繰り返しのテストを要する生成AIテストの効率化

類似度算出と同義文判定による検証コスト削減の検討

多田 麻沙子（TIS）
中川 桂（東京海上日動システムズ）
金丸 優介（AGEST）
石川 冬樹（国立情報学研究所）
徳本 晋（富士通）
栗田 太郎（フリー）

目次

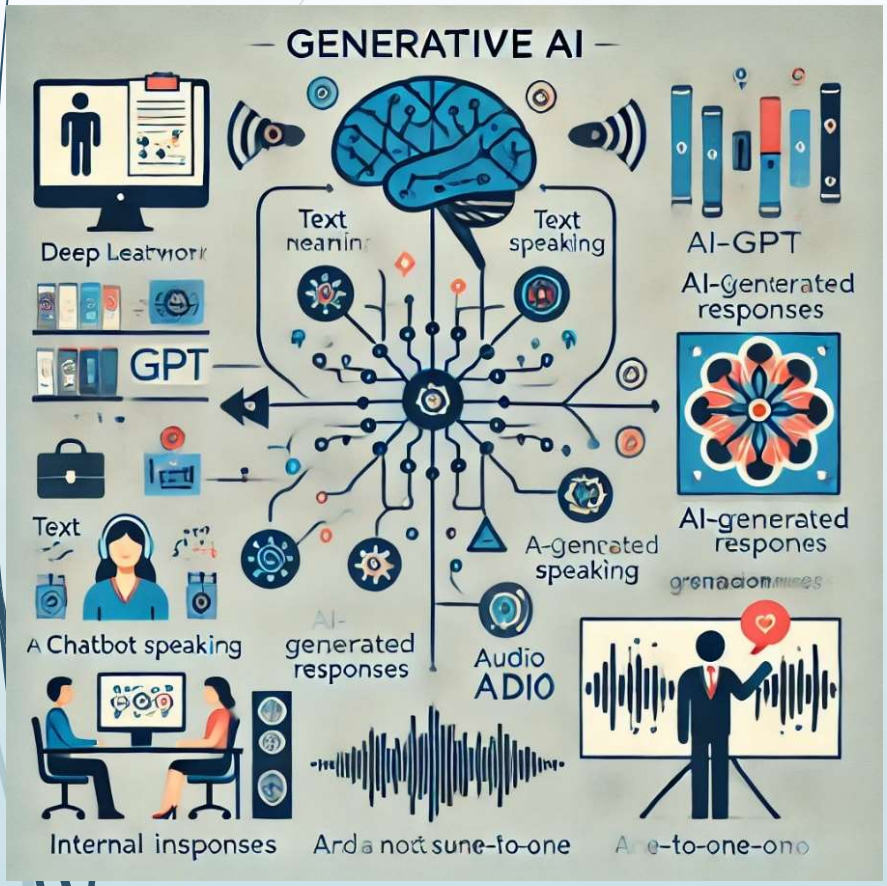
０．前提

本発表の位置づけ

　本発表はテスト効率改善を目標に実験したもので、実業務では未実践のものです。

・実務上における工数等は算出できていません。

1．研究の背景①：「生成AIって何？」



生成AIとは

- 生成AIは、文章・画像・音声などを自動生成するAI技術である
- 自然言語で回答を生成し、同じ質問にも異なる表現で応答する柔軟性がある
- 企業ではチャットボットやコンテンツ作成などに活用されている

1. 研究の背景①：「生成AIって何？」

主な活用例

- カスタマーサポートチャットボット
  - 企業の問い合わせ対応を自動化し、顧客の質問に自然言語で回答する
    - 例：ECサイトのカスタマーサポート、銀行の問い合わせ対応
- 社内問い合わせ対応
  - 社員向けに、社内規定や業務手順を回答するチャットボットを導入
    - 例：人事・総務関連の問い合わせ対応、ITサポートデスク
- マーケティング・コンテンツ生成
  - 広告コピーやブログ記事、SNS投稿の文章を自動生成
    - 例：商品のキャッチコピー作成、メールマーケティングの文面作成

1. 研究の背景①：「生成AIって何？」

生成AI（LLM）の特徴

- 自然言語で回答を生成可能
  - 人間が使う言葉でスムーズに応答できる
  - 専門知識がなくても直感的に利用可能
- ランダム性のある回答
  - 同じ質問に対しても異なる表現で回答することがある
- 迅速な応答
  - 人間と比べ、短時間で回答を生成
- 正確性は保証されない
  - 生成された回答に誤りが含まれる可能性がある
  - 事実確認が必要な場面では注意が必要

2. 研究の背景②：「生成AIのテストは何が大変？」

生成AIの出力は同一の入力でも異なる回答が生成される事がある。

回答の都度文章を読解して処理の正否を判断する必要がある。

効率化の文脈での生成AIの役割の回答イメージ（人間が説明用に作成）

No		評価
1	生成AIは、創造的なプロセスを効率化する技術である。	模範解答
2	生成AIは、創造的なプロセスを効率化する事が出来る技術です。	模範解答と同義
3	生成AIとは、新しいデータやコンテンツを自動的に生成する人工知能のことです。	違う文脈の答えである

同様の回答をする場合でも、文章表現は異なる場合がある。

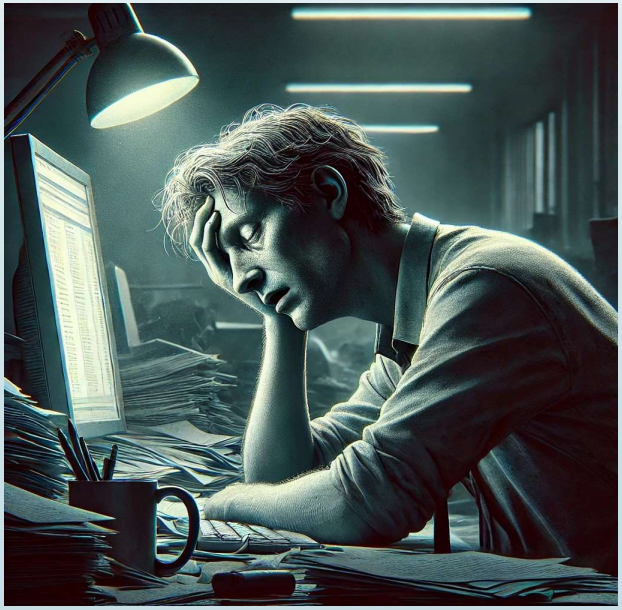
初回は成功した場合でも複数回実施すると誤った回答が出力する可能性がある。

2. 研究の背景②：「生成AIのテストは何が大変？」

生成AIのテストは繰り返しテストが必要。
なぜなら生成AIは回ごとに回答が変わることがある

- それゆえに以下のようなことが大変
  - テスト回数が増えることの手間
  - ユーザテスト(UAT)ではユーザに同じ検証項目でも複数回の検証を依頼
- ユースケースとして
  - LLMのバージョンアップ時の動作確認、
  - LLMを組み込んだアプリのリグレッションテスト

今回範囲ではないが、同じ仕様書を元に人とAIでコーディングし、同内容であるか、などに発展させても興味深い



3. 本研究の目的

繰り返し行うテストの手間を削減したい

- 膨大なテキストを人間だけで評価するのは大変
- そこで「同義文判定」を検討した。
  - 似ていれば「同義」としてOK → 人間が何度も文章を読む手間削減

No		評価
1	生成AIは、創造的なプロセスを効率化する技術である。	模範解答
2	生成AIは、創造的なプロセスを効率化する事が出来る技術です。	模範解答と同義
3	生成AIとは、新しいデータやコンテンツを自動的に生成する人工知能のことです。	違う文脈の答えである

似ている文章の判定を機械的に支援する。
今回の研究対象。

3. 本研究の目的

例えば・・・
チャットボットに同じ質問をして常に同じ回答をするか確認したい。

試行No1は人間がインプットに対して回答が正しい事を検証する。

効率化の文脈での生成AIの役割の回答確認ケース:

試行No	回答	人間検証	類似度判定	閾値(80設定)
1	生成AIは、創造的なプロセスを効率化する技術である。	✓	-	
2	生成AIは、創造的なプロセスを効率化する事が出来る技術です。		90	合格
3	創造的なプロセスを生成AIの技術で効率化します。		85	合格
4	生成AIは、創造的なプロセスを効率化する技術では**ない**。		5	**不合格**

試行No2～4は同一インプットを機械的に実行し回答を得る。
回答が人間が検証したものと同様か「同義文判定」を元に自動的に評価する。

3. 本研究の目的

- 長文の場合に同義判定が「同じ」になりやすいとの仮説

自動翻訳の評価指標において、長文ほど「同義」と判定されやすいという知見を得た。これを踏まえ、類似文を作成する際に、意味が異なっていても共通する単語が多い長文は、AIに「同義」と判断されやすいのではないか、という仮説を立てた

長文例:

試行No	回答
1	創造的なプロセスの効率化とは、アイデアの発想やコンテンツの生成にかかる時間や労力を削減し、より迅速かつ効果的に成果を得る方法を指す。生成AIは、創造的なプロセスを効率化する技術である。
2	創造的なプロセスの効率化とは、アイデアの発想やコンテンツの生成にかかる時間や労力を削減し、より迅速かつ効果的に成果を得る方法を指す。生成AIは、創造的なプロセスを効率化する技術ではない。

二つの文は最終的に文末の「ある」と「ない」で全く逆の意味の文となっている。
しかし、その前の文章は一致している。

4．研究課題

同義文判定の方法として，以下を研究課題としました.

類似度の数値化方法

- **研究課題1:** 文章の類似度の数値化と人間の感覚への適合性
  - **課題1-1:** 埋め込み表現のコサイン類似度による評価
  - **課題1-2:** 生成AI(ChatGPT)による類似度評価

自然言語を扱うのが得意なため選定

長文化の影響

- **研究課題2:** 長文の場合、類似度の検出性能がどのように影響を受けるかの評価

期待効果

効率的な評価手法の確立による、評価工数の削減

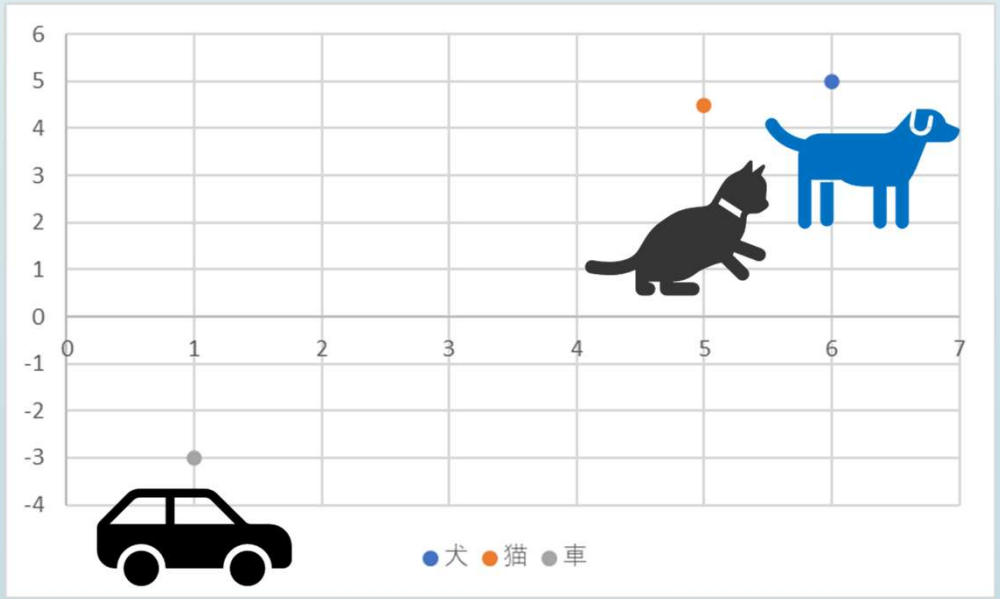
5. アプローチ①：埋め込み表現 + コサイン類似度

埋め込み表現とは？

埋め込み表現（embedding）はテキストを数値ベクトルに変換する手法の一つです。これにより、機械的に扱いやすくなります。

埋め込み表現では意味合いに応じてベクトルが決まります。

例えば「犬」、「猫」、「車」を[x,y]の形で表現すると、犬と猫は動物であり近くに、対して車は機械であり遠い位置になります。



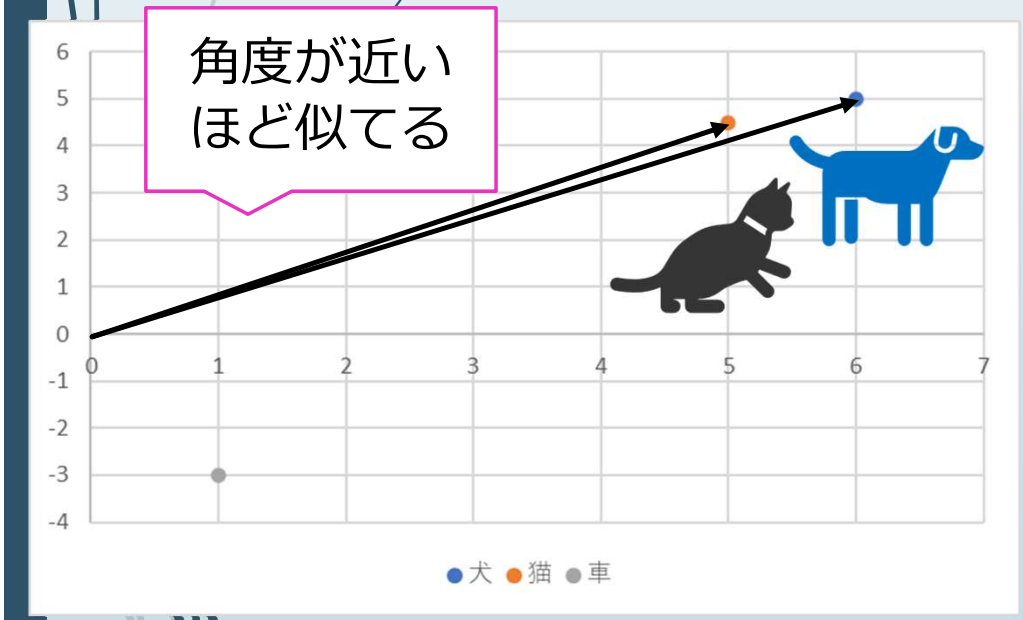
5. アプローチ①：埋め込み表現 + コサイン類似度

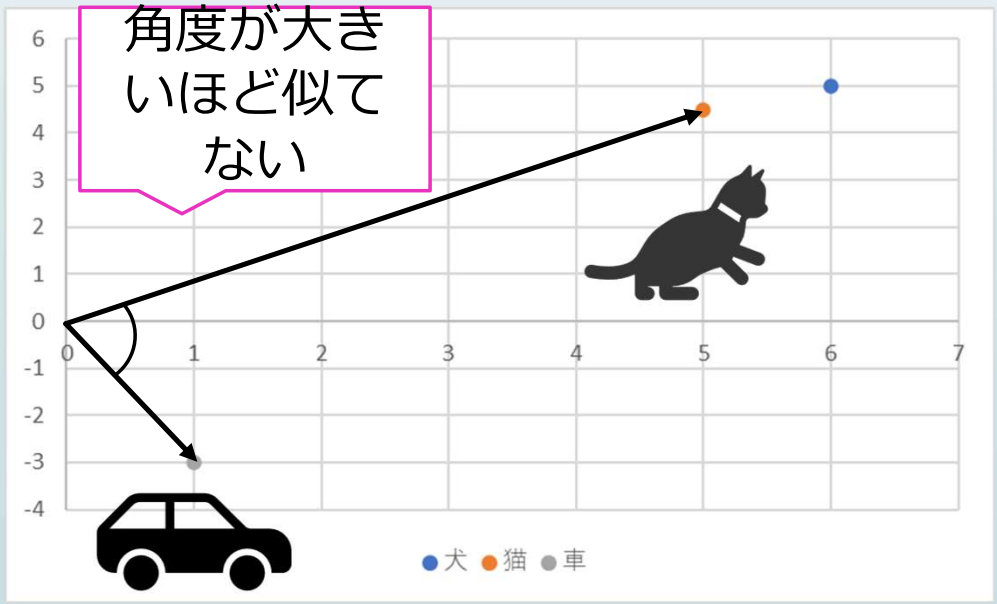
コサイン類似度とは？

先ほどのベクトルの類似性を測る物差しです。

以下図の様に角度が近いほど１に近づき、似ている事を示す。

逆に確度が大きいほど-1に近づき、似ていない事を示す。





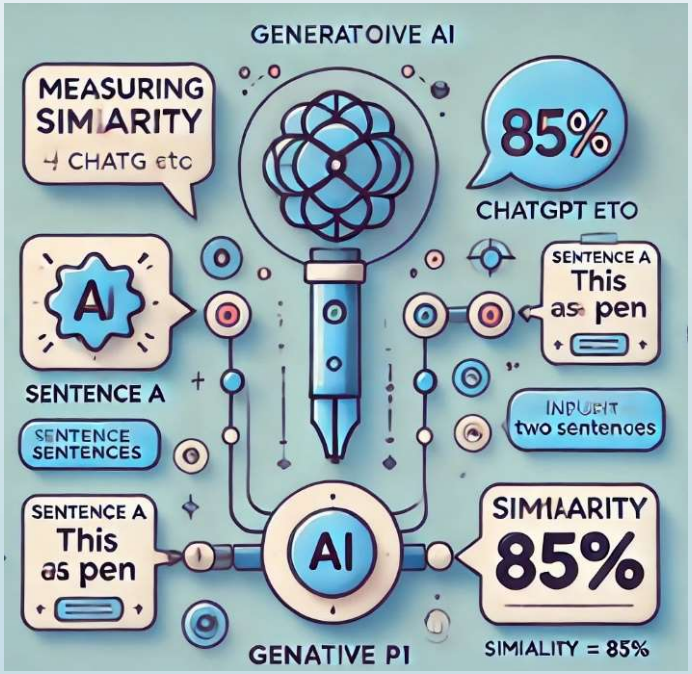
6. アプローチ②：生成AIによる類似度判定

生成AIに数値化させる

「二つの文章の類似度を0～100で評価して」等とChatGPT等に指示する方法。

プロンプトで判断基準を指示したり、生成AIが持つ知識で文脈や背景を推論して判定してくれることを期待した。

特に公式情報としてこの様な使い方は示されていないが、本研究ではできるであろうとの仮説を元に実験、検証を行った。



7. 実験概要①実験データ

実験データ

- IT関連文章で実施
- ２つのペアとなる文章を複数準備
- ２文が「同じ意味」か「別の意味」かを人間がラベリング
- 研究課題２向けに短文ペアを用意し，文意を変えない文章を両文に追加し長文化，短文ペアと長文ペアの傾向を確認し，同様にラベリング

[短文A]　[短文B]
↓（ラベリング：同じ意味 or 別の意味）
[長文化：短文A' + 共通文]　[長文化：短文B' + 共通文]
↓（再ラベリング）
[短文ペアと長文ペアの比較]

長文の文字数は定義しておらず、元文章から、
1.5倍以上の文章量を目安に作成

8. 実験概要②実験方法・評価方法

当初はGINZAという埋め込み表現のコサイン類似度も試行

実験方法

- ４分類のモデルで類似度算出
  - **text-Embedding-small**（コサイン類似度）：廉価版
  - **text-Embedding-large**（コサイン類似度）：高機能版
  - **ChatGPT独立類似度**：単純に２文の類似度を算出
  - **ChatGPT比較類似度**：文章構成の類似度と意味合いの類似度を算出

評価方法

平均，標準偏差，AUC, ROC曲線, 箱ひげ図で評価

9. 実験概要③プロンプト

- ChatGPT独立類似度

#命令:以下の条件で入力文1と入力文2を比較して意味合い的な類似度をそれぞれ評価してください。
#制約条件:・意味合い的類似度は1～100のスコアで示すこと・評価の根拠を簡潔に記載すること・入力文1と入力文2は出力しない
#入力文1:{text1} #入力文2:{text2}
#出力文:スコア、根拠

- ChatGPT比較類似度

#命令:以下の条件で入力文1と入力文2を比較して文章構造の類似度と意味合い的な類似度をそれぞれ評価してください。
#制約条件:・文章構造の類似度と意味合い的類似度は1～100のスコアで示すこと・それぞれの評価の根拠を簡潔に記載すること・入力文1と入力文2は出力しない
#入力文1:{text1} #入力文2:{text2}
#出力文:スコア、根拠

10. 実験結果と考察（研究課題1）

平均，標準偏差

同じ，別の数値の平均の差：ChatGPT系の方が大きく出る⇒判断しやすい

同じ場合の標準偏差：ばらつきが小さい方が良い

⇒ChatGPT比較類似度がばらつきが大きい

表1-実験結果の平均，標準偏差

人間の評価	text-embedding-3-small(100倍)	text-embedding-3-large(100倍)	ChatGPT比較類似度	ChatGP類似度
平均（同じ）	94.4	94.6	93.3	93.6
平均（別）	83.7	81.1	73.4	74.8
平均（同じ）-（別）	10.7	13.5	19.9	18.8
標準偏差（同じ）	5.0	4.1	11.4	4.6
標準偏差（別）	12.6	13.5	20.4	19.0

「別」はデータ自体が全く異なるものから似ている部分があるものまで様々なため、標準偏差が多くとも納得感がある。

10. 実験結果と考察①研究課題1

ROC曲線　分類モデルの性能を視覚化するための曲線

➡ROC曲線：縦軸TPRは見逃さない性能，横軸FPRは誤認の度合い

➡TPR（真陽性率）：

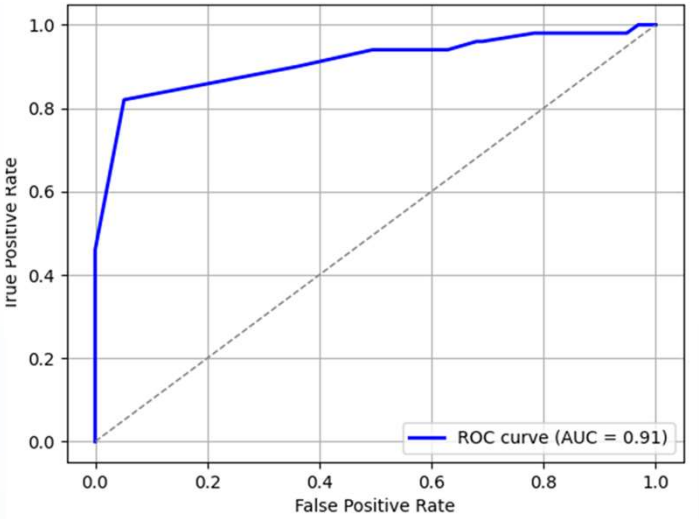
人間が同じと定義したデータのうち

「同じ」と判断できた割合

➡FPR（偽陽性率）：

人間が別と定義したデータのうち，

「同じ」と判断した割合



7-ROC曲線(ChatGPT意味合い的類似度)

➡閾値をどこにとるかによってFPR，TPRが変化する.

AUC　分類モデルの識別能力を表す指標

➡AUC：数値が高い程，モデルの性能が高い，

➡ROC図の線の下の面積を表す

10. 実験結果と考察（研究課題1）

ROC曲線

- ４モデルの結果は下図の通り
- text-Embedding-small、text-Embedding-large、ChatGPT独立類似度、ChatGPT比較類似度

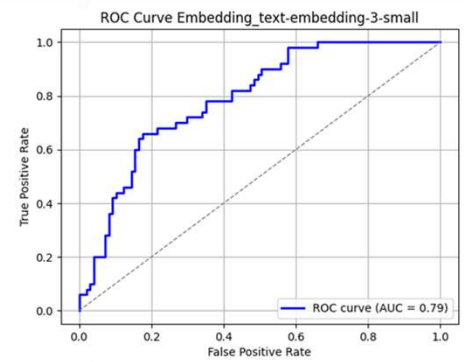


図5-ROC曲線(text-embedding-3-small)

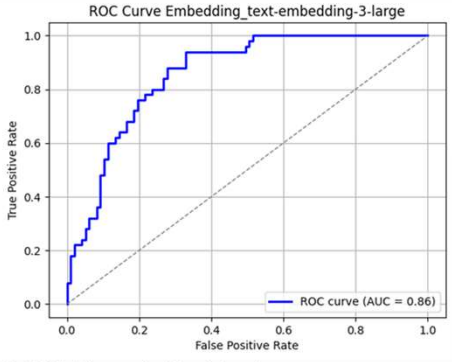
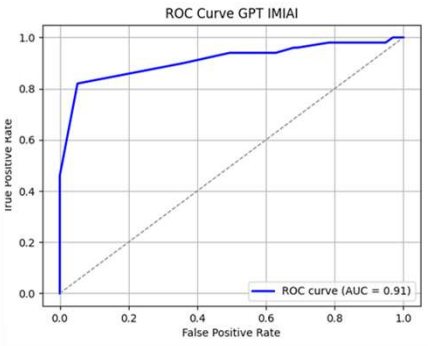


図6-ROC曲線(text-embedding-3-large)



7-ROC曲線(ChatGPT意味合い的類似度)

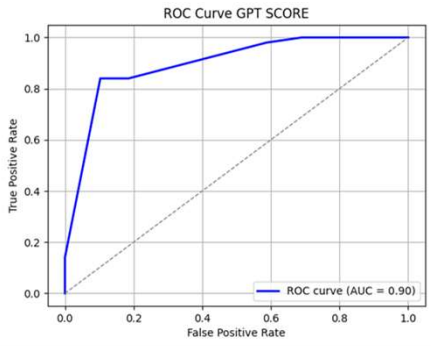


図8-ROC曲線(ChatGPTスコア)

AUC

➡全体的にChatGPT系モデルの方が高く、
ChatGPT比較類似度が0.91で最も良好な結果

表2-各モデルのAUC

モデル名	AUC
text-embedding-3-small	0.79
text-embedding-3-large	0.86
ChatGPT比較類似度	0.91
ChatGPT独立類似度	0.90

11. 実験結果と考察②研究課題2

- 長文から短文の類似度を
  減算した値を箱ひげ図で記載

- 長文化により，全モデルで類似度が上昇
- ChatGPT系は特に影響が大きかった
- 長文は類似度が高くなる傾向にある.

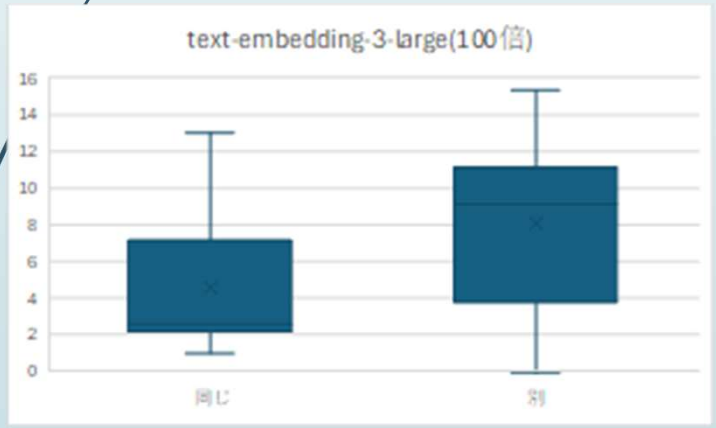


図10-長文から短文の類似度を減算した際の同じ，別の分布
text-embedding-3-large（100倍）

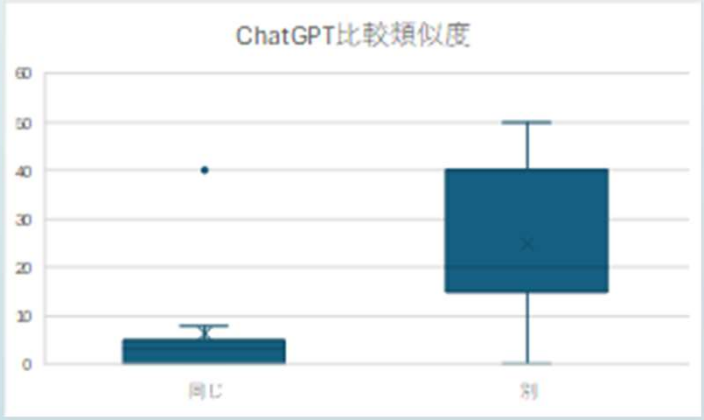


図11-長文から短文の類似度を減算した際の同じ，別の分布
ChatGPT比較類似度

12. 研究課題の結果

研究課題に対する結果は下記の通りです.

類似度の数値化方法

- **研究課題1:** 文章の類似度の数値化と人間の感覚に合うか

⇒数値化でき，その数値が人間の感覚ともある程度合うものだった.

- **課題1-1:** 埋め込み表現のコサイン類似度評価
- **課題1-2:** 生成AI(ChatGPT)による類似度評価

⇒両者で評価すると生成AIの方が人間の感覚に合うものだった

長文化の影響

- **研究課題2:** 長文の場合、類似度の検出性能がどのように影響を受けるかの評価

⇒全体的に類似度が高く評価され，検出性能が低下した.

対策としては，長い文章をそのまま評価せず，短い単位に分割して判定するなど工夫が必要

13. まとめ

類似度評価の結果

- ChatGPTによる類似度評価は、埋め込み表現よりもAUCにおいて優れた識別性能を示した
- 「同じ」と「別」の平均スコア差が大きく、**判断のしやすさ**も確認された
- ChatGPT比較類似度は「同じ」と判定されたデータのばらつき（標準偏差）がやや大きい。他モデルは比較的安定していた点には注意が必要

長文化による影響

- 長文になると類似度スコアが上昇しやすく、意味の違いを見逃すリスクがある
- 実運用では、文を適切な粒度で分割して評価する工夫が求められる

業務適用に向けて

- 大量テキストの“初期フィルタ”として、同義文判定は有効活用できる可能性がある
- 実際の業務では、目的や許容リスクに応じた閾値設定のチューニングが重要となる

14. 今後の展望

課題解決に向けて

- 独自の名詞や動詞を含む文章では精度が変動する可能性がある
  - 業務における対象のインプット文のバリエーションを定義した上で、精度差異を検証する。
  - 専門性の高い分野（社内規定や法令）に特化させた生成AIの検証を行う
- 長文化したデータでは類似度が上昇しやすい傾向がある
  - プロンプト改善（両文の主旨を要約した上で追加判定する等）
  - より新しい生成AIモデル、他のモデルでの検証
- 生成 AI 出力と人間判断の類似性（精度）
  - 安定性向上のため、プロンプトの改良に取り組む

謝辞

- 本研究を進めるにあたり，
  学びと有意義な議論の場をご提供頂いた一般財団法人日本科学技術連盟に心より感謝申し上げます．

END

- ご清聴ありがとうございました

想定問答）埋め込み表現のコサイン類似度とChat-GPT使い分け例

➨Chat-GPTの持つ知識を元にした推論が見られる例

例えば「春はあけぼの」と「春は早朝が美しい」を評価すると前述の埋め込み表現と比べ以下の様な差異が生じた。

➨
埋め込み表現による評価: 0.60
Chat-GPTによる評価:90

➨枕草子の引用であることを理由として挙げている。（Chat-GPTが持つ知識の活用）

Chat-GPTが考える理由：

➨二つの文章は意味的に似ています。

➨『春はあけぼの』は清少納言の『枕草子』の一節で、「あけぼの」とは夜明けや早朝を指し、「春の早朝が最も趣がある（美しい）」という意味です。

➨『春は早朝が美しい』も同様に「春の早朝の美しさ」を述べていますので、内容的には非常に近く、類似性が高いと言えます。