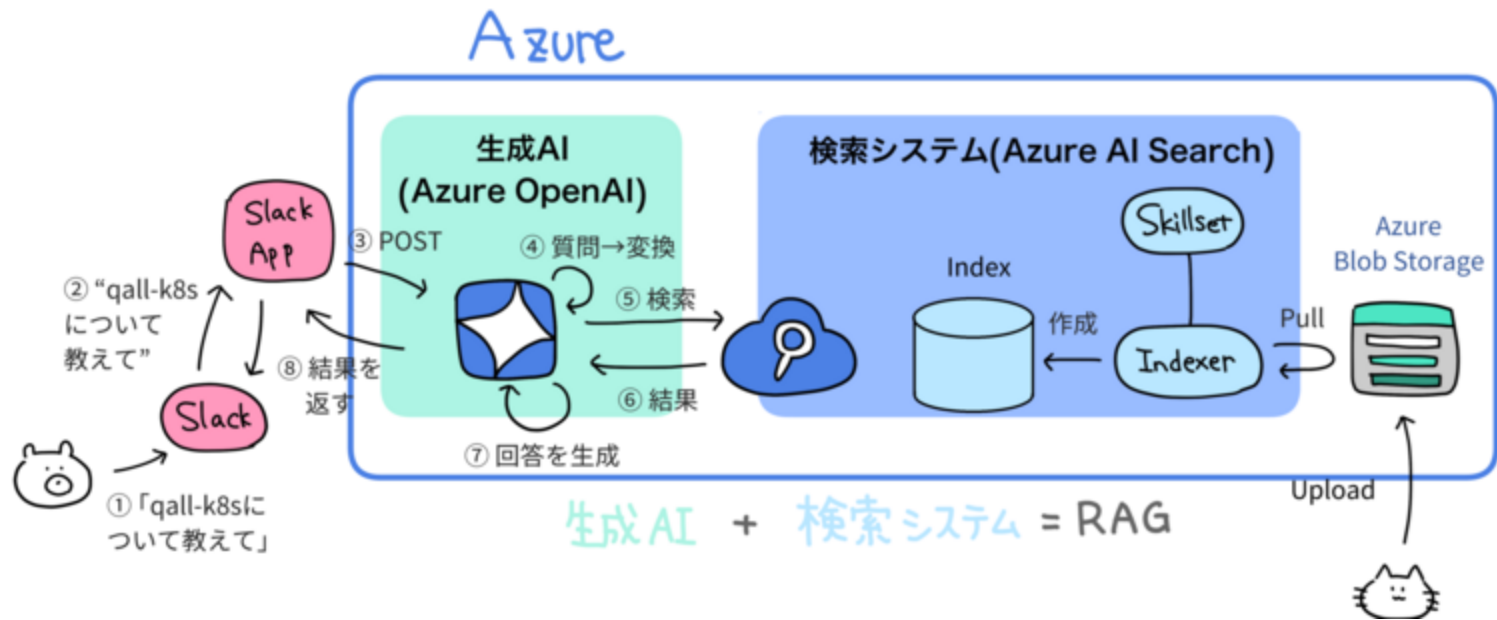
簡易フィードバックの
お願い

@chaspy

この回答が満足であったら 👍 リアクションを、解決しない場合はスレッドでその理由とともに @zeus-devs をメンションしてください

["prompt_tokens":12405,"completion_tokens":349,"total_tokens":12754]

アーキテクチャ



素敵な図はチームメンバーの aoi さん作のものをお借りしました。詳細はブログ [RAGを使って社内のGitHubリポジトリに散らばっているドキュメントを自然言語で検索できるSlack botを作りました - スタディサプリ Product Team Blog](#) もご覧ください！

RAG の評価をする上での前提となる考え

- 現状ベータ版であり、ユーザからフィードバックを高速に得て、改善を繰り返すかが重要
- やりたいことは**検索**であり、ソースドキュメントにたどり着くことをゴールとしているので、高度な評価指標は必要ない
 - ◆ ※ドキュメントは monorepo で markdown + docsify などの SSG でビルド & Amazon S3 で配信しているものが多い

簡易的に評価する手法

- (1)回答に満足したかどうかを👍リアクションで促す
 - 生成 AI のアウトプットに固定文を追加しているだけ
- (2)ローカルで同一の API を実行し、求める検索結果が得られたかどうかを確認する
 - E2E でのリグレッションテスト相当
 - テストケースと回答のセット (質問/url) を csv で持つ

(1)回答に満足したかどうかを👍リアクションで促す



chaspy 🍷 2 minutes ago

@zeus DarkLaunch ってなんですか？

Also sent to the channel



zeus APP 2 minutes ago

DarkLaunchとは、新機能のリリースを特定のユーザーに対してのみ行うことで、全体に公開する前にその機能のテストや評価を行う手法です。DarkLaunchV2は、スタディサブリのWebアプリやネイティブアプリが自由にアクセスできる新機能のリリースなどの粒度を司るシステムであり、公開APIと管理画面から成り立っています[doc1]。

このシステムを用いることで、リクルート社員または提携業者によるスタディサブリの開発をより安全かつ高速に行えるようにすることが目的です[doc1]。特定のユーザーに対して特定の新機能がリリースされているかどうかをtrue/falseで返すAPIを提供し、ユーザーの同定はUserIDやorganizationIDなどで行います[doc1]。

また、DarkLaunchV2はインターネットフェイシングなAPIであり、誰でもアクセスが可能ですが、実際に新機能のリリース状況を把握するためにはユーザーの同定情報が必要です[doc1]。管理画面では、特定の機能を有効化する条件を簡単に設定することができ、例えば全ユーザーに対して有効、特定のユーザーIDにのみ有効などの設定が可能です[doc1]。

回答生成に利用されたドキュメントはこちら:

- doc1
- doc2
- doc3

これです!!!

@chaspy

この回答が満足であったら 👍 リアクションを、解決しない場合はスレッドでその理由とともに @zeus-devs をメンションしてください

["prompt_tokens":12405,"completion_tokens":349,"total_tokens":12754]

(1)回答に満足したかどうかを👍リアクションで促す

- 意外と押してくれる
 - 将来、満足度を計測するのにも使えそう
 - インターフェースが Slack だったからこそその利点
 - チャンネルはウォッチしてるので満足していなさそうだったら @chaspy が回答していますw
 - このタイミングで、ソースドキュメントの追加・編集をやっている

(1)回答に満足したかどうかを👍リアクションで促す

yskttm 2:10 PM
@zeus Platonとはなんですか？
1 reply 12 days ago

zeus APP 2:11 PM
replied to a thread: @zeus Platonとはなんですか？
申し訳ありませんが、提供された文書の中には「Platon」に関する情報は含まれていません。他の情報源をご覧になるか、具体的な文脈を提供していただけると、より適切な情報を提供できるかもしれません。この回答が満足であったら👍リアクションを、解決しない場合はスレッドでその理由とともに @zeus-devs をメンションしてください。

回答は得られなかったが...



chaspy ❤️ 12 days ago
https://data-div-handbook/#/product-development/bigquery_guide_ja/things_good_to_know_first
スタディサブリ、スタディサブリEnglishでは、分析基盤としてBigQueryを採用していますが、サービス毎に異なるプロジェクトにまとまっています。
スタディサブリK12の基盤を Platon、スタディサブリEnglishの基盤を Organon と呼んでいます。
👍 1 🗨️

chaspy ❤️ 12 days ago
多分これが答えだが

人間が回答しつつ、ドキュメント追加



Also sent to the channel

chaspy ❤️ 11 days ago
@zeus Platonとはなんですか？

zeus APP 11 days ago
Platonは、スタディサブリK12の分析基盤として使用されているシステムです。このシステムは、BigQueryを採用しており、サービスごとに異なるプロジェクトでデータが整理されています[doc3]。また、Platonは、DBからの個人情報またはそれに準ずるものをマスクした後のデータがまとめられています[doc3]。



(1)回答に満足したかどうかを👍リアクションで促す

- フィードバックはタイミングと簡単さが重要
 - google map の経路案内後の UI にヒントを得た
 - タスクを終えて一息ついたタイミングで、押すだけ
 - 逆にタスク中で邪魔なタイミングだと反感を買う



(2)ローカルで同一の API を実行し、求める検索結果が得られたかどうかを確認する

- 質問と期待するソース url のセットを csv で持つ
 - Qall-k8sって何?, <https://docs.xxx.com/xxxx/yyy/qall-k8s>
- 1行ずつ呼んで、api を呼び出し、最終回答に url が含まれるかを確認するだけ

(2)ローカルで同一の API を実行し、求める検索結果が得られたかどうかを確認する

- ドキュメントが増えていくにつれて検索精度が落ちることが予想される
 - それに管理者側で事前に気付けるようにする
 - 検索システムと割り切ることで、高度な評価をしない

課題 (つらみ)

- (1)コスト高い問題
- (2)ソース url の伝播難しい問題
- (3)index 再生成うまくいってない問題

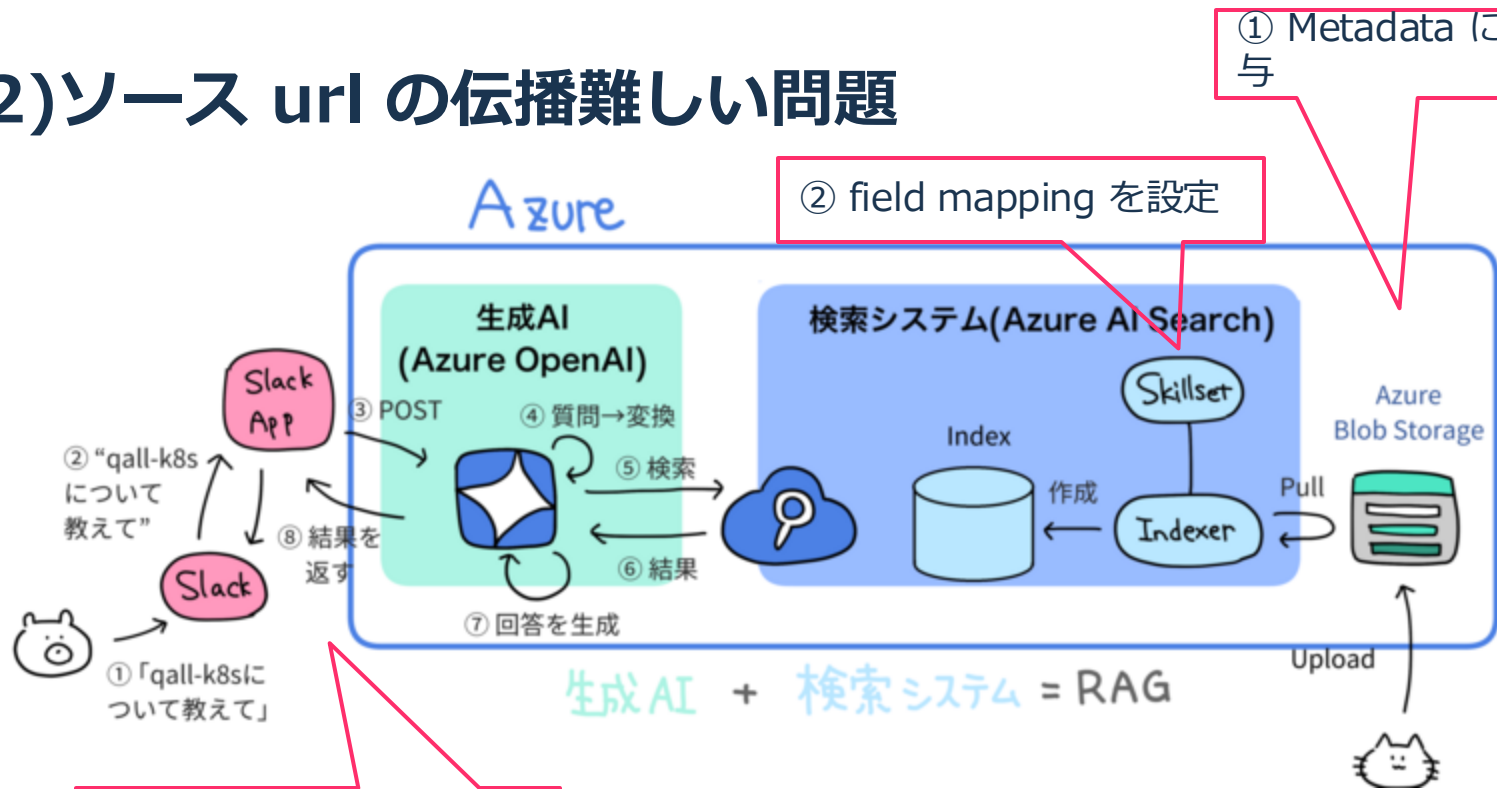
(1)コスト高い問題

- 最初に日本リージョンを選んでしまったため、最新モデルが使えるまで時間がかかる
 - gpt-4-32k をしばらく使っていたが、冷静に考えると高い
 - input ¥9.637 / 1000 token (Azure OpenAI)
 - 調べると **Input token 数がかなり多い**
 - 1回の問い合わせで10000token使っていた、つまり1回あたり100円

(1)コスト高い問題 - 解決策

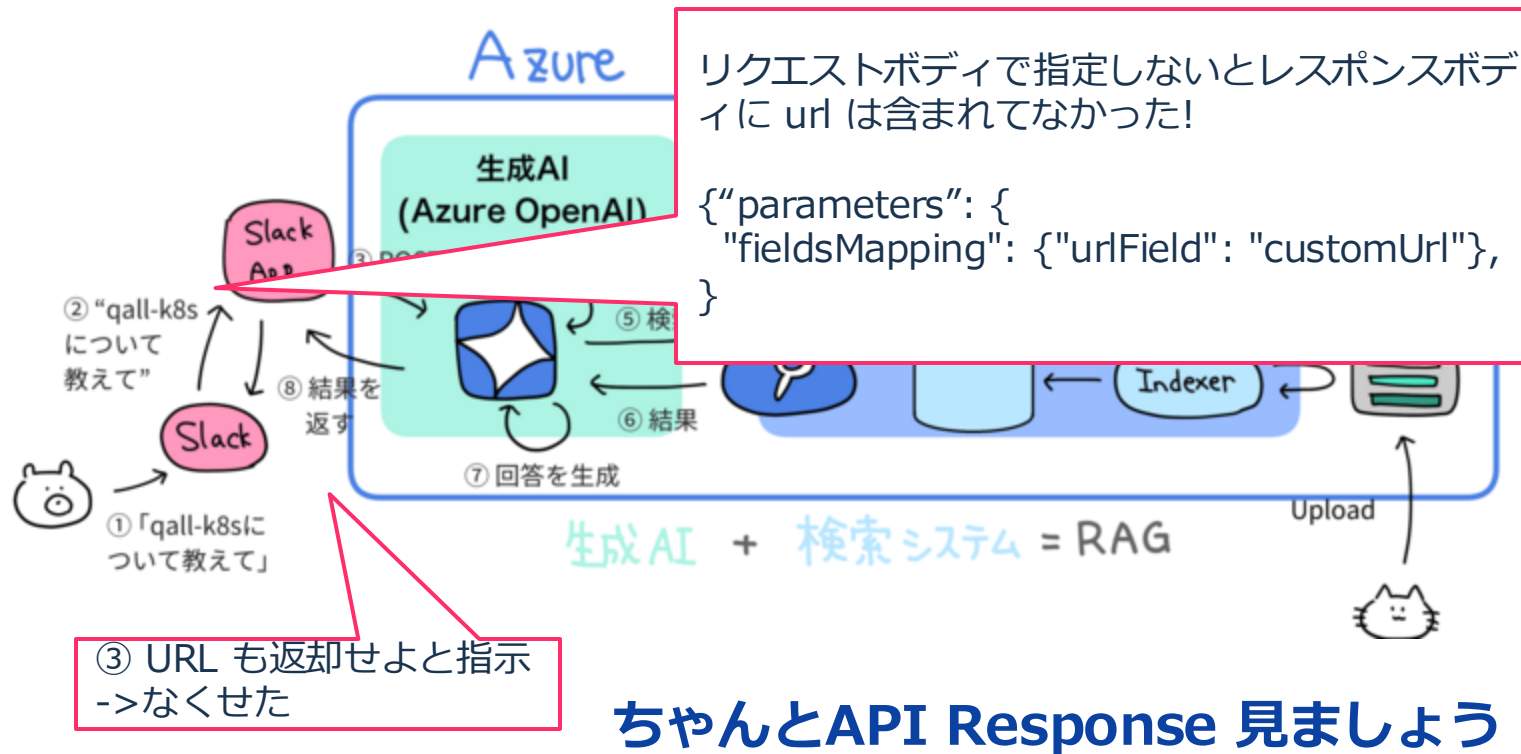
- まず計測 - 回答時に token を出力
- 採用する検索結果を5件から3件に
- 米国 Region で作り直そうと思ってたら先週日本 Region でも gpt-4o 使えるように 🎉
 - ¥0.8031 / 1000 input token, 1/10 以下に
- 最初から最新モデルが使えるリージョンで作ろう

(2) ソース url の伝播難しい問題



しかし回答結果は不安定...

(2) ソース url の伝播難しい問題



(3)index 再生成うまくいっていない問題 (未解決)

- 当時試行錯誤していた際、blob のメタデータに付与した url が間違っていたり、null のままだったりした
- その場合、indexing すると別の index が生成されてしまう
- index の Reset で解決するはずだが、AI Search で検索すると過去の誤ったチャンクが引っかかってしまう... (なんで?)
 - とはいえ E2E で実行して問題ないので、問題ないのかも

まとめ

- Slack 上で問い合わせる RAG を使った Bot を作った
 - ◆ 実質ドキュメント検索システムとして作っている
- 回答に 👍 をつけてもらったり、E2E でリグレッションテストを実施することで評価している
 - ◆ 簡易な評価でもフィードバックサイクルを回すのに十分役立っている
- 課題としてコストや、特有のソース url の伝播や、index の reset に関するつまづきを紹介しました

Thank you for listening!

Takeshi Kondo (@chaspy)

Director of Engineering
StudySapuri K12 at Recruit Co., Ltd.

今日紹介した課題は LT 発表決まってから解決しました！LT 駆動進捗 😊



chaspy



chaspy_



<https://chaspy.me>

