女子大学生の悩み相談相手としての生成 AI の比較評価

Comparative Evaluation of Generative AI as a Supportive Listener for Female University Students' Concerns

飯干莉緒 [1]　珠久澄香 [1]　安藤祐介[1,2,3]

Rio Iiboshi[1], Sumika Shuku[1], Yusuke Ando[1,2,3]

[1] 清泉女子大学
[1]Seisen University
[2]NPO 法人みんなのコード
[2] Specified Nonprofit Corporation Code for Everyone
[3]NPO ビジネス・ブレークスルー大学
[3] Business Breakthrough University

Abstract: 若年層を中心に生成 AI を相談相手とする利用が広がっている. 女子大学生を対象に, 5 カテゴリの悩みに対し 5 種の生成 AI モデルを用いて比較評価を行った．結果，Gemini 2.5 Pro が最高評価を得た一方，旧モデルの ChatGPT 4o は全モデル中で最低評価となり，冒頭の共感や視認性を重視していることが示された．悩み相談においては，知名度に頼らず，共感に長けた最新モデルを選択すべきであることが示唆された.

1　背景

近年，生成 AI の普及に伴い，ChatGPT 等の対話型生成 AI が若年層の生活に浸透している. 特に大学生の間では，学業や日常の課題解決だけでなく，友人や家族には話しにくい個人的な悩みを打ち明ける「相談相手」として生成 AI を利用するケースが増加傾向にある[1] . しかし, 現在市場には複数の生成 AI モデルが存在し，出力特性が異なる[2]. ユーザーは多くの場合, 知名度に基づいて AI モデルを選択しているが,「最も利用されている AI」が，必ずしも悩み相談という文脈において最も適した AI であるとは限らない. 相談相手としての適性を評価する上では, 情報の正確さ以上に，ユーザーの感情に寄り添う共感性や，会話のトーンといった定性的な要素が重要となるからである.

2　目的

本研究は，代表的な複数の生成 AI モデルを対象に, 女子大学生の悩み相談相手としてどの生成 AI が適切なのか，各モデルの回答に対する定量的なスコアリングに加え，自由記述によるアンケート調査から得られたユーザーの主観的な印象や信頼形成のプロセスを分析し比較検討することを目的とする．以下の 3 つのリサーチクエスチョン (RQ) を検討する.

RQ1：チャッピーは最良の相談相手なのか？

日本で最も利用され[3],「チャッピー」の愛称[4]でも知られる ChatGPT は，悩み相談という文脈においても最も高い評価を得るのか.

RQ2 : 相談ジャンルによる各 AI モデルの評価スコアにはどのような差異が生じるか

特定の悩みに特化した「得意分野」を持つモデルが存在するかを明らかにする.

RQ3：正解のない対話にどのような要素が好まれるか

ユーザーは AI のどのような返答を「好ましい」と判断するのか．自由記述の分析を通じ，単なる情報の正確さではなく，信頼形成につながる傾向を検討する.

3　先行研究

Z 世代を中心とした若年層において，精神的な拠

[1] 連絡先: 清泉女子大学地球市民学部
〒141-8642 東京都品川区東五反田 3-16-21
E-mail: yando@seisen-u.ac.jp

り所として利用されている．株式会社 MIXI が 2026 年に実施した「20 歳の AI 利用実態調査」[1]によると，20 歳の AI 利用経験率は 72.6%に達しており，その利用目的として「課題・レポート作成」や「検索」に次ぎ，約 7 割（67.6%）が「悩み相談やおしゃべり相手」を挙げている．

国内では，複数の AI モデルに同一設問を行い回答の正誤を測定する，客観的な性能評価が進められている．言語処理学会（2025）[2]の研究では，共通の設問セットを用いて論理的妥当性や知識の正確性が検証されている．これら「正解のある問い」への評価に対し，悩み相談のような「正解のない対話」におけるユーザー満足度や信頼形成のメカニズムについては，依然として検証の余地が残されている．

４　実験

女子大学生が抱える悩み相談の実態を反映させるため，先行する統計調査[1]を参考に，人間関係，趣味，日常のできごと，恋愛，将来への不安・人生の 5 つのカテゴリに分類される計 25 の相談シナリオを作成した．これらのシナリオに対する回答の生成には，複数の大規模言語モデルを同時並行で動作させ比較可能なツール「天秤 AI」[5]を使用した．比較対象として 5 種類の代表的な生成 AI モデルを選定し，同一のプロンプトを入力することで，計 125 件（5 モデル×5 カテゴリ×5 シナリオ）の回答データを作 成した，実験の概要は表１，シナリオの代表例は表２の通り．

表 1:実験の概要

参加者	ゼミ生　9 名(うち著者 2 名)
データ作成日	2026/1/21
合計シナリオ数	25
利用 AI モデル	GPT-4o，GPT-5.2，Gemini 2.5Pro，PLaMo2.1，DeepSeek V3(Azure)
合計返答数	125

表 2:シナリオの例

人間関係	男女の友情って成立しますか？
趣味	趣味に時間・お金を使うのは無駄なことですか？
日常の出来事	500 円硬貨が落ちていました．拾って自分の物にしても良いですか？
恋愛	彼氏って必要ですか？
将来への不安・人生	就活のエントリーシート，AI で書いて良いですか？

各評価者は，各シナリオに対する 5 つの AI モデルの回答を比較し，相談相手としての好ましさを定量的に評価した．評価は，**「0（好ましくない）」「1（どちらともいえない）」「2（好ましい）」の 3 段階評価**で行った．定量的な数値評価が完了した直後に，特定の回答に対して抱いた印象について自由記述形式による定性評価を行った．

５　結果

実験参加者が回答を採点した結果の集計結果は表３のとおりである．

表 3:採点結果の内訳

	ChatGPT 4o	ChatGPT 5.2	Gemini 2.5 Pro	Plamo 2.1	DeepSeek V3
平均	7.24	9.8	**10.24**	8.32	9.36
人間関係	7	9.2	**11**	9	10.6
趣味	8.4	9.2	9.6	9.6	**10.4**
日常のできごと	7	**10**	**10**	8.4	8.2
恋愛	7.6	9.8	**10.4**	7.6	7.8
将来への不安・人生	6.2	**10.8**	10.2	7	9.8
合計得点	181	245	**256**	208	234

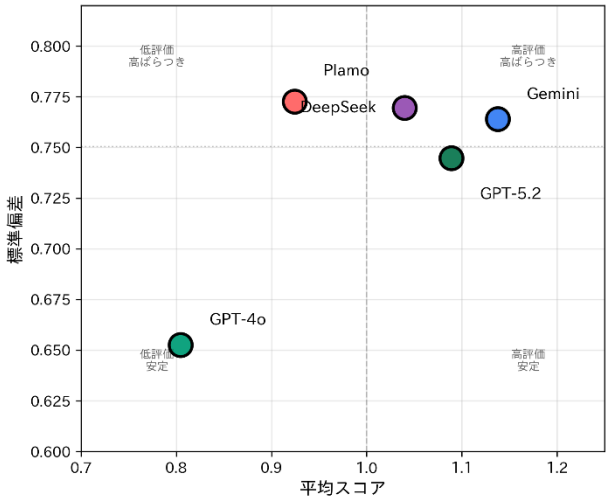

図 1:評価とばらつきの散布図

全体で最も高い評価を得たのは Gemini 2.5 Pro（合計 256 点，平均 10.24 点）であり，次いで ChatGPT 5.2（合計 245 点），DeepSeek V3（合計 234 点）の順となった．一方で，ChatGPT 4o は合計 181 点（平均 7.24 点）と，比較した 5 モデルの中で最も低い評価となった．

また，OpenAI 社のモデル間において，顕著な性能向上が確認された．旧バージョンである ChatGPT 4o が最低評価（181 点）であったのに対し，最新モデルである ChatGPT 5.2 は合計 245 点を獲得した．これは首位の Gemini 2.5 Pro（256 点）に次いで，全体 2 位の評価であり，ChatGPT 5.2 において，バージョン更新に伴う，大幅なスコアの上昇が示された．

国産モデルである PLaMo 2.1 は合計 208 点で 4 位となり，ChatGPT 4o を上回るスコアを記録した．図 1 の散布図（平均スコア vs 評価のばらつき）において，Gemini 2.5 Pro は「高評価・低ばらつき」の領域に位置した．対照的に，ChatGPT 4o は「低評価・安定（低ばらつき）」の領域に留まる結果となった．

図 2:ヒートマップ

各カテゴリにおける平均スコアと標準偏差の内訳を図 2 に示す．

- **人間関係・恋愛**: Gemini 2.5 Pro が「人間関係（1.22 ±0.77）」および「恋愛（1.16 ±0.71）」において，他モデルを上回る最高スコアを記録した．
- **将来への不安・人生**: ChatGPT 5.2 が「将来への不安・人生（1.20 ±0.73）」のカテゴリで最も高い評価を得た．
- **趣味**: DeepSeek V3 が「趣味（1.16 ±0.77）」において，Gemini に次ぐ高いスコアを記録した．一方で，散布図（図 1）において DeepSeek V3 は「高評価・高ばらつき」の領域に位置しており、回答者によって評価が大きく分かれる傾向が見られた．

自由記述の回答の中では各モデルについて下記のような言及があった．

- **Gemini 2.5 Pro**: 「『寝坊焦りますよね』など，寄り添っている感じがする」「最初に共感してくれるので，友達に相談している気分」といった，共感的な振る舞いに対する好意的な意見が多く見られた．
- **ChatGPT 5.2**: 「結論から簡潔に説明してくれるのでさっぱりしている」と評価される一方，「少し意見が強い」と感じる層も見られた．
- **ChatGPT 4o**: 「答えを曖昧に出していた．白黒はっきりさせたい人にはモヤモヤする」といった，回答の具体性や断定を避ける傾向に対する不満が見られた．
- **PLaMo 2.1**:「結論ファーストでめちゃいい」「最終的にこちらに任されている感覚が強い」という簡潔さを支持する声がある一方で，「感情に寄り添うよりは論理的」「（言い切りが強く）正しい情報か判断する前に押し切られる感覚があった」といった懸念も示された．
- **DeepSeek V3**: 「場合分けされていて読みやすい」という肯定意見と，「長たらしくて読む気にならない」という文章量への否定意見で評価が分かれた．

6 考察

RQ1:チャッピーは最良の相談相手なのか？

結果から，「最も利用されている AI が，必ずしも悩み相談という文脈において最良の相手ではない」という傾向が示唆された．Gemini 2.5 Pro が「高評価・低ばらつき」の領域に位置したことは，多くのユーザーに対して安定して質の高い対話を提供できていることを意味している．

同一開発元（OpenAI 社）のモデル間で見られた劇的な性能差から，生成 AI の進化が単なる知識量だけでなく，悩み相談に不可欠な「対話の自然さ」や「ユーザーへの寄り添い」の面でも急速に進んでいることがわかる．常にアップデートされた最新のモデルを選択することこそが，悩み相談という正解のない対話においても恩恵があることが示唆される．

また，国産モデルの PLaMo 2.1 が ChatGPT 4o を上回った点については，悩み相談のような文脈依存性が高いタスクにおいて，国内で開発されたことによる設計や学習データの選択における文化的な親和性が，ユーザーの評価に影響を与えている．

RQ2 : 相談ジャンルによる各 AI モデルの評価スコアにはどのような差異が生じるか

Gemini 2.5 Pro は，感情的な揺らぎが生じやすい情緒的カテゴリにおいて，標準偏差が比較的小さく抑えられている．これは相談者の個人差に左右されず，安定して「共感的な対話」を提供できていることを示唆している．ChatGPT 5.2 が将来相談で高評価を得た要因としては，感情的なケア以上に具体的なキャリアパスや解決策の提示が求められる場面において，

その論理構成力が肯定的に作用したと考えられる．また，DeepSeek V3 の「趣味」における評価の二極化については，自由記述に見られる「長文で丁寧」という賛辞と「長すぎて読む気がしない」という批判の混在からも裏付けられる．趣味の領域では，情報の詳細さを求める層とライトな対話を求める層で，AI に求める情報量に大きな隔たりがあることが推察される．

RQ3：正解のない対話にどのような要素が好まれるか

自由記述の結果から，女子大学生が「正解のない対話」において重視するのは，解決の早さよりも「心の距離感」と「ストレスのない視認性」であると言える．Gemini 2.5 Pro が高評価を得た要因は，回答冒頭に「それは大変でしたね」といった労いの言葉を配置する「共感ファースト」の構造にある．自由記述に見られる「寄り添ってくれる」「安心感がある」という評価のように，正論の前にまず感情を受容する姿勢が信頼形成の決定的な要因の一つとなった．対照的に，即座に解決策を提示する ChatGPT 4o は「機械的」「冷たい」と評価され，心理的な乖離を生んだ．

PLaMo 2.1 や ChatGPT 5.2 の分析からは，回答のトーンによる受容性の違いが示唆された．PLaMo 2.1 のような「結論ファースト」や「言い切り」の強い回答は，悩み相談においては「視野が狭い」「押し切られる」といった威圧感を与えかねない．女子大学生は AI に答えを断定されることよりも，Gemini のように選択肢を示しつつ「一緒に考えてくれる」という伴走的なスタンスを好む傾向があるといえる．

また，回答の「長さ」も重要な要素である．DeepSeek V3 のように情報量が過多なモデルは，詳細さが評価される一方で「長すぎて読む気がしない」と忌避される傾向も見られた．「適度な文章量」は，日常的な相談相手として受け入れられるために重要であることが推察される．

7　おわりに

本研究では，女子大学生の悩み相談における生成 AI の受容性を明らかにするため，5 つのモデルを用いた比較実験を行い，カテゴリごとに分析した．その結果，Gemini 2.5 Pro が首位（256 点）を獲得した．今回の研究においては，知名度や利用者の多さが必ずしも悩み相談の実用性と比例しないこと，そして現時点では Gemini 2.5 Pro が女子大学生の相談相手としてより優れた性能を持っていることが示された．

この結論は，AI の実力を格付けする世界的な指標「LMSYS Chatbot Arena[6]」の日本語向けランキングにおいて，Gemini や最新の ChatGPT 5.2 が上位を占めている現状とも整合している．情緒的な配慮が必要な相談には Gemini を，論理的な整理には最新の ChatGPT 5.2 を使い分けることで，満足度がさらに向上する可能性がある．以上の点から，実利用においては単なる知名度のみで判断せず，最新のモデルを使用することや対話のトーンを考慮してモデルを選択することが，生成 AI を良き相談相手とするために必要であると言える．

8　今後の展望と課題

本実験では，評価参加者が筆者らの所属するゼミ生を中心とした少人数であったため，参加者を増やし属性を広げることで，より一般的な知見が得られると考えられる．また，本実験では「天秤 AI」を用いてモデル本来の応答を比較したが，実利用されている ChatGPT 等にはメモリ機能やカスタム指示が含まれるため，ユーザーごとにカスタマイズされた環境下においても同様のモデル間差異が見られるかを検証する必要がある．さらに，今回は一回の対話評価であったが，長期的な継続利用における信頼関係の構築過程についても，今後の調査課題としたい．

9　参考文献

[1] 株式会社 MIXI: 20 歳の AI 利用実態調査, PR TIMES, 2026 年 1 月 7 日, https://prtimes.jp/main/html/rd/p/000000754.000025121.html, (参照日 2026-02-13)

[2] Namgi Han, 岡本拓己, 石田茂樹, Akim Mousterou, Bowen Chen, 林俊宏, 宮尾祐介: オープン日本語 LLM リーダーボードの構築と評価結果の分析, 言語処理学会 第 31 回年次大会 発表論文集,(2025 年 3 月)

[3] マクロミル: 生成 AI 利用に関する意識調査, 2026 年 1 月 7 日, https://www.macromill.com/press/release/20260107.html, (参照日 2026-02-13)

[4] 日本経済新聞,「古古古米」や「チャッピー」流行語大賞候補 30 語 , 2025 年 11 月 6 日 , https://www.nikkei.com/article/DGXZQOUD055EV0V01C25A1000000/ , (参照日 2026-02-13)

[5] GMO インターネットグループ: 天秤 AI by GMO, https://tenbin.ai/workspace, (参照日 2026-02-13)

[6] LMSYS Org: Chatbot Arena Leaderboard (Japanese), https://chatbot-arena.apps.llmc.nii.ac.jp/, (参照日 2026-02-13)