ユーザ感情に整合するLLM応答方針の設計に向けて：VAD制御とReasoning量がLLMの感情理解・生成能力に与える影響

小野 聡† 持田 航†† Xin Fan††† 山名 早人††††

† 早稲田大学基幹理工学部 〒169–0072 東京都新宿区大久保3丁目4–1
††,††† 早稲田大学大学院基幹理工学研究科 〒169–0072 東京都新宿区大久保3丁目4–1
†††† 早稲田大学理工学術院 〒169–0072 東京都新宿区大久保3丁目4–1
E-mail: †,††,†††{s.ono,wmochida,yamana}@yama.info.waseda.ac.jp, †††fan_xin@fuji.waseda.jp

あらまし 生成AIは医療相談・教育など情動応答が成果に影響する領域で利用が拡大しており，感情能力の定量評価が求められる．先行研究では，感情をVAD（V：快／不快・A：覚醒度・D：支配性）で表すと，生成AIの感情理解はVは高相関である一方，A・Dは0.5以下と十分でないことが示されている．この要因として，入力文脈の解釈が浅いことがA・Dの推定に影響している可能性がある．そこで本稿では，感情能力を入力文からVADを推定する「感情理解」と，指定した目標VADに整合する応答文を生成する「感情生成」の2課題に分解して定義し，推論の深さが性能に与える影響を検証する．具体的にはGPT-5 miniおよびGPT-4.1を対象に，CoT（Chain-of-Thought）指示やパラメタ指示により推論深度（Reasoning量）を増減させ，EmoBankで正解VADとの相関（Pearson）を評価した．結果，理解はVが0.76以上である一方A・Dは0.5以下であった．生成では数値／段階指示が感情語指示より高相関であり，Reasoning量増加による改善は両課題で小さかった．

キーワード LLM, 感情理解, 感情生成, Reasoning, VAD

1 はじめに

大規模言語モデル（LLM; Large Language Model）は医療相談，メンタルヘルス支援，教育，カスタマーサポートで利用が増えている[1], [2]．各領域では事実に基づく情報提供に加え，利用者の感情状態に沿う応答が求められ，感情トーンは不安の緩和，学習意欲，サービス継続利用に影響する[3]．

LLMの感情能力は，入力から感情状態を推定する感情理解と，指定した感情状態に整合する文を生成する感情生成に分けられる．両者を別尺度で評価するとモデル間比較が難しくなるため，同一尺度で比較できる評価枠組みが必要である．

本稿は感情理解と感情生成を同一尺度で比較するため，感情をValence–Arousal–Dominance（VAD）の3軸で連続的に表すVADモデル[4]を評価枠組みとして採用する．しかしEmoBank[5]を用いた先行研究では，Valenceは高相関で推定できる一方，ArousalとDominanceは0.5未満と報告されている[6], [7]．Arousalは切迫度など状況的強度，Dominanceは立場関係や制御感に依存し，語彙の表層からは読み取りにくい．このためA/Dの低精度は，LLMの感情能力評価において文脈解釈の深さがボトルネックになり得るという問題を示す．

近年，CoT（Chain-of-Thought）など中間推論を促す手法や推論量（Reasoning量）の制御により，数学・論理タスクの性能が変化することが報告されている[8], [9]．一方で，理解と生成を同一枠組みで定量評価し，プロンプトによる感情条件の提示形式と推論深度の寄与を同時に比較する体系的検討は十分ではない．そこで本稿は，推論深度（CoTの有無およびReasoning量）を操作し，感情理解（Text→VAD）と感情生成（VAD→Text）の双方における性能，とくにArousal/Dominanceを含むVAD空間上の誤差・追従性への影響を定量的に検証する．

本稿では文レベルVAD注釈を持つEmoBank[5]を用い，GPT-5 miniとGPT-4.1の感情理解・感情生成を比較する．感情理解は出力VADと正解VADのPearson相関で評価する．感情生成では，生成文のVADは直接観測できないため，生成文を独立VAD回帰評価器でVADへ写像し，推定VADと目標VAD（または感情条件）のPearson相関で評価する．追従性は目標VADへの整合性として評価し，CoTの有無，Reasoning量，感情条件の提示形式（数値／段階（Very Low/Low/Moderate/High/Very High）／感情語）を操作する．

以上の背景から，本稿では以下の研究課題（Research Questions）を設定する．

- RQ1：LLMにおける「感情理解（テキスト→VAD推定）」と「感情生成（VAD指定→テキスト生成）」の性能差（理解・生成ギャップ）は，VAD空間上でどの程度生じるか．
- RQ2：感情生成において，(i) VADを数値／段階で指定する指示と (ii) 感情語で指定する指示のどちらが，同一条件下で目標VADへの追従性を高く達成できるか．
- RQ3：推論深度（CoTの有無や推論ステップ指示の強さ）を増加させることは，感情理解・生成の性能を改善するか．

本稿の貢献は次の3点である．

- VAD空間で感情理解と感情生成を統一評価し，性能の非対称を定量化した．
- 感情生成において，数値指示／段階指示／感情語指示の差を比較し，追従性が高い条件を示した．

- 推論量と CoT の効果を比較し，感情タスクにおける推論深度の寄与を整理した．

本稿の構成は次の通りである．2 節で背景，3 節で関連研究，4 節で手法，5 節で実験，6 節で結果，7 節で考察，8 節で結論を述べる．

2 背景知識

本節では感情表現（VAD）と推論深度（CoT・推論量制御）を整理する．

2.1 感情の表現形式と VAD 空間

感情表現は，感情を「喜び」「悲しみ」などのカテゴリーへ割り当てる離散的表現と，多次元連続空間上の点として表す連続的表現に分けられる．離散的表現は解釈しやすいが，中間状態や複数感情の混合を扱いにくい．一方，連続的表現は複雑な感情を表現できる．

本稿は連続的表現として Valence–Arousal–Dominance（VAD）モデル [4] を採用する．VAD は 3 軸で感情状態を表す．各軸の意味は次のとおりである．

- Valence（快／不快）：快・不快（高いほど快）．
- Arousal（覚醒度）：興奮・沈静（高いほど興奮）．
- Dominance（支配性）：制御感・優位性（高いほど優位）．

VAD は感情強度を同一尺度で比較できる．本稿では VAD 空間で理解（Text→VAD）と生成（VAD→Text）を統一評価する．感情生成では目標 VAD を条件として文を生成する．

2.2 Chain-of-Thought（CoT）

CoT は中間推論を段階的に出力させて解答を導くプロンプト手法である．「ステップごとに考える」指示で問題を分割して検討させる．先行研究では，数学・論理推論課題で性能向上が報告されている． [8].

2.3 Reasoning model と推論深度の操作

Reasoning model は最終回答の前に内部推論表現を用いる設計であり，推論深度は計算量により変化する．商用 LLM では推論量制御が提供され，GPT-5 以降の GPT モデルは API の reasoning_effort で推論の度合いを調整できる．本稿は CoT と推論量の操作が感情理解と感情生成に与える影響を検証する．

3 関連研究

本節は RQ1–RQ3 に沿い，LLM 感情能力評価の先行研究（VAD 理解／生成評価，提示形式，推論深度）を整理する．

3.1 感情表現と VAD（連続次元表現）

VAD は感情強度や近接感情の差異を連続的に扱えるが，状況的切迫と生理的強度に依存する A は，立場関係と自己効力感に依存する D は表層に現れにくく推定が難しい．文レベル VAD を提供する代表的なデータセットとして，EmoBank がある [5]．本稿では，EmoBank を LLM の VAD 理解・生成能力の評価に用いる．

表 1 EmoBank における VAD 推定の Pearson 相関（教師あり・既存研究） [7]

モデル（設定）	Valence	Arousal	Dominance
BERTL (EB←AIT)	0.765	0.583	0.416
VADR	0.821	0.553	0.493
VADEC (SenWave)	0.823	0.553	0.485

表 2 AEB-2（zero-shot）に含まれる EmoBank 回帰タスクでの Pearson 相関 [6]

モデル	Valence	Arousal	Dominance
ChatGPT	0.554	0.320	-0.121
GPT-4	0.723	0.364	0.193

3.2 VAD に基づく感情理解評価（テキスト →VAD 推定）

感情理解は入力文の VAD を推定し，正解 VAD との一致度（相関）で評価する．Mukherjee らは EmoBank に対する教師あり回帰で V の高相関・A/D の低相関傾向を示した [7]．Liu らは affective evaluation benchmark（AEB）を構築し，AEB-2 の EmoBank 回帰を zero-shot 評価し，GPT-4 で V=0.723, A=0.364，D=0.193 を報告した [6]．zero-shot 値は手順に依存するため，教師あり結果と単純比較には設定の明示が必要である [6]．代表値を表 1 と表 2 に示す．

先行研究は，V に比べ A/D が低相関になりやすい傾向を示した [6], [7]．特に，GPT-4 は回帰モデルである Roberta に比べ A/D の相関が低い傾向がある．ただし，理解を主対象とする研究が多く，生成を VAD 空間上で定量評価し，理解と同一指標で比較する設計や推論深度操作の検討は少ない．

3.3 感情生成・感情制御と評価（条件 → テキスト）

LLM 研究では連続条件を与え，生成文を独立評価器で感情空間へ写像し整合を評価する研究がある [10]．Ishikawa らは Russell の円環モデル（VA）で valence/arousal の連続条件を与え，GoEmotions [11] 学習モデルで評価して整合性を測定した [10]．一方で評価器（人手評価／分類器／回帰器／ LLM 評価）と感情空間（離散／ VA ／ VAD）が研究ごとに異なり，横断比較が難しい．追従性差の議論は，生成能力と評価器・写像特性を切り分ける必要がある．本稿（RQ1）は理解（出力 VAD と正解 VAD）と生成（推定 VAD と目標 VAD）の一致度を同一指標で比較する．生成では推定 VAD を得るために独立 VAD 回帰評価器（EmoBank 学習）を固定する．

3.4 感情条件の提示形式（数値／段階／感情語）

Choudhury らは条件付き生成におけるモデル表現と人間期待のギャップを，人手判断で定量化する Representation Alignment を提案し，Words，VAD（Lexical/Numeric），Emoji を比較した [12].

表 3 で示すように，Choudhury らは，Words 条件が高整合で，Numeric より Lexical VAD が合意を改善すると示した [12]．Lexical VAD は Numeric を段階語（Very Low/Low/Moderate/High/Very High）へ写像する．Representation Alignment は人間期待との整合を扱い，本稿の追従性を同一評価器・指標で比較する設計とは異なる．本稿（RQ2）

表 3 条件提示形式と Representation Alignment（人手評価）[12]

条件	Self-Alignment (%) (GPT-4)	Self-Alignment (%) (LLaMA-3)	Entropy (GPT-4)	Entropy (LLaMA-3)
Words(単語)	61.9	57.5	0.32	0.42
Lexical(段階的)	52.0	31.2	0.61	0.72
Numeric(数値的)	—	—	0.70	0.63
Emoji(絵文字)	29.7	—	0.67	0.52

は数値／段階／感情語指示を同一 VAD 評価器で比較し，提示形式の影響を定量化する．

3.5 推論深度（CoT・推論量制御）と感情タスク

感情領域では，Emotional Chain-of-Thought（ECoT）による生成改善が報告されている [13]．一方，理解と生成を VAD 空間で同一評価枠組みに揃え，推論深度操作の効果を比較した研究は少ない．本稿（RQ3）は CoT と推論量指示を操作し，理解・生成の効果を同一評価枠組みで比較する．

3.6 本稿の位置づけ

VAD 推定としての感情理解評価は蓄積があり，A/D が難しい傾向が報告されている [6], [7]．一方，感情生成研究は条件表現と評価器が多様で横断比較が難しい [10]．提示形式（RQ2）と推論深度（RQ3）を含め，理解・生成を VAD 空間で同一指標により一括比較する設計は標準化されていない [12], [13]．本稿は EmoBank 正解 VAD を目標とし，生成では独立回帰評価器を固定したうえで，理解・生成を同一尺度（相関）で評価し，RQ1–RQ3 の寄与を定量比較する．

4 手　　法

本節では，感情理解（Text→VAD）と感情生成（VAD→Text）を VAD 空間で統一評価する枠組みを示す．

4.1 統一的な評価枠組み

本項は入出力形式と評価手順を定義する．感情は VAD（Valence-Arousal-Dominance）の 3 次元ベクトル $\mathbf{y} = (v, a, d)$ とする．v は Valence，a は Arousal，d は Dominance を表す．入力テキストを x，生成テキストを $\hat{x}$ とする．感情理解は x から推定値 $\hat{\mathbf{y}}$ を出力する処理である．感情生成は目標条件 $\mathbf{y}$ から $\hat{x}$ を生成する処理である．生成文 $\hat{x}$ の VAD は直接観測できないため，Text→VAD 評価器 E により $\tilde{\mathbf{y}} = E(\hat{x})$ を推定する．これにより，感情生成においても目標 $\mathbf{y}$ と推定値 $\tilde{\mathbf{y}}$ の一致度を同一尺度で定量化できる．

本稿は一致度指標として Pearson 相関係数を用い，各次元（V/A/D）で算出する．

- 感情理解：$\hat{\mathbf{y}}$ と正解 $\mathbf{y}$ の相関
- 感情生成：$\tilde{\mathbf{y}}$ と目標 $\mathbf{y}$ の相関

本稿では V/A/D の各次元ごとに相関を算出し，次元別に報告する．正解データとして，EmoBank の各文に対応した VAD を用いる．

生成文の人手評価は英語文から VAD への変換を要するため，本稿では Text→VAD 評価器 E による自動評価を採用する．この方式は条件間比較の再現性を担保する一方，評価器の推定誤差やバイアスの影響を受ける．

そこで本稿では評価器 E を固定し，条件間の相対比較に限定する．

4.2 感情理解タスク（Emotional Understanding）

4.2.1 タスク定義

感情理解は，EmoBank の文 x から VAD（Valence, Arousal, Dominance）である $\hat{\mathbf{y}}$ を推定し，推定値 $\hat{\mathbf{y}}$ と正解値 $\mathbf{y}$ の一致度を各次元で評価する．

4.2.2 出 力 形 式

感情理解では，v, a, d は 1.00–5.00 の実数とし，小数点以下 2 桁に統一する．

4.3 感情生成タスク（Emotional Generation）

本項は感情生成タスクを定義する．感情生成は VAD 条件を入力とし，英語 1 文を生成する．

4.3.1 タスク定義

感情生成は，目標 $\mathbf{y}$ から英語 1 文 $\hat{x}$ を生成し，$\tilde{\mathbf{y}} = E(\hat{x})$ を用いて $\mathbf{y}$ と $\tilde{\mathbf{y}}$ の一致度を各次元で評価する．

4.3.2 生成評価器 E（Text→VAD 評価器）

評価器 E は入力文から VAD を推定する教師あり回帰モデルである．ベースモデルは 事前学習モデルである RoBERTa-large [14] とし，Valence/Arousal/Dominance を同時回帰する．目的関数は推定 VAD と正解 VAD の平均二乗誤差（MSE）の最小化である．学習データは EmoBank の train/valid に加え，VAD 極端値を含む追加テキスト（LLM 生成）を併用する．本稿は複数手法で学習した評価器の推定値を次元別中央値で統合し，統合モデルを E とする．

4.3.3 出 力 形 式

感情生成では，評価対象の LLM は英語 1 文のみを出力し，生成文を評価器 E により 1.00–5.00（小数点以下 2 桁）の VAD 数値に変換する．

5 実　　験

本節では，実験設定，感情生成評価器の選定を述べる．

5.1 実 験 設 定

本項では，比較対象となる条件と実行手順を整理する．感情理解・感情生成の両タスクにおいて同一の評価枠組み（同一指標・同一後処理）を用い，推論深度および条件提示形式の影響を同一の枠組みで比較する．感情理解は 6 条件，感情生成は 6 条件 ×3 形式を比較する．比較条件を以下に示す．

- 推論深度
  - GPT-5-mini: `reasoning_effort` を {minimal, low, medium, high} の 4 条件で比較する．
  - GPT-4.1: 推論誘導なし／推論誘導あり（CoT）を比較する．
- 指示形式（生成）: Numeric，Lexical，Word の 3 形式を

表 4 API パラメータ（代表値，未指定は API 既定値）

タスク	モデル	temperature	max_tokens	実行回数
感情理解	GPT-5-mini	—	500	EmoBank test 8,000
感情理解	GPT-4.1	—	1,000	EmoBank test 8,000
感情生成	GPT-5-mini	—	1,400–6,500	$1,728 \times 3$
感情生成	GPT-4.1	0.0	1,000	$1,728 \times 3$

比較する．

5.1.1 VAD の値域と後処理

VAD の各次元の値域は $[1.00, 5.00]$ に固定する．出力が値域外の場合は，各次元を $[1.00, 5.00]$ にクリップして評価に用いる．推定値の表記は小数点 2 桁に統一し，プロンプトにおいても同一の表記を採用する．

5.1.2 データセット

データセットには EmoBank [5] を用い，読み手である reader の視点の VAD を正解ラベルとする．データ分割は train/valid/test = 8000/1000/1000 の既定分割に従う [5].

5.1.3 内部推論と最終出力の分離

内部推論を促す条件でも，最終出力は出力形式制約（感情理解: JSON，感情生成: 英語 1 文）に限定する．加えて，設定は推論深度の操作を外部に明示的に出力させるのではなく，内部計算に限定し，出力形式への混入を避ける．

5.1.4 API パラメータ

再現性の確保のため，主要な API パラメータを整理する．表 4 に代表的な設定を示す（未指定の項目は API 既定値を用いる）．GPT-5-mini は temperature を指定できないため，空欄としている．また，GPT-5-mini の max_tokens は reasoning_effort に合わせて調整する．温度やサンプリング関連パラメータについて，指定可能な項目は表 4 のとおり固定し，未指定（または指定不可）の項目は実行時点の API 既定値に従う．とくに `top_p` 等は未指定とし，API 既定値を用いる．このため，再現には同等の API 仕様が前提となる．

感情理解は EmoBank test 8,000 件を入力とし，読み手である reader の視点で VAD を推定させる．感情生成は $12^3 = 1,728$ 条件を用意し，各条件につき 1 文を生成させる．

5.1.5 感情理解の抽出

感情理解の出力は JSON オブジェクトのみに制約する．抽出は以下の優先順で行う．

- 応答全体を JSON としてパースする．
- 応答文字列から `"valence"` を含む JSON 片を抽出してパースする．
- 応答から `V=...` ，`A=...` ，`D=...` 形式を正規表現で抽出する．

抽出に失敗したサンプルは欠損として扱い，相関計算から除外する．抽出後，各次元を $[1.00, 5.00]$ にクリップする．

5.1.6 感情生成の条件構成

VAD 条件は，各次元の値域を $[1.00, 5.00]$ とし，これを等間隔に 13 分割により 12 区間を得て，中点 12 個を代表値にする．分割点は

$$1.00,\ 1.33,\ 1.67,\ 2.00,\ 2.33,\ 2.67,\ 3.00,$$
$$3.33,\ 3.67,\ 4.00,\ 4.33,\ 4.67,\ 5.00$$

である．隣接する分割点の中点を代表値として用い，代表値集合 $\mathcal{R}$ を

$$\mathcal{R} = \{1.17,\ 1.50,\ 1.83,\ 2.17,\ 2.50,\ 2.83,$$
$$3.17,\ 3.50,\ 3.83,\ 4.17,\ 4.50,\ 4.83\}$$

と定義する．VAD の各次元に $\mathcal{R}$ の要素を割り当て，直積 $\mathcal{R}^3$ により全条件数の VAD 条件を生成する．条件数は $12^3 = 1,728$ とする．条件は 1 行 1 レコードの JSONL（JSON Lines）形式で保持する．

5.1.7 感情生成の条件提示形式（num / lex / word）

同一の VAD 条件を 3 形式で LLM に提示し，形式差を比較する．

- num: Valence/Arousal/Dominance を小数点 2 桁の数値で提示する（例: `V=2.15, A=4.20, D=1.85`）．
- lex: VAD を 5 段階カテゴリへ離散化し，段階表現として提示する．境界は全次元で共通とし，Very Low（$1.00 \leqq l < 1.67$），Low（$1.67 \leqq l < 2.67$），Moderate（$2.67 \leqq l < 3.34$），High（$3.34 \leqq l < 4.34$），Very High（$4.34 \leqq l \leqq 5.00$）とする．
- word: GoEmotions データセット [11] のうち，`neutral` 以外の 26 感情ラベル集合を入力として明示的に提示する（例：`labels = ["sadness", "disappointment", "disapproval"]`）．

事前計算には NRC-VAD Lexicon v2.1 を用いる [4]．各ラベル l についてセントロイド $\boldsymbol{\mu}_l$（中央値）と共分散行列 $\boldsymbol{\Sigma}_l$（ridge 付き）を推定する．入力 VAD $\mathbf{y}$ に対し，Mahalanobis 距離 $(\mathbf{y} - \boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}_l^{-1} (\mathbf{y} - \boldsymbol{\mu}_l)$ の上位 3 ラベルを割り当てる．

全形式で「英語 1 文，50 語未満」を指示する．判定は次の規則により行い，いずれかを満たさない出力は失敗として除外する：語数は空白区切りの単語数で数え 50 未満とする．1 文判定は改行を含まず，文末記号（. ! ?）の出現数が 1 以下であることとする．禁止語の判定は小文字化した生成文に対する単純な部分一致で行う．

- num: 数字または `valence/arousal/dominance`,
- lex: `Very Low/Low/Moderate/High/Very High`,
- word: 入力ラベル文字列

失敗を除外したうえで相関を算出する．

5.1.8 評 価 指 標

主要指標は Pearson 相関係数とする．実験では VAD 各次元の Pearson 相関係数を算出し，V/A/D の 3 次元の単純平均（macro）を要約指標とする．補助指標として Spearman 相関係数，RMSE，MAE，R^2 を算出する．

5.1.9 信頼区間と検定

本稿は Pearson 相関係数の 95%信頼区間をブートストラップ（$B = 400$，seed=12345）で推定する．条件差の検定は macro MAE を用いる．2 条件比較は対応のある Wilcoxon 符号付順位検定，4 条件比較は Friedman 検定を用いる．多重比較は Holm 法で補正し，効果量は $r = |z|/\sqrt{n}$ とする．信頼区間は評価器 E を固定した条件で推定し，評価器学習の不確実性は

表 5　Text→VAD 評価器の性能（ISEAR 事前学習 →EmoBank fine-tuning，EmoBank test）

設定	次元	Pearson	Spearman	R^2	RMSE
6 評価器中央値（median）	Valence	0.833	0.802	0.663	0.200
6 評価器中央値（median）	Arousal	0.554	0.477	0.249	0.213
6 評価器中央値（median）	Dominance	0.522	0.457	0.221	0.186

含めない．

5.2　感情生成評価器の選定

本項では，生成文を VAD に写像する Text→VAD 評価器の選定結果を示す．

5.2.1　評価器の学習設定

評価器は International Survey on Emotion Antecedents and Reactions (ISEAR) データセット [15] と EmoBank [5] を用い，ISEAR→EmoBank の 2 段階転移学習で学習する．本節はカテゴリ分布 →VAD を EMD で学習する枠組みを参考にする [16]．

- 第 1 段階（ISEAR）：vad→categories タスクを学習する．カテゴリは `joy`, `fear`, `anger`, `sadness`, `disgust`, `shame`, `guilt` とする．学習率は 3×10^{-5}，最大エポック数は 20 とする．損失関数は Earth Mover's Distance（EMD）を用いる．
- 第 2 段階（EmoBank）:VAD 回帰タスクを fine-tuning する．初期化では第 1 段階の重みを読み込み，回帰 head は再初期化する．エポック 0–4 は head のみを学習し（学習率 3×10^{-3}），エポック 5 以降は全層を学習させ，学習率を 5×10^{-6} に切り替える．出力は `sigmoid_1_5` 変換で 1.00–5.00 に制約する．

fine-tuning におけるデータ variant は以下の 6 条件とする．

- base: 標準 EmoBank
- base_balanced: 標準 EmoBank + VAD 重み付きサンプリング
- extreme: base + 極端データを 300 件追加
- extreme_balanced: 極端データを 300 件追加 + VAD 重み付きサンプリング
- expand_extreme: base + 極端データを 1,600 件追加
- expand_extreme_balanced: 極端データを 1,600 件追加 + VAD 重み付きサンプリング

6 条件の単体評価器と 2 統合手法（mean/median）の計 8 候補を EmoBank test で比較する．mean は次元別平均，median は次元別中央値とする．値域外の出力は $[1.00, 5.00]$ にクリップして集計に用いる．

5.2.2　選定結果

計 8 候補を比較し，相関や R^2，RMSE を総合的に判断した際に，最も良い 6 評価器中央値（median）を評価器 E に採用する．表 5 に主要候補の性能を示す．

6　結　　果

本節では，感情理解および感情生成の結果，ならびに推論トークン数の観測結果を述べる．

表 6　感情理解の Pearson 相関（EmoBank test）

設定	Valence	Arousal	Dominance
GPT-5-mini (think high)	0.773	0.486	**0.474**
GPT-5-mini (think medium)	0.776	0.494	0.469
GPT-5-mini (think low)	0.778	0.495	0.471
GPT-5-mini (think minimal)	0.760	0.485	0.387
GPT-4.1 (推論誘導なし)	**0.792**	0.479	0.447
GPT-4.1 (推論誘導あり（CoT）)	**0.792**	**0.494**	0.446

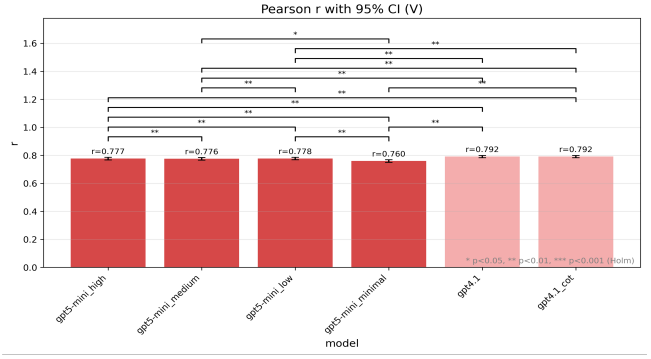

図 1　感情理解の Pearson 相関（Valence，95%CI）

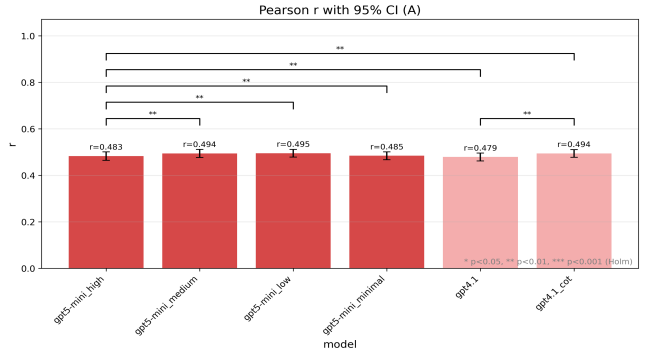

図 2　感情理解の Pearson 相関（Arousal，95%CI）

6.1　感 情 理 解

本項では，感情理解の相関結果を示す．推定値と正解値の一致度を次元別に評価し，推論深度の操作が各次元へ与える影響を確認する．表 6 に，EmoBank test における Pearson 相関係数を示す．Valence は全設定で $r \geqq 0.760$ を示す一方，Arousal は全設定で $r \leqq 0.495$，Dominance は全設定で $r \leqq 0.474$ に留まる．図 1–図 3 に，次元別の相関と 95%信頼区間を示す．図中の有意差は macro MAE に対する条件差検定の結果であり，多重比較は Holm 法で補正する（手順は実験節に従う）．ここで，図中のエラーバーは 95%信頼区間（95% CI）を表す．また，条件間に有意差がある場合はブラケットで示し，アスタリスクで Holm 補正後の p の範囲（*: $0.01 \leqq p < 0.05$，**: $0.001 \leqq p < 0.01$，***: $p < 0.001$）を示す．　以上より，次の 2 点を示す．

- 次元差:Valence は全設定で $r \geqq 0.760$ を示す一方，Arousal と Dominance は全設定で $r \leqq 0.495$ に留まる．
- 推論深度：`reasoning_effort` と CoT の操作は，相関を単調に改善しない．

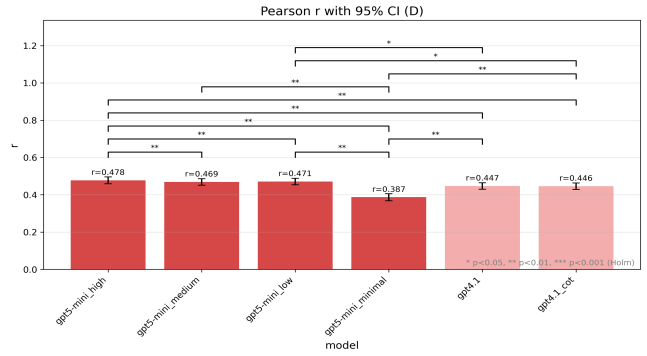

図 3　感情理解の Pearson 相関（Dominance，95%CI）

表 7　感情生成の失敗数（各形式 1,728 件，括弧内は欠損）

設定	num	lex	word	合計
GPT-5-mini (think high)	0 (0)	90 (88)	5 (0)	95 (88)
GPT-5-mini (think medium)	66 (66)	2 (2)	5 (0)	73 (68)
GPT-5-mini (think low)	20 (20)	1 (0)	3 (1)	24 (21)
GPT-5-mini (think minimal)	10 (10)	1 (1)	47 (26)	58 (37)
GPT-4.1 (推論誘導なし)	0 (0)	1 (1)	1 (0)	2 (1)
GPT-4.1 (推論誘導あり（CoT）)	0 (0)	2 (2)	0 (0)	2 (2)

6.2 感情生成

本項は，感情生成の相関結果を示す．本項の相関は固定した Text→VAD 評価器に基づく条件間の相対比較である．

表 7 に，設定別の失敗数（括弧内は欠損数）を示す．ここで失敗は出力制約違反（語数・1 文制約）または禁止語の出現を指し，欠損は API 応答の欠落や抽出不能により評価できないケースを指す．表 8 に，失敗を除外した Pearson 相関係数を示す．結果は次の 2 点を示す．

- word は Valence の相関は同程度だが，num と lex に比べ，Arousal と Dominance の相関が低い．
- 推論深度：`reasoning_effort` と CoT の操作は，相関を単調に改善しない．

図 4–図 6 に，次元別の相関と 95%信頼区間を示す．図中の有意差は macro MAE に対する条件差検定の結果であり，多重比較は Holm 法で補正する（手順は実験節に従う）．ここで，感情理解と同様に，図中のエラーバーは 95%信頼区間（95% CI）を表す．また，条件間に有意差がある場合はブラケットで示し，アスタリスクで Holm 補正後の p の範囲（*: $0.01 \leqq p < 0.05$，**: $0.001 \leqq p < 0.01$，***: $p < 0.001$）を示す．

6.3 推論深度（推論トークン）

本項は，推論深度の操作がトークン消費へ与える影響を整理する．推論トークンは API 応答 `usage` の `completion_tokens_details.reasoning_tokens` を用いて集計する．図 7 は理解，図 8 は生成の推論トークン分布を示す．GPT-5-mini は `reasoning_effort` の上昇に伴い推論トークンが増加する．一方，表 6 と表 8 は，推論深度の上昇により相関の改善が小さいことを示す．ここで，GPT-5-mini の think minimal は用いられた推論トークン数が 0 であるため，推論なし条件とみなす（以降，推論なし）．以上より，推論深度の増加は推論トークン数を増加させる一方，相関の改善は限定的である傾向が見られる．

7 考　　察

本節は，実験結果から設計上の示唆を整理する．

7.1 感情理解の結果

表 6 より，Valence は全設定で $r \geqq 0.760$ を示す一方，Arousal と Dominance は全設定で $r < 0.5$ を示す．このことは，Arousal と Dominance の推定精度が十分でないため，推定値のみからユーザ状態（例：介入の要否や対話方針を決める内部状態）を確定すると不安定になり得ることを示唆する．したがって，対話システムでは Arousal，Dominance 両次元の推定値を状態決定に直接用いることは避け，ユーザによる自己申告や追加質問で補完する運用が望ましい．

7.2 感情生成の結果

表 8 より，word は Arousal と Dominance の相関が低い一方，num と lex は Arousal と Dominance の相関が高い．具体的には，Valence は num と lex の差の絶対値が 0.000–0.035 である一方，word 条件は Arousal が $r = 0.437$–0.589，Dominance が $r = 0.356$–0.487 と相関が低い．指示形式差は誤差指標 e（macro MAE）でも検証した．対応のある Wilcoxon 符号付順位検定の結果を表 9 に示す．表 9 は Holm 法で補正した p_{Holm} を示す．表 9 より，GPT-5-mini は全推論設定で num の e の中央値が lex より低い．また，GPT-4.1 は no-cot で num の e の中央値が lex より低い一方，cot では num と lex の差は示されない．

以上より，VAD 制御では指示形式として num/lex を選択することが望ましい．表 9 より，macro MAE に対する対応のある Wilcoxon 符号付順位検定（Holm 補正）の結果，モデル・推論設定に依存するものの num が lex より誤差が小さい条件が確認された．よって，生成条件入力および内部状態（VAD 条件）の保持には num を推奨する．一方，提示形式の理解容易性は主観・行動指標が必要であり，本実験ではこれらを収集していないため，検定による比較は行っていない．したがって提示形式は運用要件に基づいて lex/num 等から選択することが望ましい．例えばユーザには lex で感情を表示し，内部状態と生成条件は num で行う構成が取り得る．

7.3 感情理解と感情生成の比較

本項は，感情理解と感情生成の性能差を整理する． 表 6 と表 8 より，理解（ユーザ状態推定）と生成（トーン反映）は同程度の性能ではない．とくに理解では Arousal/Dominance が $r < 0.5$ に留まる一方，生成では num/lex 条件でより高い相関が得られる設定がある．両者を一体で設計すると，特に Arousal/Dominance の推定誤差が応答トーンへ直接伝播し，誤った強度・支配性の応答を生みうるうえ，失敗要因（推定か制御か）の切り分けも困難になる．したがって実運用では，(i) 推定は状態を単一値で確定せず候補（不確実性）として扱い，必要に応じて自己申告や追加質問で補完・補正した上で状態を決定する．(ii) 決定した状態を入力として応答トーンを制御す

表 8　感情生成の Pearson 相関（生成 1,728 件，失敗除外，Text→VAD 評価器）

設定	V(num)	V(lex)	V(word)	A(num)	A(lex)	A(word)	D(num)	D(lex)	D(word)
GPT-5-mini (think high)	0.856	0.859	0.708	0.877	0.869	0.538	0.728	0.743	0.356
GPT-5-mini (think medium)	0.870	0.871	0.734	0.883	0.869	0.547	**0.752**	0.759	0.375
GPT-5-mini (think low)	**0.871**	**0.872**	0.755	**0.890**	0.873	**0.589**	0.752	**0.781**	0.366
GPT-5-mini (think minimal)	0.867	0.863	**0.785**	0.831	0.862	0.568	0.648	0.753	**0.487**
GPT-4.1 (推論誘導なし)	0.866	0.832	0.781	0.847	**0.898**	0.500	0.703	0.732	0.451
GPT-4.1 (推論誘導あり（CoT）)	0.850	0.865	0.777	0.810	0.887	0.437	0.701	0.636	0.470

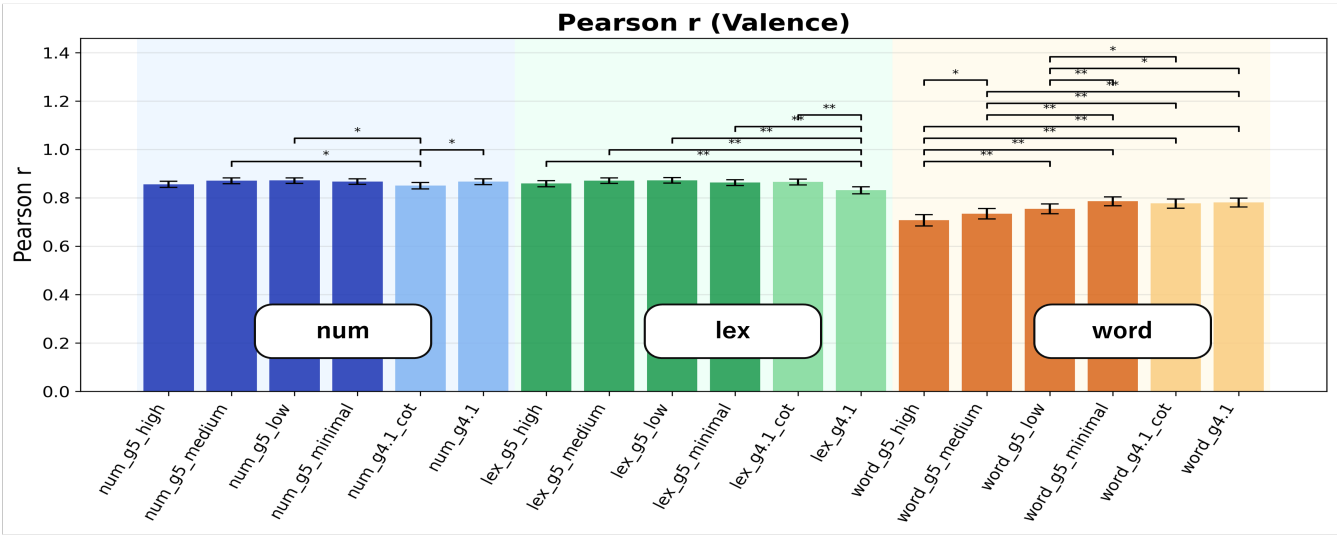

図 4　感情生成における Valence 相関（条件別，95%CI）

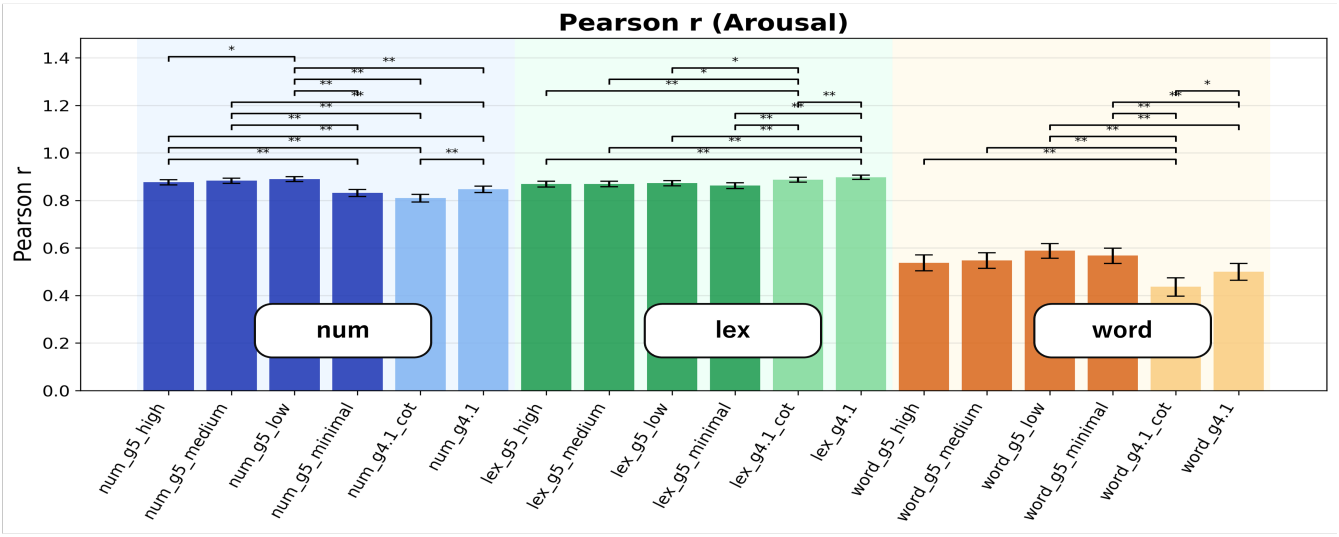

図 5　感情生成における Arousal 相関（条件別，95%CI）

る．このように推定と制御を分離して設計することが望ましい．

7.4　推論深度

表 6 と表 8 より，推論深度の増加は相関の改善を示さない．一方，図 7 と図 8 より，`reasoning_effort` の増加は推論トークンを増加させる．加えて，表 6 と表 8 より，minimal と low は high と同程度の相関を示す．以上より，感情タスクでは推論深度を上げ，計算量（トークン消費）を増やしても性能向上が限定的である．すなわち，推論トークンを増やす運用は費用対効果に乏しい．実際に，minimal または low でも high と同等の相関が得られるため，推論深度を high まで引き上げる必要は少ない．

8　結　　論

本稿は，LLM の感情能力を VAD 空間上の 2 つのタスクで定義し，定量評価した．対象タスクは，感情理解（テキスト→VAD）と感情生成（VAD→ テキスト →VAD）である．評価は EmoBank を用い，GPT-5-mini と GPT-4.1 を対象とした．推論深度は GPT-5-mini の `reasoning_effort` と GPT-4.1 の CoT 有無で操作した．感情生成の指示形式は num/lex/word を比較した．

実験結果から，次の 4 点が明らかとなった．

- 感情理解は Valence の相関が高い．感情理解は Arousal

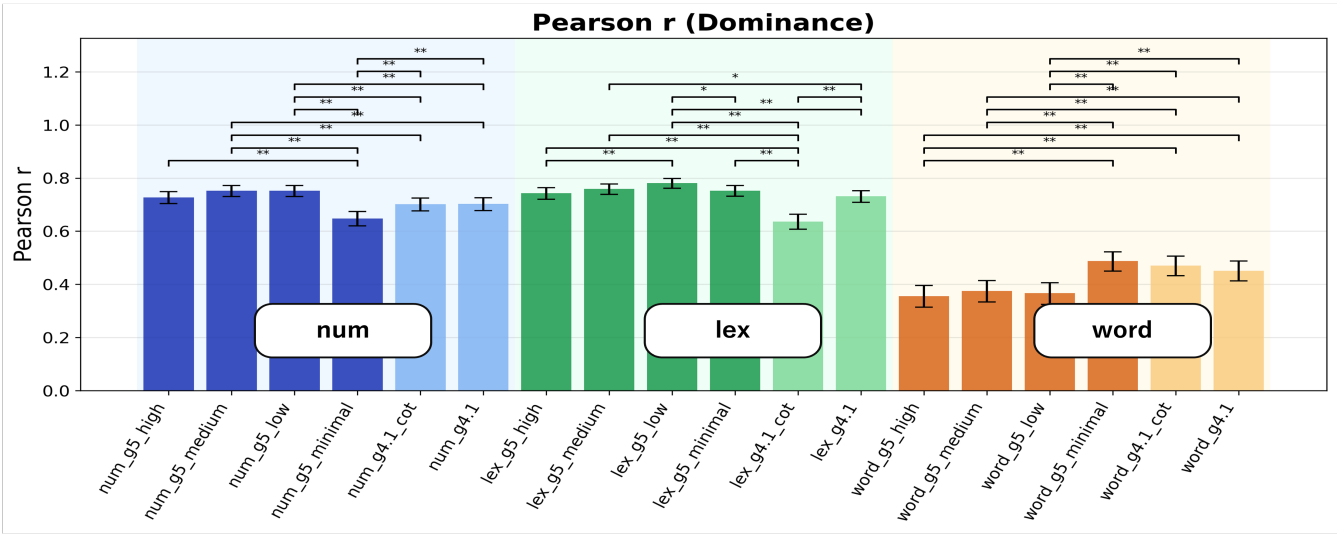

図 6　感情生成における Dominance 相関（条件別，95%CI）

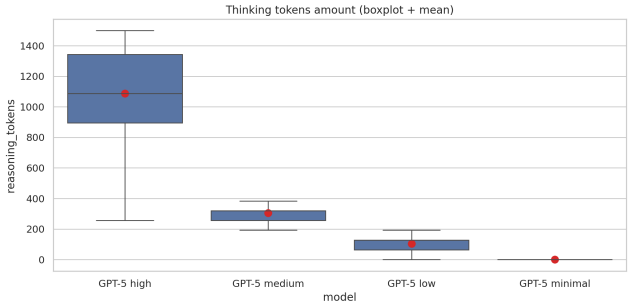

図 7　推論深度設定別 推論トークン数（理解時）

表 9　指示形式差の Wilcoxon 符号付順位検定（誤差 e，macro MAE）

モデル	推論設定	比較	n	Δ	p_{Holm}	r
gpt-4.1	no-cot	num vs lex	1727	+0.024	3.05×10^{-8}	0.133
gpt-4.1	no-cot	num vs word	1728	+0.322	3.74×10^{-188}	0.705
gpt-4.1	no-cot	lex vs word	1727	+0.273	9.36×10^{-152}	0.632
gpt-4.1	cot	num vs lex	1726	-0.005	0.642	0.011
gpt-4.1	cot	num vs word	1728	+0.285	7.46×10^{-138}	0.602
gpt-4.1	cot	lex vs word	1726	+0.256	5.79×10^{-146}	0.620
gpt-5-mini	high	num vs lex	1640	+0.035	1.32×10^{-11}	0.167
gpt-5-mini	high	num vs word	1728	+0.335	2.78×10^{-199}	0.725
gpt-5-mini	high	lex vs word	1640	+0.284	6.42×10^{-159}	0.664
gpt-5-mini	medium	num vs lex	1660	+0.053	1.7×10^{-20}	0.228
gpt-5-mini	medium	num vs word	1662	+0.341	1.49×10^{-195}	0.733
gpt-5-mini	medium	lex vs word	1726	+0.282	2.52×10^{-168}	0.666
gpt-5-mini	low	num vs lex	1708	+0.062	9.6×10^{-32}	0.284
gpt-5-mini	low	num vs word	1707	+0.331	2.94×10^{-202}	0.735
gpt-5-mini	low	lex vs word	1727	+0.246	3.83×10^{-145}	0.618
gpt-5-mini	minimal	num vs lex	1717	+0.037	3.75×10^{-8}	0.133
gpt-5-mini	minimal	num vs word	1692	+0.247	2.58×10^{-151}	0.638
gpt-5-mini	minimal	lex vs word	1701	+0.217	6.94×10^{-123}	0.572

と Dominance の相関が 0.5 未満である．

- 感情生成では，固定した Text→VAD 評価器に基づく評価において，num 条件は Valence と Arousal で $r > 0.84$ を示す．
- 指示形式は追従性に影響する．word 条件は Arousal と Dominance の相関が低下する．誤差 e（macro MAE）では num が lex より低い設定がある．
- 推論深度の増加は推論トークンを増加させる一方，相関の改善に寄与しない．

実験結果は，感情理解と感情生成の間に非対称性があることを示す．対話システムでは Arousal と Dominance の推定値を状態決定へ直接用いることは避けるのが望ましい．応答トーンを制御する場合は，VAD 条件を用いた設計が有効である．推論設定は minimal または low とする選択肢を持つことが望ましい．

今後の課題は次の通りである．

- ユーザ自己申告を組み込む対話設計を実装し，Arousal と Dominance の不確実性を補う効果を評価する．
- テキスト以外の情報を用いる推定を導入し，Arousal と Dominance の推定精度を改善する．
- VAD 条件の提示と応答品質の関係を人手評価で検証し，制御性と自然さのトレードオフを整理する．

文　献

[1] Keivalya Pandya et al. Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for Organizations, 2023. arXiv:2310.05421.

[2] Marcos Belver et al. Requirements Elicitation Based on Psycho-Pedagogical Theatre for Context-Sensitive Affective Educational Recommender Systems. *IEEE Access*, 11:76284–76299, 2023.

[3] Yupei Li et al. Artificial Emotion: A Survey of Theories and Debates on Realising Emotion in Artificial Intelligence, 2025. arXiv:2508.10286.

[4] Saif Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of ACL 2018*, pages 174–184, 2018.

[5] Sven Buechel and Udo Hahn. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of EACL 2017 (Short Papers)*, pages 578–585, 2017.

[6] Jinghan Liu, Tianyu Sun, Lei Hou, and Juanzi Li. EmoLLMs: Emotion Understanding and Generation in Large Language Models, 2024. arXiv:2401.08508.

[7] Rajdeep Mukherjee, Niloy Bhattacharya, and Animesh Mukherjee Ghosh. Understanding the Role of Affect Dimensions in Detecting Emotions from Tweets: A Multitask Approach. In *Proceedings of the 44th Int. ACM SIGIR Conf. (SIGIR '21)*, pages 2303–2307, 2021.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten

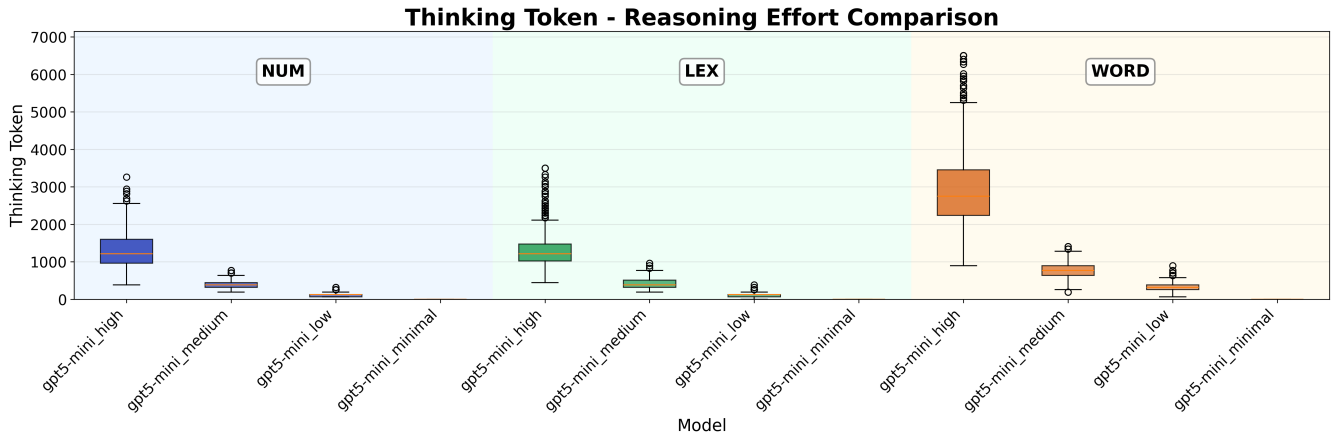

図 8　推論深度設定別 推論トークン数（生成時）

Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645:633–638, 2025.

[10] Shin-nosuke Ishikawa and Atsushi Yoshino. AI with Emotions: Exploring Emotional Expressions in Large Language Models. In *Proceedings of NLP4DH 2025*, pages 614–627, 2025.

[11] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of ACL 2020*, pages 4040–4054, 2020.

[12] Shadab Choudhury, Md. Masum Rahman, Dipto Sarkar, et al. GPT's Devastated and LLaMA's Content: Emotion Representation Alignment in LLMs for Keyword-Based Generation, 2025. arXiv:2503.11881.

[13] Zaijing Li et al. Enhancing Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought, 2024. arXiv:2401.06836.

[14] Yinhan Liu et al. RoBERTa: Robustly Optimized BERT Pretraining Approach, 2019. arXiv:1907.11692.

[15] Klaus R. Scherer and Harald G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328, 1994.

[16] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. Dimensional Emotion Detection from Categorical Emotion. In *Proceedings of EMNLP 2021*, pages 4367–4380, 2021.