



AI議事録の開発で感じた
ユーザー目線のAIの精度
を測る難しさ

自己紹介

高山 雄貴

- Nishika株式会社 リードAIエンジニア
 - 自社のAI議事録プロダクトのAIインフラの運用
 - 企業向けAIソリューションの開発/PoC案件のPM
 - AIの研究開発
- Kaggle Expert
- Cursorを使っているが、社内がClaude Code1択になりつつあるので乗り換え予定
- マイブームは半年前に始めたボルダリング





Gemini



皆さんはどのAIチャットの
サービスをよく使いますか？

 perplexity

 Claude



AIの精度の良さを定量評価するのは難しい

- 皆さん、どのAIサービスを使っていますか？→「精度がいいから」
- でも「精度」を言語化するのは難しい
- 測る物差しが多種多様
 - 例：コーディングの質問にどれだけ適切に回答できるか
 - 例：曖昧に質問しても、背景や意図を補ってよしなに回答できるか



AI議事録サービスのAI開発でも、ユーザーにとっての精度を測ることは難しい

SecureMemo

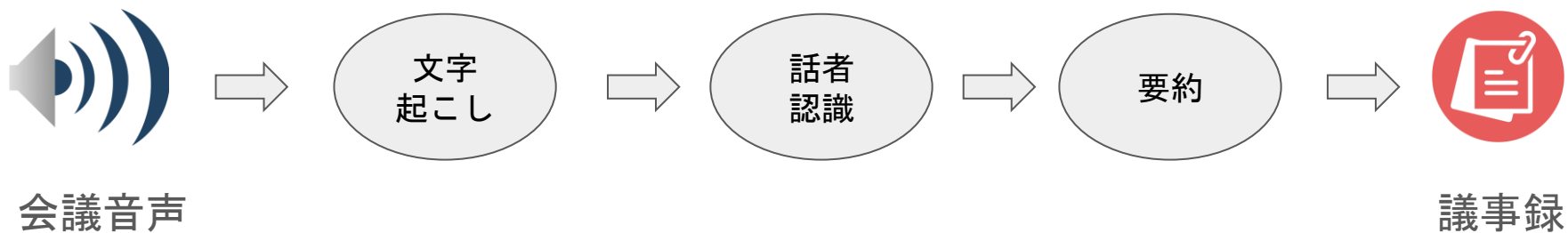
会議音声をもとに、AIが文字起こし・話者認識・議事録作成を一気通貫で行い会議の情報を資産化し有効活用できるAI議事録サービス

The screenshot displays the SecureMemo interface for a meeting titled "製造業デモ_250808.m4a". The main area shows a transcript with timestamps and speaker labels (e.g., "Nishikaデモ_発声"). The transcript content includes:

- 00:02- Nishikaデモ_発声: 皆さん今日は生産管理会議にお集まりいただきありがとうございます。
- 00:05- Nishikaデモ_発声: まず今月の生産目標に対する達成状況を確認しましょう。田中さん現状はいかがでしょうか。
- 00:14- Nishikaデモ_発声: 今月は生産量が計画より10%ほど上回っています。
- 00:19- Nishikaデモ_発声: タクトタイムを少し短縮したことが功利的だと思います。
- 00:23- Nishikaデモ_発声: ただ一部の部品在庫が溜まってきているので、
- 00:26- Nishikaデモ_発声: 後工程や搬送方式の改善を再検討した方が良さそうです。
- 00:34- Nishikaデモ_発声: 保管庫内の確保や在庫管理コスト面の問題が出ないようにしたいですね。
- 00:40- Nishikaデモ_発声: 品質面については先週プレス社で発生した異状がまだ解決にできていない状態です。
- 00:46- Nishikaデモ_発声: サーチ検査でNGが出たロットを追跡したところプレス後のフレームの歪みがあるのではないかと聞いています。
- 00:55- Nishikaデモ_発声: 寸法のばらつきを抑えたいという要望があり引き続き部品を詳しく調べていきます。
- 01:02- Nishikaデモ_発声: フレームの歪みですが、メンテナンスチームには報告済みです。フレームの歪みですが、メンテナンスチームには報告済みです。
- 01:07- Nishikaデモ_発声: どこまで進んでいるか把握していますか。
- 01:09- Nishikaデモ_発声: メンテナンスチームに連絡して、X光測定器で寸法を測定中です。
- 01:15- Nishikaデモ_発声: もし部品の交換が必要となると保守部品の手配に時間がかかる恐れがあります。
- 01:22- Nishikaデモ_発声: 現時点では保守で対応できるかどうかを見極めていこうです。

The sidebar on the left contains navigation options such as "文字起こしを開始", "ファイル一覧", "メンバーとグループ", "AIカスタマイズ", "利用状況", "ワークスペース設定", "プロフィールを設定", "Web会議連携", "使いかマニュアル", "リソースノート", and "サポートフォーム". The bottom right corner shows a "日次開催事項" section with a list of items including "タイトル", "日時", "参加者", "議定事項", and "ToDo".

✓ AI議事録を作成するAIのパイプラインは複雑





ユーザー目線の文字起こし精度は、 CER（文字単位の誤り率）と一致しない

正解の文字起こし

「アサヒ商事との案件については予算は、予算上限を超過する以上、見直しが必要です」

不正解の文字起こし

「えーっと、アジア商事との案件については予算は、予算上限を超過する以上、見直しが必要です」

- 大事な情報である会社名（固有名詞）を誤り、スコア以上に全体的に誤っている印象を持ちやすい
 - CER/WERと合わせて名詞誤り率も見るようにする
- 「えーっと」というフィラー（繋ぎ言葉）が含まれる
 - 含めてほしいユーザーとそうでないユーザーで意見が分かれる



ユーザー目線の話者認識精度は、 DER/IER（話者誤り率）と一致しない

DER/IERは

「誰がいつ話していたか」の正解と、
AIの認識結果を時間軸で比較して、話
者がズレていた時間の割合を示す指標

話者情報

時間	話者	発言内容
00:00	speaker_1	はい。この前マイザークリームを出してもらって背中ですね。ここなんですけども塗って良かったので、インブロンクリームに変えたんですけども、最近汗がひどくなるようになって、ちょっとひどくなってきているんですけども、このままでもいいですかね。
00:20	speaker_0	マイザークリームちゃんと塗ってます。塗っているよね。でも汗で悪化しているんだよ。
00:26	speaker_0	だいぶ良かったけども、一部この辺が少しサルと集まってくデコボコくれますよね。
00:34	speaker_0	そういうところは。汗も結構悪化意思になっているから、ちゃんとシャワーを浴びて浴びた後、すぐにマイザー。
00:42	speaker_0	それがある程度いいところはもうちょっとリンで飲んでいいと思うんですけども、マイザーを塗る場所はしっかりまだ続けた方がいいですね。
00:51	speaker_1	まだマイザーを続けた方がいいですね。
00:55	speaker_1	何日くらいに塗った方がいいですか。
00:56	speaker_0	その質問なかなか難しいよね。でも今手で触ってデコボコしたところをちゃんと塗りましょうって。
01:03	speaker_0	皮膚の症状が炎症があるところはどうしても触ると、デコボコして触れるから、それが消えるまでやっぱりちゃんと塗らなきゃいけないし消えにくいようだったら塗りが足りないと思わなきゃいけないから。しっかりとその部分の触った
01:19	speaker_0	大体
01:22	speaker_1	
01:26	speaker_0	
01:26	speaker_1	だい

- ユーザーから見て目立つ誤りとそうでない誤りがある
 - 1人がずっと喋っているところに他の人が混じっていないか
 - 誰か1人全く認識できてない人がいないか
 - **話者ごとの誤り率も見るようにする**



ユーザー目線の要約精度は、 内容・形式がユーザー自身の理想に近いか

- 議事録の要約精度は最も評価方法が悩ましい
- 弊社ではAzure OpenAI Serviceを採用
- 内容面のゆれ
 - 人によって、重要視している会議の議論の内容のポイントは様々
- 形式面のゆれ
 - ユーザーによって求める議事録の形式は様々
 - 文字起こしの情報をあまり落とさない逐語録
 - 決定事項やTODOのみの箇条書き
- 弊社での運用
 - 評価観点を決めて、LLM as a Judgeと呼ばれるLLMによる定量評価の手法も導入しているが、見切れない観点もある
 - 人の目を見て、定性的に評価することも大事にしてモデルの選定を行っている

SecureMemoCloud 商談_議事録

日時

- 2024年10月25日13:25

参加者

- 松田、鈴木、山本

決定事項

プランの決定

- チームプランを採用することを決定した。
 - 25時間の利用時間を含む。
 - 要約機能を使用する。
 - 従量課金は適用しない。

契約形態の決定

- 年間契約を基本とするが、月契約の見積りも依頼することを決定した。

議事要旨

利用時間とアカウント数の確認

- 録音を開始しました。ありがとうございます。早速ですが、本題に入ります。何時間利用するか、何人で利用するかの希望を伺い、それに基づいたカスタムプランを考えたいと思います。(松田)
 - 最大で実際は9時間程度ですが、バッファを見て1ヶ月で12時間程度になります。(山本)
 - 12時間ということで了解しました。(松田)
 - 実際の文字起こしした結果をウェブ画面上で共有したい人数、アカウント数のイメージは何人でしょうか。(松田)
 - 台数は一台で良いですが、アカウント数によって料金は異なりますか？(山本)
 - 例えばチームプランのアカウント数は10名となっております、ここまでは固定課金ですが、10名以上増やす場合は従量課金となります。(松田)
 - 承知しました。であれば1アカウントで良いです。(山本)
 - 承知しました。(松田)

利用時間の再確認とプランの選定

- すいません。もう1回会議の時間を計算し直したら、7月は18時間とかになるんですけど。(鈴木)
 - わかりました。(松田)
 - 25時間でいいんじゃないですか。(山本)
 - そうですね。25時間で本当に重い時だとそれぐらいかかってしまうので、ない時はないんですけど、ある時は18時間だなど思ったんですけど、いろんな会議があるなと思って。だったら25時間でいいのかもしれないですね。(鈴木)
 - 25時間で1名で固定でしていただけると助かります。(山本)



LLMの新しいバージョンのリリース後、
その日に本番環境のモデルを
更新する会社も時々見る。
すごいと思っている。

モデルの出力の定性的な評価をする
時間はどう考えてもする時間がない
はず・・・

新しいモデルだったら基本精度は良
いという思い込みのもとやるしか
ないのか



評価用のデータを用意することも難しい

公開されている音声データと
データの分布が異なるリアルな
会議音声を収集することが必要がある

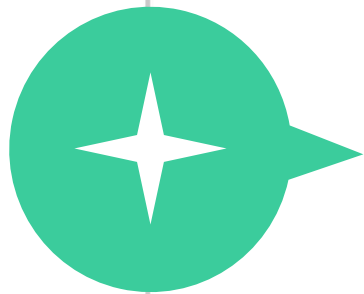
会議音声の文字起こしと
議事録の形式・内容の
正解データを作成する難しさ

だれがいつ発言をしていたかの
話者のラベリングは
膨大なコストがかかる



宣伝

- NishikaではAI議事録の開発を推進するアプリエンジニアを中心に積極採用中です。詳細は[こちら](#)
- [Nishika Tech Blog](#)では毎週のAIの論文情報発信を中心にホットな技術ネタを配信していますので、ぜひチェックしてください！



ご清聴ありがとうございました！

