多次元幾何学的推論における大規模言語モデルの性能評価

阿部桃大 [1] Namgi Han[1] 宮尾祐介 [1,2]
[1] 東京大学
[2] 国立情報学研究所大規模言語モデル研究開発センター
{tota_abe,hng88,yusuke}@is.s.u-tokyo.ac.jp

概要

人間の幾何学的推論能力は物理世界の制約を強く受けており，四次元以上の高次元空間において顕著な限界を示す．一方，推論において物理世界への身体化を伴わない大規模言語モデル（Large Language Model; LLM）においては，四次元の問題が二次元や三次元の問題と本質的に同様の形で処理可能である可能性がある．本研究では，二次元・三次元・四次元にわたる多次元幾何学ベンチマーク MDGeom を提案し，次元数の増加に伴う LLM の性能変化を観察する．既存の幾何学ベンチマーク GeomRel を次元的に拡張することで MDGeom を構築し，GPT-4o ファミリーや代表的なオープンモデルを用いて性能評価を行う．実験の結果，多くの LLM において人間の場合と同様，四次元以上の高次元空間の認識において困難が生じることが示唆された．

1 はじめに

多くの推論タスクにおいて，人間の推論能力は問題の構造的複雑性が増すにつれて顕著な限界を示すことが知られている．その代表例が幾何学的推論である．人間は二次元や三次元の空間構造を比較的容易に把握できる一方で，四次元以上の空間構造を理解することは，特別な訓練を受けない限り極めて困難である [1]．この困難の一因は，我々が生活する物理世界が三次元空間で構成されていることであり，四次元以上の高次元空間を扱う際には物理世界を参照した推論が不可能となるためである．

これに対し，大規模言語モデル（Large Language Model; LLM）の推論は物理世界への身体化を伴わない．LLM は大量のテキストデータを学習することで統計的・構造的規則性を獲得し，それに基づいて現実世界を抽象的に表現した世界モデルを内部に形成する [2]．この世界モデルは，人間が知覚や行動体験を通じて獲得する三次元的な世界理解とは異なる性質を持つと予想される．したがって，幾何学的対象が三次元であろうと四次元であろうと，LLM の世界モデルにおいては本質的な差異が存在しない可能性がある．また，近年では推論に特化した LLM を開発する取り組みが盛んに行われている [3, 4]．こうした推論に強いモデルにおいては，先に述べた LLM 固有の世界モデルの特性と合わせて人間には困難とされる四次元以上の高次元の幾何学空間の理解を実現できることが期待される．

本研究では，二次元・三次元・四次元にわたる幾何学ベンチマーク MDGeom (**M**ulti**D**imensional **Geom**etric Benchmark) を提案し，次元の増加に伴う LLM の性能変化を観察することで，人間に見られる三次元から四次元にかけての大幅な性能低下が LLM においても認められるかを検証した．実験の結果，一般的に幾何学的推論性能は次元数の増加とともに低下する傾向を示すことが確認された．また，多くのモデルで三次元から四次元にかけての性能低下が二次元から三次元にかけての性能低下と比較して大きい傾向が観察された．

本研究の実験コードおよびデータセットは，GitHub[1)] で公開している．

2 関連研究

2.1 幾何学的推論ベンチマーク

数学的推論は人間の知性と深く結びついており，LLM の性能を評価する上でも重要なタスクとして位置付けられる [5]．言語モデルにおける数学的推論能力を定量的に把握することを目的として，MATH データセット [6] をはじめとしたベンチマークが数多く開発されてきた [7, 8, 9]．

本研究の対象である幾何学的推論も例外ではなく，Geometry3K[10] や GeomRel[11] など，幾何学的

1) https://github.com/anasker88/Multidimensional_Geometric_Evaluation

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

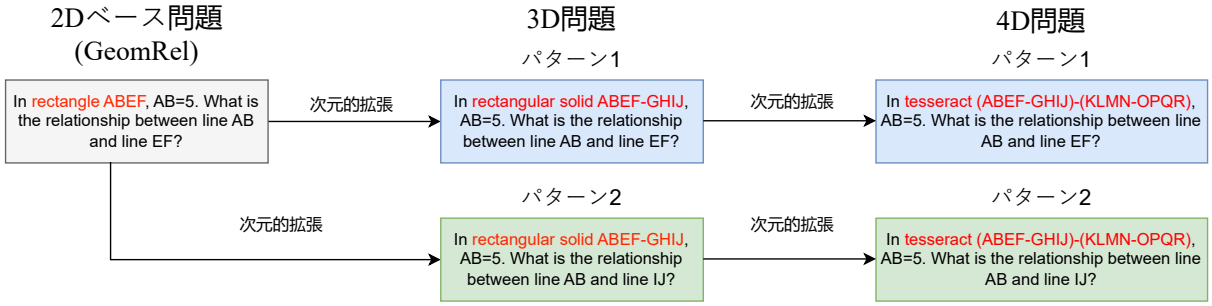


図 1　次元的拡張による選択肢問題生成の手順．二次元の問題を答えを保ったまま自然に次元的拡張することで，三次元・四次元の問題を生成する．次元拡張時の変更箇所を赤字で示す．次元の増加に伴い自由度も増加するため，1 つのベース問題から複数の三次元・四次元問題が生成されることもある．

推論に特化した評価ベンチマークが提案されている．一例として，GeomRel はすべて二次元幾何に関する選択肢形式の問題から構成され，直線，角度，図形の三つのカテゴリに分類される．また，直線に関する問題は，点と直線の関係，および直線同士の関係（平行・垂直・交差）に細分化されている．

こうした数学的推論や幾何学的推論に関するベンチマークの多くは，教科書やインターネット上の問題を収集することで構築されており，その内容は主として高校生レベルの問題から構成されている．その結果，収録されている問題の大半は二次元幾何に限定されており，三次元幾何を体系的に扱うデータセットは極めて少ない．さらに，四次元以上の幾何学的推論を対象とした評価ベンチマークは，我々の知る限り存在しない．

このような背景のもと，本研究は，大規模言語モデルに対して四次元を含む多次元にわたる幾何学的推論を体系的に評価する，初めての試みである．

2.2　LLM における推論強化

近年，LLM は性能の向上に伴い，言語処理にとどまらず多様な応用領域で活用されるようになっている．こうした背景の下，LLM の推論能力を強化する手法の開発が活発に進められている．代表的な手法として，思考の連鎖（Chain of Thought; CoT）[12] が挙げられる．CoT では，Few-shot プロンプトを用いて推論過程を明示的に生成させることで，LLM の推論性能を大幅に向上させる．

さらに近年では，モデル自体を推論に特化させる取り組みも増加している．たとえば，o1[3] は推論能力を強化するために強化学習を導入し，CoT を活用して深い推論を行った後に出力するよう設計されている．

表 1　次元ごとの基本的図形．これらの図形に対し，頂点数，辺長，面積/体積/超体積，周長/表面積/表体積，対角線長などを問う問題を作成する．

二次元	三次元	四次元
長方形	直方体	四次元超直方体
正方形	立方体	四次元超立方体
三角形	三角錐	四次元単体
円	球	四次元超球

また，Qwen-Math シリーズは数学的推論に特化したモデルであり，Qwen2.5-Math[4] では Qwen2-Math を用いて高品質な数学データセットを構築し，学習に利用している．さらに，教師あり学習による報酬モデルの改善を行い，これを用いて出力のサンプリングを制御している．加えて，推論時には CoT に加え，Python コードの生成・実行を組み込むツール統合型推論を採用することで，より効果的な推論を実現している．

こうした推論特化モデルは数値計算や形式的推論による一般的な数学的推論ベンチマークによって評価されている．しかし，幾何学的推論を対象とした評価は手続き型コードなど特定の形式に限られ [13]，十分な検証がなされたとは言えない．本研究では推論特化モデルを対象に多次元にわたる幾何学的推論能力を検証する．

3　MDGeom の構築

3.1　構築手順

MDGeom は，GeomRel[11] をベースとして作成されている．まず，GeomRel の問題を次元的に拡張することで選択肢問題を作成する．選択肢問題の作成手順を図 1 にしめす．次元的拡張を容易にするた

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 2　各次元における問題数．次元的拡張によって問題を生成しているため，各次元の問題数は完全には一致しない．

問題区分	二次元	三次元	四次元
PPC	180	258	270
IC	180	258	270
CSC	138	138	138
NUM	118	118	118
合計	616	772	796

め，ベースとする問題は線分に関するものに限定し，拡張が困難と判断される問題は除外する．こうして選定されたベース問題を答えを保ったまま自然に次元拡張することで，二次元・三次元・四次元にわたる多次元幾何学問題を作成する．次元的拡張において，ベース問題で扱う図形によって拡張される問題の数は異なる．選択肢問題については，問題数を確保するため，頂点名を入れ替えることで複製を行っている．

さらに，問題の多様性を担保するため，選択肢問題に加えて数値問題を作成する．数値問題の作成にあたっては，GeomRel を参考に次元ごとに基本的図形を定義する．その一覧を表 1 に示す．基本的図形に対して，頂点数などの基本的性質を問う問題，面積・体積・超体積や周長・表面積・表体積を計算する問題，対角線長を求める問題を作成し，辺長などを変更することで複製を行う．

各次元における問題数は，表 2 に示す通りである．

3.2　問題区分

MDGeom は，平行・垂直分類（Parallel Perpendicular Classification; PPC），交差分類（Intersection Classification; IC），共通部分空間分類（Co-subspace Classification; CSC），および数値問題（Numeric Problem; NUM）から構成される．このうち，PPC，IC，CSC は選択肢問題であり，NUM は整数値による回答を求める数値問題である．PPC と IC は GeomRel の直線同士の関係に関する問題に，CSC は点と直線の関係に関する問題に由来し，NUM は問題の多様化を目的として本研究で独自に設計された問題集である．PPC，IC，CSC の選択肢は付録 4 の表 4 に示す．

PPC は，与えられた 2 つの線分が平行か垂直かを判定する問題である．IC は，与えられた 2 つの線分が交差するか否かを判定する問題である．ここで，PPC と IC で LLM に与える図形の情報は同一であり，選択肢のみが異なっている．CSC は，与えられた点群が共通部分空間上に存在するか否かを判定する問題である．具体的には，二次元では 3 点が同一直線上に存在するか，三次元では 4 点が同一平面上に存在するか，四次元では 5 点が同一三次元超平面上に存在するかを問う．NUM は，基本的図形に関する頂点数，辺長，面積・体積・超体積，周長・表面積・表体積，対角線長などを問う問題であり，整数値による回答を求める．

4　実験

4.1　モデル

分析対象とするモデルとして，一般に広く利用され，各種ベンチマークにおいて高い性能を示している GPT-4o ファミリーを選択した．さらに，モデルの軽量化の影響および推論特化の強化学習の効果を検証するため，GPT-4o，GPT-4o-mini，o1 の性能を比較する．

また，代表的なオープンモデルとして，Qwen2.5-7B-it（以下，Qwen2.5）および Gemma2-9B-it（以下，Gemma2）を検証の対象とする．さらに，数学特化の強化学習の効果を評価するため，Qwen2.5-Math-7B-it（以下，Qwen2.5-Math）における性能も検証し，Qwen2.5 との比較を行う．

4.2　プロンプト

他の推論タスクでは，CoT などのプロンプト設計により性能が向上することが確認されている [12]．そこで本研究では，プロンプト内に推論することを明示することによる効果を検証するため，推論あり，推論なしの 2 種類のプロンプトで実験を行う．プロンプトの詳細は付録 B の図 2 に示す．

4.3　評価指標

各設定における性能は精度 (Accuracy) によって評価する．

5　結果と考察

推論ありのプロンプトを用いた実験結果を表 3 に示す．本研究は LLM に対する推論強化が多次元幾何学的推論能力に与える影響の検証を目的としているため，本文では推論ありとした場合のみを対象に議論を行う．推論なしのプロンプトを用いた実験結果は付録 C の表 5 を参照されたい．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3　推論ありとした場合の問題区分ごとの精度 (%).

区分	次元	GPT-4o	GPT-4o-mini	o1	Qwen2.5	Qwen2.5-Math	Gemma2
PPC	2D	80.56	85.56	86.11	85.00	81.11	65.56
	3D	62.02	68.60	73.64	72.09	64.34	58.14
	4D	49.26	56.30	67.41	64.07	57.41	57.04
IC	2D	84.44	72.78	99.44	82.78	87.22	73.89
	3D	77.52	65.50	91.47	69.38	73.26	74.42
	4D	69.63	57.41	85.93	67.04	74.44	66.30
CSC	2D	85.51	84.78	94.93	71.01	73.91	66.67
	3D	86.23	80.43	94.20	90.58	76.81	63.04
	4D	76.09	65.94	85.51	67.39	68.84	57.25
NUM	2D	98.31	98.31	100.00	94.92	98.31	94.92
	3D	100.00	100.00	100.00	100.00	100.00	90.68
	4D	85.59	54.24	95.76	61.02	67.80	40.68
Overall	2D	86.20	84.09	94.64	83.12	84.58	73.86
	3D	77.33	74.48	87.31	78.76	75.00	69.43
	4D	66.21	58.04	81.03	65.20	66.71	57.79

全体的な傾向として，次元の増加とともに精度の低下が確認された．一方で，Qwen2.5 における CSC など，一部のケースでは二次元から三次元にかけて精度の向上が見られた．次元が増加すると問題を回答するために考えるべき図形や法則の数も増加するため精度は低下するというのが自然な仮説であるが，LLM においてはその仮説が通用しない可能性があることが示唆された．

二次元から三次元にかけての性能低下と三次元から四次元にかけての性能低下を比較すると，ほとんどのケースで三次元から四次元にかけての性能低下が大きい傾向を示した．特に，NUM では GPT-4o と o1 を除く全てのモデルが三次元から四次元にかけて 30%以上の性能低下を示しており，四次元において顕著な限界に達していることを示唆している．

モデルの軽量化の影響として，GPT-4o-mini は GPT-4o と比較して二次元・三次元においては全体として 3%前後の性能差にとどまった．一方で，四次元においては 8%ほど性能が低下する結果となった．このことから，GPT-4o-mini への軽量化において，三次元以下の幾何学的推論能力は維持された一方，四次元以上の高次元の幾何学的推論能力はある程度損なわれているものと考えられる．

また，推論強化の影響として，o1 は各次元で最も高い性能を示した．特に，次元ごとの精度を GPT-4o と比較すると，次元数の増加とともに精度の改善幅は大きくなっている．このことから，o1 における推論強化により，特に高次元における幾何学的推論能力が大幅に強化されたと言える．

一方で，数学特化の推論強化の影響として，Qwen2.5-Math の精度は Qwen2.5 と比較してほとんど改善を示していない．このことから，Qwen2.5-Math においては o1 とは異なり高次元における幾何学的推論能力は改善していないものと考えられる．この原因に関しては今後の分析が求められる．

6　おわりに

本研究では，大規模言語モデル（LLM）における多次元幾何学的推論能力を体系的に評価するため，二次元・三次元・四次元にわたる幾何学ベンチマーク MDGeom を新たに構築し，複数のモデルに対して性能を比較した．その結果，次元数の増加に伴い性能の低下が確認された．さらに，多くのモデルにおいて人間と同様，三次元から四次元にかけての性能低下が二次元から三次元にかけての性能低下と比較して大きい傾向が観察された．また，推論強化によって一部のモデルでは効果的に高次元における幾何学的推論能力を向上することができることが確認された．

今後の課題として，(1) 次元数のさらなる拡張，(2) 問題区分ごとに異なる傾向を示す原因の究明，(3) 次元数増加に伴う性能低下の機械論的解釈可能性による分析などが挙げられる．こうした取り組みを通じて，幾何学的な観点から見た LLM の世界モデルの構造的限界とそのメカニズムをより深く理解するとともに，LLM を用いた効果的な高次元空間理解の実現への指針を得られることが期待される．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです.

参考文献

[1] Michael S Ambinder, Ranxiao Frances Wang, James A Crowell, George K Francis, and Peter Brinkmann. Human four-dimensional spatial intuition in virtual reality. **Psychonomic bulletin & review**, Vol. 16, No. 5, pp. 818–823, 2009.

[2] Cole Robertson and Philip Wolff. Llm world models are mental: Output layer evidence of brittle world model use in llm mechanical reasoning. **arXiv preprint arXiv:2507.15521**, 2025.

[3] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. **arXiv preprint arXiv:2412.16720**, 2024.

[4] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. **arXiv preprint arXiv:2409.12122**, 2024.

[5] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. **arXiv preprint arXiv:2402.00157**, 2024.

[6] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. **arXiv preprint arXiv:2103.03874**, 2021.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[8] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. **arXiv preprint arXiv:2210.03057**, 2022.

[9] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In **Proceedings of the 58th annual meeting of the Association for Computational Linguistics**, pp. 975–984, 2020.

[10] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Intergps: Interpretable geometry problem solving with formal language and symbolic reasoning. **arXiv preprint arXiv:2105.04165**, 2021.

[11] Xiaofeng Wang, Yiming Wang, Wenhong Zhu, and Rui Wang. Do large language models truly understand geometric structures? **arXiv preprint arXiv:2501.13773**, 2025.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, Vol. 35, pp. 24824–24837, 2022.

[13] Shixian Luo, Zezhou Zhu, Yu Yuan, Yuncheng Yang, Lianlei Shan, and Yong Wu. Geogrambench: Benchmarking the geometric program reasoning in modern llms. **arXiv preprint arXiv:2505.17653**, 2025.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

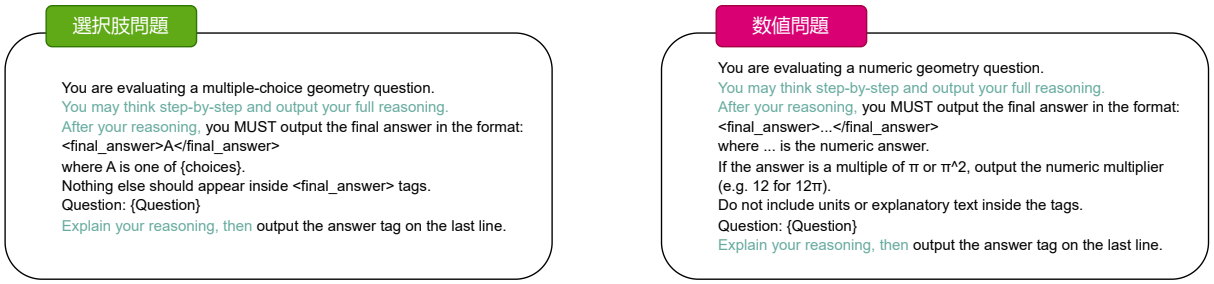
A　選択肢一覧

表 4　選択肢問題の選択肢

問題区分	選択肢 1	選択肢 2	選択肢 3	選択肢 4
PPC	A. Parallel	B. Perpendicular	C. Neither Parallel nor Perpendicular	D. Cannot be inferred
IC	A. Intersecting	B. Not Intersecting	C. Cannot be inferred	-
CSC	A. Yes	B. No	C. Cannot be inferred	-

B　プロンプト

図 2　実験に使用するプロンプト．左が選択肢問題に使用されるプロンプト，右が数値問題に使用されるプロンプトである．{choices}には選択肢一覧が，{Question}には問題文がそれぞれ挿入される．示されているのは推論ありのプロンプト全文であり，推論なしのプロンプトは緑で示される箇所を削除したものである．



C　推論なしとした場合の実験結果

表 5　推論なしとした場合の問題区分ごとの精度 (%).

区分	次元	GPT-4o	GPT-4o-mini	o1	Qwen2.5	Qwen2.5-Math	Gemma2
PPC	2D	86.11	71.67	85.00	63.89	80.00	53.89
	3D	65.12	65.89	82.17	62.79	63.57	36.05
	4D	60.37	62.59	66.67	49.63	57.41	34.07
IC	2D	74.44	62.78	99.44	41.11	89.44	65.00
	3D	77.52	63.95	93.41	27.52	74.81	67.44
	4D	68.52	55.19	87.04	26.67	64.07	46.67
CSC	2D	71.01	71.74	97.83	43.48	65.22	40.58
	3D	81.88	69.57	94.93	43.48	76.09	40.58
	4D	67.39	63.04	88.41	44.20	68.84	47.83
NUM	2D	97.46	99.15	100.00	88.14	97.46	81.36
	3D	100.00	90.68	100.00	83.05	100.00	59.32
	4D	81.36	53.39	92.37	55.08	75.42	31.36
Overall	2D	81.49	74.35	94.97	57.31	82.79	59.42
	3D	77.59	69.69	90.93	50.65	75.13	50.91
	4D	67.46	58.79	81.16	41.71	64.32	40.33

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).