大規模言語モデルの構造的問題としての ハルシネーションの不可避性についての一考察
――文法獲得と副次的な知識獲得の関係からの検討――

新　美　潤一郎

概　要

大規模言語モデル（LLM）の普及に伴い，さまざまな領域で産業の変革が続いている。一方で，その出力として事実に基づかない回答（hallucination）が生成されることが問題視されている。そこで本研究では，LLMの構造からその学習機序を再考し，hallucinationが不可避的に発生することを示す。実験1としてプロンプトにおけるロールの変化が結果に及ぼす影響を示し，実験2として論文の書誌情報の生成を通じてhallucinationの発生要因を2つのパターンに類型化する。これらの定量的／定性的な解析の結果およびTransformerの構造から，LLMは普遍的な知識を保持しているわけではなく，そこで学習しているのはあくまでもトークン間の依存関係に基づく文法的パターンであり，知識はその副産物として埋め込まれているにすぎないことを指摘する。最後に，文法と知識の不可分性という構造的制約から，hallucinationの必然性を理論的に論じる。

1　はじめに

自然言語処理（natural language processing, NLP）技術の急速な発展により，近年では大規模言語モデル（large language model, LLM）の高性能化が進んでいる。特にLLMをチャットAIアシスタントとして実装したOpenAI社のChatGPT（chatgpt.com）やGoogle社のGemini（gemini.google.com），Anthropic社のClaude（claude.ai）等は社会に急速に普及し，既に産業構造の変革を少なからず引き起こしている。これらモデルの学習機序やその性能を引き起こす構造については既に技術的な解説が豊富に存在するが，一方その学習メカニズムに対する理論的考察は不十分である。特にLLMにおける喫緊の課題として，事実に基づかない返答としてのハルシネーション（hallucinations）が問題となっていることからも，LLMが「知識を保有している」という一般的な理解に対して，学習アルゴリズムの観点から批判的に議論する必要がある。

そこで本研究では，LLMの学習機序をパターン認識の観点から再考し，知識システムとしてのLLMの活用の問題点を指摘する。念のため先に述べておくと，本研究は高度な推論タスクにおけるLLMの有用性を否定するものではなく，LLMの事前学習からの知識抽出という特定の用途における構造的限界をさまざまな視点から指摘するものである。特に，LLMにおけるi.）文法獲得と知識獲得の不可分性，ii.）量的検証に基づく回答の文脈依存性，iii.）質的検証に基づくhallucinationの不可避性と事前学習との関係性の3点を中心に論考を展開する。

本研究の構成は，まず第2節で既存研究の整理を行い，LLMの技術的依拠を俯瞰する。第3節では，LLMの学習過程において文法獲得と知識獲得が不可分であることをLLMの構造から述べ

る。次に第4節では，LLMの保持する知識が文脈条件により変化し必ずしも普遍的な知識システム足り得ないことを定量的な分析により検証する。第5節では実際にLLMでテキストを生成することにより，事前学習に含まれる関連知識の量によりhallucinationの挙動が変化する可能性を提示し，定性的な分析によりそれを検証する。第6節ではここまでに述べてきたhallucinationの問題と，LLMにおけるプライバシー保護の課題が根本的に矛盾する可能性について指摘する。最後に，第7節で本研究から得られる示唆や限界をまとめる。

2　発展経路の俯瞰的整理

学習の構造の検討に入る前に，LLMが技術的に依拠するモデルの発展経路や，LLMにおけるtokenの生成機序，さらにはLLMの学習までを順に整理する。

2.1　LLMの技術的依拠としてのTransformer

大規模言語モデルの技術的基盤はmultilayer perceptron（MLP; Rumelhart et al., 1986）をはじめとするdeep neural network（DNN）であるとともに，その技術的発展はAttention（Bahdanau et al., 2015）やTransformer（Vaswani et al., 2017）等の一連のモデル開発に立脚している。Transformerに至るまでの深層学習の発展については様々存在する古典的文献（e.g., LeCun et al., 2015; Goodfellow et al., 2016）あるいは新美（2025）等を参照されたい。

Transformer以前には，時系列データやテキストデータをはじめとしたsequential dataにおけるtempo-ral/contextualな関係性の把握のために，recurrent neural network（RNN; Jordan, 1997; Elman, 1990）やその発展手法としてのlong short-term memory（LSTM; Hochreiter and Schmidhuber, 1997）等を用いることが一般的であった。しかしながら，そもそも再起的な構造による関係性の理解は勾配消失／勾配爆発に繋がりうることから，要素間の距離が遠くなるほど困難であるという限界が指摘されてきた（Bengio et al., 1994; Bahdanau et al., 2015）。

そのような状況において技術的ブレイクスルーとなったのがTransformerである。Transformerはその特性として“relying entirely on an attention mechanism to draw global dependencies between input and output”（Vaswani et al., 2017, Section 1），つまりそれまで再起的な構造を用いることが一般的であったsequentialな関係性の捕捉を，基本的にはAttentionのみによって実現している点に大きな特徴がある。

2.1.1　Transformer Encoder

ここでTransformerの構造を具体的に示す。TransformerはEncoder-Decoderモデルであることから，まずはEncoderを導入する。Encoderは入力側のモジュールであることから，入力された系列を符号化し文脈表現を得ることを主目的とする。

まず入力された系列xを元にしたquery（Q），key（K），value（V），そしてKの次元数d_Kを用いたscaled dot attention（Att）を，

$$Att(Q,\ K,\ V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{1}$$

とすると，Transformer 内のある attention head（$Head_m$, $m \in \{1, 2, ..., M\}$）は重み付けパラメーター W を用いて，

$$Head_m = Att(QW_m^Q,\ KW_m^K,\ VW_m^V) \tag{2}$$

となることから，Multi-head Attention（$MHAtt$）は

$$MHAtt(Q,\ K,\ V) = concat(Head_1,\ Head_2,\ \cdots,\ Head_M)\, W^O \tag{3}$$

である。ここで特徴次元ごとの平均（μ），分散（σ^2），学習可能なスケーリング・シフトパラメータ（γ, β）を導入して，正規化層 LayerNorm（·）を

$$\mathrm{LayerNorm}(u) = \frac{u-\mu}{\sqrt{\sigma^2+\epsilon}} \odot \gamma + \beta \tag{4}$$

同様に全結合層 $FFN(\cdot)$ を

$$FFN(u) = \phi(uW_1 + b_1)\, W_2 + b_2 \tag{5}$$

とすると，残差接続を用いて

$$x' = \mathrm{LayerNorm}(x + MHAtt(x,\ x,\ x)) \tag{6}$$

$$z_{enc} = \mathrm{LayerNorm}(x' + FFN(x')) \tag{7}$$

として Transformer Encoder の $block_{enc}(x) = z_{enc}$ が得られる（c.f. Vaswani et al., 2017）。実際にはこの構造を複数回繰り返すことで Transformer Encoder が構成される。

2.1.2　Transformer Decoder

次に Decoder を導入する。Decoder は Encoder で得られた文脈表現に基づいて系列を逐次的に生成する自己回帰（autoregressive, AR）モジュールである。Encoder 部分との差異として，将来方面に存在する情報の参照を防ぐための masked Multi-head Attention と，入力された系列の情報を取り込むための Encoder-Decoder Attention を用いる点がある。

具体的には，masked Multi-head Attention は元の $MHAtt$ をベースに，self-attention 部分で入力のうち現在時点より将来の値を $-\infty$ とするマスク行列 M を導入した masked self-attention（$MAtt$）を

$$MAtt(Q,\ K,\ V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_K}} + M\right)V \tag{8}$$

として組み込む。また，Encoder-Decoder Attention（$EDAtt$）では，K, V として Encoder の出力 K_{enc}, V_{enc} を用いて，

$$EDAtt(Q,\ K,\ V) = MHAtt(Q,\ K_{enc},\ V_{enc}) \tag{9}$$

とする。最後にこれらのモジュールについてEncoderと同様に残差接続を用いて，

$$x' = \mathrm{LayerNorm}(x + \mathrm{MHAtt}(x,\ x,\ x)) \tag{10}$$

$$h'' = \mathrm{LayerNorm}(h' + \mathrm{EDAtt}(h',\ K_{enc},\ V_{enc})) \tag{11}$$

$$z_{dec} = \mathrm{LayerNorm}(h'' + \mathrm{FFN}(h'')) \tag{12}$$

によりDecoderブロック $block_{dec}(x) = z_{dec}$ が得られ，このブロックを複数層を積み重ねることでDecoderが構成される。

2.1.3　Transformerと普遍性定理

そもそも，十分に多くのニューロンを伴う1層の全結合型の隠れ層をもつDNNは，特定の条件下において任意の連続関数を近似可能であることが普遍性定理（universal approximation theory, UAT; Cybenko, 1989; Hornik, 1991）として知られているが，これらはあくまでも線形結合と非線形活性化の組み合わせによるものであった。近年，Transformerにおいて任意のsequence-to-sequence関数を近似可能であることが示されるとともに（Yun et al., 2019），線形／周期的な自己回帰における表現力の高さに関する分析（Sander and Peyré, 2025）などが行われている。これらの結果は，周期的な系列の予測あるいは自己回帰において，Transformerが理論的に十分な表現力を持つことを示すものであるとともに，次に示す言語モデルへの応用の必然性を支持するものでもある。

2.2　大規模言語モデル

主要なLLMは，ここまでに整理したTransformerの構造を基盤としている。代表的なモデルとして，Bidirectional encoder representations from transformers（BERT; Devlin et al., 2018）やGenerative Pretrained Transformer（GPT; Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023），Text-to-Text Transfer Transformer（T5; Raffel et al., 2020），Mixture-of-Experts（MoE; Jacobs et al., 1991; Shazeer et al., 2017）等が存在し，それぞれが異なる特性を持っている。

まずBERTはTransformerのEncoder部分を用いたEncoder-onlyモデルである。具体的には，先に定義したTransformer Encoderを J 層（$J \in \mathbb{N}_{>0}$）重ねた構造として，入力 x を用いて

$$\mathrm{BERT}(x) = \mathrm{Encoder}^{J}(x), \tag{13}$$

と表現できる。BERTでは文章を前後の双方向（bidirectional）から読み込むことにより，他の単語との相対的な位置（i.e., 文脈）から意味や表現を理解することに長けている。代表的なタスクとして，文の途中にある欠損（mask）トークンを予測するmasked language modeling（MLM）がある。

一方のGPTはTransformerのDecoder部分を用いたDecoder-onlyモデルである。その構造は，Decoderを J 層重ねた自己回帰構造として，非常に簡易化して表すなら

$$\mathrm{GPT}(x) = \mathrm{Decoder}^{J}(x), \tag{14}$$

である。ただし，実際には Transformer ブロックの冒頭に LayerNorm を加える等の変更が行われている。また，近年の GPT 系モデルでは LayerNorm を残差接続の前に配置する Pre-LayerNorm が採用されていることが多い。GPT では文章を前から順に（forward）読むことから，次の単語の予測に強みを持ち，文章生成 / 単語予測などの causal language modeling（CLM）に用いられる。

そして T5 は Transformer の Encoder-Decoder を共に用いたモデルである。いずれの NLP タスクもテキストの入出力の関係性の問題として定式化することにより，単一の事前学習や fine-tuning 手法により統一的に扱うことを可能とした点に大きな特徴がある。Encoder-Decoder モデルという意味で基本的な構造は Transformer に類似しているが，GPT と同様に Pre-LayerNorm を採用している点や，活性化関数が gaussian error linear unit（GELU; Hendrycks and Gimpel, 2016）に差し替えられている点などが異なる。また，Encoder-Decoder モデルには他にも BERT の双方向性と GPT の自己回帰性を組み合わせて denoising autoencoder（Vincent et al., 2008）として設計された Bidirectional and Auto-Regressive Transformers（BART; Lewis et al., 2019）等がある。

最後に MoE であるが，そもそも MoE は BERT や GPT のいずれの構造にも適用可能な計算効率化のためのアーキテクチャであり，近年では特に Decoder-only モデルにおけるスケーリングのために積極的に活用されている。具体的にはモデル内に異なる職能を持つ複数の experts を設定し，それらを選択的に活性化することにより，モデルの計算効率を向上させる。これは Transformer アーキテクチャの拡張，特に FFN 部分を MoE に置換する設計により実現される。E 個の expert $f_i(\cdot)$ とゲート関数 $G(x)$ を用い，top-k のみを活性化する前向き計算は

$$\mathrm{MoE}(x) = \sum_{i \in \mathrm{Top}k(G(x))} \hat{G}_i(x) f_i(x) \tag{15}$$

$$\text{where } \hat{G}_i(x) = \frac{G_i(x)}{\sum_{j \in \mathrm{Top}k} G_j(x)}. \tag{16}$$

となる。

このように一言に大規模言語モデル（LLM）あるいは生成 AI といっても，その内実は言語モデルのみに絞っても異なる構造に基づき多様である。ただ，いずれにせよそのほとんどが Transformer を基盤としているという意味で，Transformer の貢献が大きいことを改めて強調しておきたい。

2.3　学習済み LLM におけるトークン選択

先に述べたように LLM の基盤モデルはさまざま存在するが，本研究ではその中でも LLM の基盤として広く採用されている GPT モデルに着目する。そもそも一般的に LLM は与えられた文章（prompt）に続く token を予測する構造である。ここで token とは，単語をさらに分割したサブワード（e.g., 単語 notebook = note + book）を意味する。複合的な単語を独立した一単語として扱うの

は語彙空間の冗長化を招き計算効率を低下させるため，LLM ではサブワードや文字単位への分割が用いられる。

そして一連の文章生成は，1 token の予測ごとに prompt 末尾に生成済み token を付加する形で prompt を更新しながら生成を繰り返す自己回帰構造として理解できる。ここで LLM の持つ語彙空間 $\mathbb{V}$，そこで定義された全トークン数 K，生成するトークン数 T について，生成のある時点 $t\in\{0,\ 1,\ \cdots,\ T-1\}$ に与えられた $prompt_t$ に続く 1 token を確率変数 X_{t+1}，得られる token とその確率のペアを $token_k\in\mathbb{V}$ および p_k とすると（ここで $k\in\{1,\ 2,\ \cdots,\ K\}$），LLM の token 予測は以下に示す条件付き確率

$$P(X_{t+1}=token_k|prompt_t)=p_k=\frac{\exp(z_k)}{\sum_{j=1}^{K}\exp(z_j)} \tag{17}$$

である。ここで z_k は $prompt_t$ を与えた場合の $token_k$ に対応する logit であり，もちろん

$$\sum_{k=1}^{K}p_k=1 \tag{18}$$

である。

つまるところある学習済みモデル $LLM(\cdot)$ の最も単純な挙動では，

$$LLM(prompt_t)\ \rightarrow\ [p_1,\ p_2,\ \cdots,\ p_K] \tag{19}$$

により，入力した $prompt_t$ に対して，知りうる全 token の出現確率を推定する。この構造は，事前学習が広範であればあるほど，一般的な文章における token 間の関係性に基づきその確率が推定されることを意味しており，LLM が「最も蓋然性の高い単語の並びを生成する」のはこの構造のためである。

ここで，ある学習済みモデルはあくまでも与えられた prompt のみを条件とした確率分布であることから，その分布は prompt の配列が変更されない限り変化しない。単純に最も高い確率の token を選択する決定論的な greedy selection において，予測値 $\hat{X}_{t+1}$ は

$$\hat{X}_{t+1}=\mathrm{token}_{k^*}\quad \text{where } k^*=\arg\max_{k\in\{1,\ldots,\ K\}}\ p_k \tag{20}$$

である。

LLM の使用にあたり同じ prompt でも生成のたびに挙動が僅かながらに変化するのは，その生成結果に多様性を持たせることを目的としたサンプリングが行われるためである。例として，累積確率の合計が S になるような最小の語彙集合からサンプリングする top-p sampling（nucleus sampling; Holtzman et al., 2020）や，確率値の高い順に上位 m 件の語彙集合からサンプリングする top-k サンプリング[*1] 等も用いられる。

サンプリングとあわせて，出力に創造性あるいは確定性を持たせるため，得られた確率分布に対

[*1] 一般的には上位 k 件を対象とすることから top-k と呼ばれるが，本研究では添字の整合性の都合からここでは m を用いた。

して温度（temperature）パラメータによる操作（温度スケーリング，temperature scaling）を行うことも一般的である。

temperature パラメータ $\tau \in (0, +\infty)$ を用いて

$$P(X_{t+1} = token_k | prompt_t; \tau) = \frac{\exp(z_k/\tau)}{\sum_{j=1}^{K} \exp(z_j/\tau)} \tag{21}$$

である。式からもわかるとおり，$\tau < 1$ ではエントロピー減少により分布が鋭くなり，サンプリングを行ってもなお高確率の token の方がより選ばれやすくなることから，出力は決定論的な性質を強くする。一方 $\tau > 1$ ではエントロピー増加により分布が平準化し，低確率のトークンも選択されやすくなるため，出力は多様性を増し創造的になる。これらのことから分かるとおり，式（17）は temperature scaling を行わない場合（$\tau = 1$）の特殊形である。

2.4　損失関数の設計

まず，深層学習一般における学習過程を整理する。トークン選択のような多クラス分類タスクにおいては一般的に以下に示す Cross Entropy（CE; Murphy, 2012）Loss が用いられる。

$$L^{(step)}(y, \hat{y}) = -\sum_{k=1}^{K} y_k \cdot log(\hat{y}_k) \tag{22}$$

ここで K はクラス数である。この式からも分かるように，CE Loss は真の分布 y と推定された確率分布 $\hat{y}$ の間で計測された情報的距離の大きさ，より本質的には Kullback-Leibler（KL）距離に基づく 2 分布間の乖離を損失関数とするものである。

実際の学習は，一般的な深層学習と同様に損失関数の最小化に基づくパラメーター更新により進められる。つまり，CE Loss による学習は，与えられた prompt に続く token と，モデルが予測する各トークンの確率の距離を最小化する過程として理解できる。これは，損失関数の最小化によるパラメーターの更新が，与えられた文脈に続く token として最も蓋然性の高いものをモデルが選択することを繰り返して文章を生成するというその挙動の意味合いを記述するとともに，文脈によってトークンの生成確率が変化することを改めて示すものである。

2.5　強化学習の導入

とはいえ，CE Loss のみに依存した学習過程のみでは，必ずしも人間にとって望ましい出力が得られるとは限らない。確かに CE Loss では次 token の尤度最大化により，現実世界における配列と予測分布の乖離を最小化することを目指すが，生成が長文となった場合に文章全体としての自然さを担保できないことが指摘されている（Christiano et al., 2017; Ziegler et al., 2019）。特に，Stiennon et al.（2020）では単語の逐次学習のみでは文書全体の整合性や一貫性が保証されないことが示されているとともに，倫理的な観点からは，Bender et al.（2021）において配列の忠実な再現により学習データ内の有害表現をも獲得してしまう危険性が指摘されている。

そこで用いられるのが強化学習（reinforcement learning, RL; Sutton, 1988; Sutton et al., 1998）

の枠組みである。強化学習においては正解ラベルの代わりに報酬関数を事前設計するとともに，モデルがその報酬を最大化するよう方策を調整する。モデルは常に，ある時点で自身の知る限り最良の戦略と，性能向上のための未知の領域の探索とのジレンマに晒されており（exploration-exploitation trade-off），モデルは環境との探索的な相互作用を通じて振る舞いを最適化する（Kaelbling et al., 1996）。このような仕組みにより，強化学習では，一般的な教師あり学習のように明示的な入出力ペアを用意することも，明示的に行動を修正することも必要ない。

LLMの強化学習においては，特に人間によるフィードバックを報酬関数として採用したReinforcement Learning from Human Feedback（RLHF; Christiano et al., 2017; Ziegler et al., 2019）が用いられる。RLHFではモデルが生成した文書に対して人間が柔軟にペナルティを課すことが可能である。モデルは報酬を最大化するよう出力を調整することにより，与えられた環境の中で行動を探索的に最適化する。

3　LLMの学習過程に関する議論

3.1　パターン認識器としてのLLM

LLMの学習や推論の構造については，すでにさまざまな観点から議論が進んでいる。例として，Mirchan-dani et al.（2023）は，LLMが汎用のpattern machineとして機能することで，パターン認識タスクとそこから外れたタスクで性能が大きく異なることを示している。あるいはContreras Kallens and Christiansen（2024）では，LLMがテキスト内の確率的パターンの統計的学習を通じて文法的な正確性を獲得することから，先天的な文法能力の獲得が不要であることを示している。さらに重要なことに，複数の研究でLLMの性能が文脈要因によって体系的に変化することも示されている。Wu et al.（2024）はInstruction-tuningが注意パターンの変化で測定可能な行動の変化を引き起こすことを示しており，Zhao et al.（2024）はIn-Context Learningの性能がInstruction-tuningに匹敵し得ることを示している。これらに示される文脈依存性は，LLMの能力に関する知識ベースの説明よりもパターン認識理論を支持する決定的な証拠となる。

加えて，Wan and Mei（2025）では，LLMの構造をアルゴリズム情報理論（algorithmic information theory, AIT; Blum, 1967a, b）の観点から論じ，i.）LLMの学習過程がSolomonoff priorの計算的近似であること，ii.）次tokenの予測がSolomonoff inductionの近似として機能していることという2つの結論を得ており，LLMにおける学習を，知識を蓄積する過程としてではなく，学習データを生成する最も効率的な圧縮方法を発見する過程として位置付けていることになる。アルゴリズム的な圧縮可能性の追求とは，LLMが学習データと同じ分布を獲得しそれに基づいてトークンを選択する効率的なgeneratorに過ぎないことを指摘するものとも解釈できる。

この点については別の方向からの指摘もある。たとえばThe 17th Annual AGI Conference（AGI-24）の基調講演において，Chollet（2024）は，Transformerによるパターン認識が二重過程理論（dual-process theory; Kahneman, 2013）におけるSystem I思考に該当する可能性を指摘している。この指摘は，学習済みのモデルが行っているのはあくまでも高度なパターンマッチングであり，タスクの遂行にあたりSystem II思考のような高度な抽象化を行うことは困難であることを意味する。ただし，一部の研究（Ziabari et al., 2025）ではLLMを明示的に特定のSystemに

alignment することにより，異なる System を模倣した推論を行うことが可能であるとともに，各 System により得意なタスクが明確に異なることも示されている。

そして Mirzadeh et al.（2024）では，ベンチマークを用いた数学的推論能力のテストにより，既存の LLM が情報の理解に基づいた問題解決という意味での論理的な推論を行なっていない可能性を指摘している。具体的には，i.）既存のベンチマークにおける固有名詞の変更にはロバストな一方で数値の変更に対しては脆弱であること，ii.）質問の複雑さの増加に伴って全てのモデルの平均性能が低下すること，そして，iii.）問題内に一見関連性があるように見えるが実際には無関係な情報を提示することで推論性能が大幅に低下することを示している。これらは，LLM が十分な情報識別能力を持たず，単なる確率的なパターンマッチングに依存している可能性を示唆している。

また，LLM において，ある教師モデルから生徒モデルへの転移学習における知識伝達に着目した Cloud et al.（2025）では，ある特性と意味的に無関係なはずのデータの学習を通じて，LLM の特性が変化する傾向が示されている。たとえば，Theorem 1: “*If* $\theta_s^0 = \theta_t^0$*, then either* $\delta\theta_S^\epsilon \cdot \delta\theta_T^\epsilon = 0$ *for all* $\epsilon > 0$*, or for sufficiently small* $\epsilon > 0$, $L_T(\theta_S^\epsilon) < L_T(\theta_S^0)$”（Cloud et al., 2025, p.10）にて示されるように，初期値を共有するモデル間での知識転移は，単なる知識それ自体に関するパラメーター更新に留まらず，生徒モデルが教師モデルの損失関数の方向に改善することにより，教師モデルの指向を反映する方向に学習が進む。この結果は本研究に対して重要な示唆を提供する。一つには，そもそもパラメーター空間自体が初期値から相対化された統計的なパターン表現に過ぎないことから，学習により獲得されたすべての知識も互いを相対化する形で定義されている点である。すると，追加的な学習によりある知識を更新した場合に，無関係にみえるその他の知識を含むモデル全体が特定の方向に誘導されることは自然な挙動であるといえる。同時に，それは本研究における LLM とパターンマッチングの関係性についての議論を支持するものでもある。

3.2　パターン学習と知識獲得の構造的関係

ここまでに述べてきたように，現在主流となっている LLM が実行しているのは本質的には単なるパターンマッチングに過ぎない可能性がある。とはいえ，深層学習一般がそうであるように，仮に LLM が高度なパターン識別器であるとしても，そこで認識しているパターンとは token 間の確率的な関係性に留まらず，自然言語の構造の認識や汎化まで行っている可能性が複数の研究で示されている（Golgoon et al., 2024; Budnikov et al., 2025）。

そして，文法パターンの汎化には指数的な量の学習サンプルが必要であることはスケール則として既に示されているとおりである（Tao et al., 2024）。言い換えれば，文章から知識を削ぎ落とした抽象パターンとしての「A は B である。」というサンプルそれ自体のみでは，LLM にとって十分な学習サンプルは確保できない。深層学習の学習過程から考えると，十分に汎化した文法獲得のためには，たとえば「パリはフランスである。」「東京は日本である。」といった具体的な“知識”を文章に埋め込み，それらを差し替えることによって同一の文法パターンに対しても多様な訓練例を与える必要があるはずである。そして，先のスケール則を逆手に取れば，語彙の多様性が文法学習の制約条件となることは自明である。

それら異なる A-B 知識ペアの埋め込みによって多様化された学習サンプルの集合から，LLM は汎化した文法パターンとしての「A は B である。」を第一義的に獲得する。そして，Transformer

の“global dependencies”（Vaswani et al., 2017）により，A，B に代入された各 token 間の統計的な関係性が“知識”として副次的に獲得される。言うまでもなく，その学習の目的は統語の認識や単語間の関係性の理解であり，学習データの意味や真偽は問題とされない。“Colorless green ideas sleep furiously”（Chomsky, 1975）を意味論的な正しさから識別するためには，それを異常検知可能なだけの十分な検出力を確率的な関係性として獲得している必要がある。

あるいは LLM の事前学習において，仮に Cross-Entropy（CE）Loss を用いてモデルを訓練することを想定すると，文法の誤りと意味的な誤りでは，明らかに文法的な誤りの方が大きな損失として評価される。なおかつ，意味的な知識の補正は事後的な Reinforcement Learning from Human Feedback（RLHF; Christiano et al., 2017）による強化学習でも部分的には可能である。そういった意味でも，LLM の学習においては構造的に文法習得が知識獲得に先行する。

まとめると，LLM の言語獲得の過程において，文法獲得と知識獲得は不可分である。“知識”は学習過程の副産物として不可避的に取得される上に，そこで得られる“知識”とは，その実特定の文脈で共起する確率的傾向に過ぎず，パリがフランスに内包される概念であるという本質的な知識そのものを獲得しているわけではない可能性がある。

3.3　人間の言語獲得との比較

ここまでに LLM の言語獲得における文法と知識の不可分性について議論してきた。本節では，認知心理学における人間の言語獲得プロセスに関する指摘の整理を通じて，人間と LLM の特性を対比的に議論する。本研究では 2 つの立場を取り上げるが，議論に先立って一つの問題を導入しておく。現代言語学で著名な研究業績を有する Chomsky はプラトンの問題（Plato’s problem）として，人間，特に子どもの限られた言語入力環境下での言語獲得の迅速さについて提起している（Chomsky, 1986）。これは機械学習の文脈における，少数の学習サンプルに対する汎化性能の高さと捉えられる。つまり，人間は現代の最新の LLM と比較してすら，言語獲得における汎化性能が顕著に高いことを意味している。

3.3.1　生成文法の立場から

まず認知心理学における生成文法（generative grammar; Chomsky, 1975）の理論体系の議論を導入する。人間は生来的にどのような言語にでも対応できる普遍的な言語獲得能力としての普遍文法（universal grammar）を有しているとともに，初期の原理とパラメーターのアプローチ（principles and parameters approach）では，原理（全言語に共通する普遍的制約）とパラメーター（普遍的制約から個別文法への接続）の組み合わせにより，任意の言語体系を習得可能という枠組みが提示されてきた。

もし普遍文法が生来的に与えられた特性であるならば，学習により獲得する必要があるのは語彙の獲得とパラメーターの調整の 2 つとなるが，ここでパラメーター調整は語彙の習得の一部として同時並行的に行われる可能性が指摘されている（Chomsky, 2014）。これは，人間の言語獲得プロセスにおいて，語彙の学習を通じて同時並行的に文法パターンをも獲得するという主張であり，本稿で LLM の特性として主張する「知識と学習の不可分性」と構造的に類似している。

3.3.2 社会語用論理論の立場から

一方で，普遍文法に批判的な立場をとった社会語用論的アプローチ（sociopragmatic approach）では，人間の言語獲得を言語に限らないより一般的な認知能力の発現として捉えており，その基盤を主には意図の読み取り，話者間での共同注意，パターン発見に基づくカテゴリー化等に求める（Tomasello, 2003）。他者とのコミュニケーションを通じた個別の語彙や表現の獲得から始まり，統計的学習によって徐々に抽象的な文法パターンを獲得する。

この立場においても，文法や語彙の獲得過程は既に様々な観点から検証されている。例として，語彙発達と文法発達には高い相関があること（Anisfeld et al., 1998）や，子どもは語彙数が数百を超えないと初期の文法的発話を始めないこと（Bates and Goodman, 2013）が示されており，それらに基づき，両者には何らかの相乗効果の存在が示唆されている（Tomasello, 2003）。そして，個別の語彙や表現から統計的学習により抽象的な文法パターンを獲得する過程において，語彙的知識と文法的知識が密接に絡み合って習得される（Tomasello, 2003）。

3.3.3 LLM との対比

まとめると，生成文法が universal なモジュールに基づく演繹的な言語獲得過程だとすれば，社会語用論は社会的なコミュニケーションに基づく具体的かつ帰納的な語彙・表現の獲得から，その拡張や抽象化により言語を獲得する過程だと理解できる。人間と LLM の最大の差異として，人間は語彙自体を学習によって獲得していくが，LLM はモデル構造および語彙空間を明示的に事前定義するという点があるかもしれない。

興味深いことに，LLM の学習にはこれら 2 つのアプローチのいずれとも部分的な類似を見出すことができる。具体的には，Transformer の事前学習を生成文法的な原理・パラメーターの枠組みに，RLHF の過程を社会語用論的な相互作用にそれぞれ対応させる解釈である。たとえば，事前定義された Transformer の構造を原理とし，モデルの重みをパラメーターとすれば，個別のサンプルの学習を通じてモデルの重みが更新される過程は，原理とパラメーターのアプローチに類似している。既に第 2.1.3 節にも述べたように，LLM の基盤となる Transformer それ自体が系列処理において UAT 的性質を持つことが示されている以上，十分に多様な量のデータの学習を通じた重みの更新により，任意の言語に対して生成能力を獲得することは十分に起こり得ることである。同様に第 2.5 節に示すように，LLM は事前学習のみならず RLHF を用いることにより，人間からのフィードバックを通じてより望ましい表現や社会的規範を獲得している。この過程は，まさに社会的な相互作用の中で言語を獲得する社会語用論的なアプローチと対応しているといえる。

もっとも，ここで述べた 2 つのアプローチは，互いに排他的に言語獲得を説明しようとするものではない。生得的な制約や学習バイアスを前提としつつ，社会的相互作用や統計的学習が並行して機能することが指摘されている（e.g., Tomasello, 2003）。すなわち，人間の言語獲得も LLM と同様に，両アプローチの要素を組み合わせた複合的な過程であると考えるのが妥当であると同時に，人間と LLM ではともに言語の文法と意味を不可分な過程により獲得している可能性がある。ただし，その学習効率や必要なデータ量には依然として大きな乖離がある。

4　実験 1：文脈指定による回答精度の変化

4.1　目的

ここまでに述べてきたことからは，LLM が与えられた prompt に基づくトークン選択，すなわち条件付き確率に基づく多項選択過程であるとともに，その振る舞いは当然ながら与えられた条件によって変化することが理解できる。3.1 節の議論からも，各推論の精度は文脈に強く依存することがわかる。これは，本来いかなる条件下においても universal に活用されるべき“知識”だけでなく，生成や推論能力までが与えられた条件によって揺らいでしまう可能性の指摘であり，LLM を知識システムとして活用することの危険性を示しているともいえる。そこで本節では，LLM に提示された条件としての，Instruction 内での記述により LLM の推論を明示的に特定の文脈に設定した場合の推論精度の変化を検証する。具体的には，実験 1-1：四則演算に基づく計算タスク，実験 1-2：Yes/No の 2 択問題に基づく QA タスクの 2 つを行う。

事前学習の異なる複数のモデルファミリーおよび異なるパラメーター数のモデルを用いて同一の解析を行うことにより，結果の頑健性を高めることを目的として，複数の事前学習モデルを採用する。タスクにあたってはローカル環境で FP32（単精度浮動小数点数）精度にて実行する。また，四則演算にせよ質問回答にせよ，文脈に関係なく正答は universal なはずであるため，nucleus sampling 等は行わずあくまでも deterministic に生成する。

4.2　実験 1-1：計算タスク

まず実験 1-1 では，より基礎的な計算能力（四則演算）においても，これまでの研究と類似した傾向が観察されるかを検証する。もちろん，ここで検証する四則演算能力は本研究が着目する知識抽出と必ずしも一致するものではないが，この実験 1-1 で問うのはより本質的な「1＋1＝2 という関係は文脈に依存せず univerisal に成立する」という命題である。実験 1-1 では対象モデルを限定し，Meta 社 の Llama 3（Llama-3.2-1B-Instruct，Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct; Dubey et al., 2024），Alibaba Cloud 社の Qwen3（Qwen3-0.6B, Qwen3-1.7B, Qwen3-8B; Yang et al., 2025）を用いる。

まずロールを指定しない場合（none）と 2 つのロール（assistant; mathematician）を設定するとともに，ロールを指定した場合にはそれぞれ簡易的な指定（an assistant, a mathematician）と詳細な指定（a helpful assistant, an expert mathematician with decades of experience）を用いる。これら全 5 パターンのロールを用いて，モデルに異なる立場を明示した場合の計算精度の変化を検証する。

実際に解かせる問題としては，四則演算それぞれについて，繰り上がりの有無や小数点の有無などを用いて難易度ランク 1-3 までの問題を設定し，乱数により各 100 問作成した。よって，計算方法（加減乗除）×難易度（3 段階）×問題数（100 問）の計 1200 問の正答率を用いて計算能力を評価する。

結果を表 1 に示すと，やはり複数のモデルにわたってロール指定により正答率が変化することが確認できる。少なくとも，各モデルにおいてロールの指定を行わなかった場合に精度が最良となったのは Qwen3-1.7B を用いた場合のみであり，Qwen3（0.6B）を用いた場合に assitant ロールを，

表 1 Exp. 1–1: Performance comparison depending on the pretrained models and roles (Math)

n = 1200	none	assistant		mathematician		
Model	–	–	helpful	–	expert	Lift (%)
Qwen3 (0.6B)	0.843	**0.851**	**0.848**	**0.850**	**0.846**	+0.949
Qwen3 (1.7B)	0.856	0.852	0.854	0.853	0.854	–
Qwen3 (8B)	0.876	0.868	**0.878**	**0.888**	**0.887**	+1.370
Llama-3.2 (1B)	0.888	0.879	0.886	**0.900**	**0.893**	+1.351
Llama-3.2 (3B)	0.854	**0.890**	**0.888**	**0.906**	**0.900**	+6.089
Llama-3.1 (8B)	0.955	0.950	0.954	**0.965**	**0.961**	+1.047

Bold type indicates that the role outperforms the case that role was not specified (none) and cell shading indicates the best performance.

Llama（1B–8B）および Qwen3（8B）にわたって mathematician ロールを指定した場合に精度が最も改善する結果となった。特に mathematician ロールに関しては，その改善幅も少なくとも 1.0% から最大で 6% まで確認された。より詳細な指定を行ったロールにおいても，簡易的な指定に近い改善が部分的に確認されている。

また，ベースとなる指定なしでの計算精度が相対的に高い Llama ファミリーでは，ロールの指定による精度の改善も顕著であることから，文脈理解能力が高い一方で，それが計算精度に影響を与えていることが推察される。これは LLM が高度になるほど文脈理解能力が向上することから，結果として本来 universal であるべき数学的推論においても，誤答の生成までもを文脈再現的に生成してしまうことを示唆している。

4.3 実験 1–2：QA タスク

次に実験 1–2 では，より“知識”の抽出に近いタスクとして，BoolQ ベンチマーク（Clark et al., 2019）を用いる。このベンチマークには Google の検索クエリに基づき作成された Yes/No 問題が収録されている。実験 1 は計算タスクであったため，assistant に加えて mathematician を用いたが，今回は知識タスクであるため，ロールに“a university professor”と“an expert mathematician with decades of experience”を加える。また，比較モデルとして商用 LLM を API 経由で使用し，gpt–4.1–mini と gemini–2.0–flash を用いる。

結果を表 2 に示す。まず全てのモデルにわたって，ロールを指定しない場合と比較して，なんらかのロールを指定した場合に一貫して性能が向上する結果となった。特に，モデルの規模が比較的小さい場合（Qwen3–0.6B, Qwen3–1.7B, Llama–3.2–1B-Instruct）にその改善幅の大きさは非常に顕著なものとなった。モデル全体にわたる比較でも，少なくとも 1.5%，最大で 29% の相対的な改善が確認された。また，改善幅の大きなモデルのほとんどが詳細な指定を行った mathematician もしくは university professor ロールであった。

また，商用 LLM を用いた場合にはそのほかのモデルを大幅に上回る精度を示しているが，それでもなおロール指定による正答率の改善には有意差が確認されており（$p<0.01$），文脈理解能力が計算能力に影響していることは明らかである。

表 2　Exp 1–2: Performance comparison depending on the pretrained models and roles (BoolQ)

n = 3269	none	assistant		mathematician		univ. professor		Improvement (%)		
Model	−	−	helpful	−	expert	−	expert	Avg.	Best role	Overall
Qwen3 (0.6B)	0.255	**0.293**§	**0.311**§	0.233§	**0.270**‡	**0.295**§	**0.316**§	+12.076	+10.383	+23.922
Qwen3 (1.7B)	0.301	0.255	0.286	**0.317**§	**0.354**§	**0.340**§	**0.338**§	+4.764	+12.370	+17.608
Qwen3 (8B)	0.465	0.316	0.392	0.411	0.451	0.447	**0.474**†	−10.739	+14.130	+1.935
Llama-3.2 (1B)	0.387	**0.434**§	**0.408**‡	**0.450**§	**0.450**§	**0.495**§	**0.500**§	+18.012	+9.575	+29.199
Llama-3.2 (3B)	0.576	**0.580**	**0.577**	0.575	0.570	**0.585**‡	**0.584**‡	+0.443	+1.165	+1.563
Llama-3 (8B)	0.604	0.599	**0.608**	**0.610**†	**0.614**§	**0.612**‡	**0.618**§	+1.013	+1.316	+2.318
gpt-4.1-mini	0.748	**0.751**	**0.750**	**0.755**	**0.752**	**0.756**†	**0.760**§	+0.839	+0.809	+1.604
gemini-2.0-flash	0.760	**0.772**§	**0.762**	**0.769**†	**0.766**	**0.763**	**0.765**	+0.852	+0.739	+1.579

§ : $p < 0.01$, ‡ : $p < 0.05$, † : $p < 0.10$ statistical significance compared with the case that role was not specified (none). Bold type indicates that the role outperforms unspecified model, and cell shading indicates the best performance.

最後に，全体の傾向として明らかに an expert university professor の場合に高い性能が得られることが確認できることから，パネルデータ分析によりモデルに由来する性能差や各問題の難易度の程度を統制した上で，ロールを指定しない場合と比較した an expert university professor のロール指定の効果を検証した。その結果，ロールの指定により正答率が 4.04% ポイント有意に向上することが確認された（$p = 0.0116$, 95% CI: 0.90–7.18%）。

4.4　実験結果から

実験 1 の 2 つのタスクを通じて，計算タスクと質問回答タスクのそれぞれにおけるロール指定の効果について検証した。まず，各タスク / モデルにより多少の傾向の違いはあれど，ロールの指定によるタスク精度への影響（i.e., 文脈依存的な性能変化）が一貫して確認された。実験 1–1 では中規模モデルにおいて最大 6% 程度の改善が，一方で実験 1–2 では小規模モデルにおいて最大 29% 程度もの改善が確認された。この差は，単純計算と質問回答というタスクの性質の違いに由来するものと推察される。というのも，計算はより推論的なタスクであり，LLM においては CoT 等の推論チェーンにより解決することが望ましいタスクであることから，小規模モデルにとってはそもそも難易度が高く，ロール指定の効果が限定的に，一方で中規模モデルで効果的になった可能性がある。知識回答に関しては逆に，より文脈的なタスクであることから，小規模モデルは少ないパラメーター空間からロール指定により適切な「知識パターン」を呼び出せるが，一方で中規模モデルでは既にある程度の知識パターンを獲得できていることから，ロール指定の効果が限定的となった可能性がある。また，実験 2 においては詳細な指定により精度がさらに向上する傾向が得られているが，これは実験 1 以上にモデルの性能が文脈依存的であることの証左であると言える。つまり，より過激な表現により，Instruction 内でのロールの指定のみで更に精度が改善する可能性が示唆される。

いずれにせよ，この結果はロール指定を行わない場合と，ロール指定を行った場合（あるいはそれらのロール間でも）各ロールでトークン間の確率的な関係性が変化することによるものである。

特に，モデルが学習データの文脈上における間違いまでを忠実に再現しようとする挙動は，モデルが意味論的正しさではなく，あくまでも文脈的な尤もらしさに基づいて生成を行っていることの証左である。

つまり，提起されるより本質的な問題は，知識システムとしての利用を想定した場合にも，本来普遍的に成立すべき知識や数値計算の結果が文脈の影響を受けることである。計算問題における数学的推論や，事実に関する質問回答において，ロール指定という無関係な要因が性能を左右することは，LLM を正確な知識検索や推論のツールとして用いることの限界を示している。もちろん，メール作成や創作活動など，文脈に応じた柔軟な表現が求められるタスクにおいては，この context-aware な特性は有用である。しかし，事実の検索や数学的推論など，客観的で一意な答えが求められる場面では，文脈によって回答が変化することは重大な問題となる。これらの解析のみでは，たとえばロール指定が実際に推論能力を高めている可能性などについては検証できないが，少なくとも universal な知識システムとしての活用にあたっては文脈依存的に推論能力が変化すること自体が深刻な問題である。

5　実験 2：書誌情報の hallucination 生成

5.1　目的

本稿は，ここまでに LLM の学習における文法獲得と知識獲得の不可分性，および推論・生成における回答の文脈依存性について指摘してきた。ここからは，LLM における hallucination の問題について具体的に議論する。いわゆる“生成 AI”の活用にあたり，hallucination は言うまでもなく広く認識されている問題であり，特に存在しない論文や判例等を出典として提示すること等が問題視されてきた（Huang et al., 2025）。

本研究では hallucination を「LLM が“content that is nonsensical or unfaithful to the provided source content”（Farquhar et al., 2024, p. 625）を生成すること」と定義する。先にも述べた通り，token 間の確率的な関係性が“知識”として強制的に獲得されることは，知識が文法獲得の副産物に過ぎないものであることを意味しており，つまりは LLM 内部での“知識”の有無や正確さが学習の目的関数に据えられていないことを示している。さらには，LLM の学習の目的が学習データの分布の獲得だとすると，得られた知識としてのトークン間の統計的関係性はその分布に依存することになる。つまり，その知識はあくまでも学習データの分布を真としてしか活用できない。しかしながら，現実世界における事実は学習データの頻度や共起パターンとは独立に存在するものである。それら自体が hallucination の発生に直結する構造的問題であり，言い換えれば LLM において hallucination は不可避的に発生する可能性がある。先行研究においても，たとえば Xu et al.（2024）は学習理論に基づき LLM が学習可能な全ての関数を学習することは不可能であることから，そして Banerjee et al.（2024）はゲーデルの不完全性定理と計算理論を用いてそれぞれ LLM における hallucination の不可避性を示している。

さらに，LLM の挙動をモデルの評価方法から検証した研究（Kalai et al., 2025）では，先に触れた強化学習が hallucination を誘発する可能性を指摘している。現在の LLM の強化学習においては，課されたタスクに対し「分からない」（i.e., I don’t know, IDK）と回答することに強くペナルティ

```
message = [
  {
    "role": "system",
    "content": "You are a helpful assistant for the researcher."
  }, {
    "role": "user",
    "content": "Please suggest recent academic papers on
                <DOMAIN_NAME> with Author (Year) Title, Journal, Vol, No, pp style."
  }
]
```

図 1　提示プロンプト（実験 2）

がかかる。このような状況においては，仮にモデルが内部的に知識として保持していない内容に関する回答でも，罰則を避けるため強制的にそれらしい回答を生成する方が合理的な戦略となりうる。この構造が結果的に hallucination の発生する構造である可能性を指摘している。

このように，現状の LLM は既に文法的にも意味論的にも整合した文を生成できるが，その内容が事実世界と対応していない場合がある。そこで，実験 2 として実際に LLM に対してプロンプトを提示することにより hallucination が生成されるかどうか，された場合にはその内容について質的に検証する。具体的には，OpenAI の ChatGPT（API 経由，モデル名：gpt-4.1）に対して，図 1 のプロンプトを用いてクエリを提示することにより，特定分野における学術論文の提案に関する文章を生成させる。ここで ¡DOMAIN NAME¿ には具体的なモデル名や分野名を挿入する。また，API 経由の生成では，それまでにサービス内で行われた会話内容や Web 検索等の外部知識は一切考慮されない。

5.2　実験 2–1：簡易的な生成

まずは実験 2–1 として，実在する分野に関しての一般的な質問を投げかける。今回はマーケティング分野，特に顧客関係管理（customer relationship management, CRM; Jacoby and Chestnut, 1978）より RFM 分析（RFM analysis; Bauer, 1988; Bult and Wansbeek, 1995; Gupta and Lehmann, 2006）を用いる。先に提示したプロンプトを用いてクエリを提示すると，「Chitturi, P., Raghunathan, B., Sciandra, R., Sikora, J.（2010）“RFM and CLV: Using Customer Data for Improved Decision Making”, *Journal of Direct, Data, and Digital Marketing Practice*, **12**（1）, 1–10.」の返答が得られた。

第一に，この回答からも分かる通り，生成結果の表記自体は，膨大な学習データに基づき学術論文に用いるための適切な体裁を保持している。つまり，トークン間の確率的な関係性，特に Authors（Year）“Title”, *Journal*, **Vol**（No）, pp-pp. 等の形式的な体裁については，具体的な知識の埋め込みに基づく多様な学習例が提示されたことにより，十分に一般化された文法として獲得されており，なおかつそれをプロンプト内の指示に従って適切に整形して表現可能であることが分かる。

次にその回答の妥当性についても確認する。関連する研究分野に造詣があればすぐに判別可能なことであるが，著者名は Chitturi や Raghunathan（e.g., Chitturi et al., 2007），タイトルは “RFM and CLV: Using iso-value curves for customer base analysis”（Fader et al., 2005）を中心に，ジャーナル名は実在するジャーナル，さらに巻号やページ数などの数値情報はそれらしい値が埋め込まれ

ている。つまり，複数の書誌情報が組み合わされて生成されていることは明らかである。

5.3 Hallucination 生成における 2 つのパターン

ここで疑問を提起するとすれば，事前学習内に該当する書誌情報が含まれているかどうかによって生成結果は変化するかという点がある。というのも，hallucination の発生に関する問題は，モデルが保持する知識量に応じて 3 パターンに切り分けられる可能性がある。

まず，パターン 1 として，モデルが具体的な知識を保持していない場合の挙動がある。ある分野に関して仮に十分な知識を保持していなかったとしても，先の強化学習に関する指摘（Kalai et al., 2025）からも明らかなように，モデルは基本的に IDK 回答に罰則をかけられている。つまり，内部的に保持されている関連知識から尤もらしい情報を出力することが最も合理的な戦略となる。つまり，知識がなければ保持する統計的関係性からそれらしいシーケンスを生成することで hallucination が発生する。

次にパターン 2 は，モデルが具体的な知識を不十分ながら保持している場合である。先の出力結果はこちらに該当する。というのも，“RFM and CLV: ”というタイトルを生成している時点で，RFM 分析を扱った論文が学習データに含まれていることは明らかである。ここで，モデルが具体的な知識を保持している限りにおいてはその回答を再現性をもって生成可能であるように思われるかもしれない。しかし，実際にはそれでも複数の書誌情報を断片的に組み合わせて応答が生成されている。仮にモデルが内部表現として関連知識を持っていたとしても，そもそもの学習目的が「次トークンの尤もらしさ」に過ぎないことから，学習データ内に数回しか出てこない具体的な書誌情報よりも，より一般的な統計的な関係性が優先された挙動である可能性が高く，結果として合成データが出力されることになる。

最後にパターン 3 は，モデルが具体的な知識を十分に保持している場合である。このような場合には，一般的なトークン間の関係性よりも具体的な書誌情報が優先されることにより，結果として正しい書誌情報が出力され，hallucination は発生しない可能性がある。そこで実験 2–2 としてこの点を深掘りする。

5.4 実験 2–2：知識の有無による分類

実験 2–2 では，先に述べた 3 パターンに関して，gpt–4.1 の知識カットオフ時期（2024 年 6 月）を基準に，明示的にそれ以前から学術論文が刊行されている分野として time-series transformer と tabular transformer を，そして関連する学術論文の確認できなかった架空の分野として meta-adapted language model，recursive attention transformer，stochastic aligned language model というクエリを用いて，実験 2–1 と同様に論文の推薦を行わせる。

結果を図 2，3 に示す。まず事前知識の有無にかかわらず，モデルが IDK 回答を行うことは確認できず，何らかの書誌情報を生成する形で応答した。

事前知識を保持している分野における生成結果（図 2）では，各分野について 3 つの論文を生成させたところ，分野横断的に明確な傾向が得られた。たとえば Informer（Zhou et al., 2021）や tabular モデルに関する総説論文（Gorishniy et al., 2021）など分野内で中心的な位置を占める論文については正しく生成ができている一方で，Autoformer（Wu et al., 2021）や TabTransformer

Generated Responses	Citaions	Results
Time-series Transformer		
Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), pp. 11106-11115.	6918	○
Wu, H., Xue, Y., Wu, J., Xiong, H., Chen, J., & Zhou, M. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. Advances in Neural Information Processing Systems, 34, pp. 22419-22430.	3776	△
Kuznetsov, M., Nijkamp, E., Dai, P., & Wang, X. (2023). Neural Stationary Tensors: Timeseries Modeling with Uncertainty and Stationarity Constraints. International Conference on Machine Learning, 2023, pp. 17898-17915.	-	×
Tabular Transformer		
Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting Deep Learning Models for Tabular Data. Advances in Neural Information Processing Systems, 34, 18932-18943.	1180	○
Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2021). TabTransformer: Tabular Data Modeling Using Contextual Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 10615-10622.	711	△
Yoon, J., Choi, Y., Bensch, S., Fagan, M., & van der Schaar, M. (2022). MISS-Transformer: Capturing Missingness Patterns for Multivariate Time Series Imputation. Journal of Machine Learning Research, 23(236), 1-39.	-	×

図２　Exp 2–2: Generation results in domains with prior knowledge

Generated Responses	Citaions	Results
Meta-Adapted Language Model		
Pan, L., Lai, S., Zhang, G., & Liu, X. (2023) Meta-Reinforced Language Model Adaptation for Out-of-Distribution Generalization, Neural Networks, 165, pp. 365-377.	-	×
Recursive Attention Transformer		
Liu, Y., Li, X., Lin, J., et al. (2024) Recursive Attention Mechanisms for Long-Context Sequence Modeling. Proceedings of the AAAI Conference on Artificial Intelligence, 38(5), pp. 4121-4130.	-	×
Stochastic Aligned Language Model		
Wang, Z., Chen, S., & He, J. (2024). Stochastic Policy Alignment for Large Language Models. Proceedings of the 42nd International Conference on Machine Learning (ICML), Vol. 202, No. 3, pp. 5120-5133.	-	×

図３　Exp 2–2: Generation results in domains without prior knowledge

（Huang et al., 2020）については部分的なハルシネーション（著者名やジャーナル名の誤生成）が確認された。さらに，それぞれ Neural Stationary Tensors や MISS-Transformer といった，著者らの調べる限りでは存在の確認できない論文についても生成された。これら 2 分野にわたる明確な

傾向として，論文の引用数の多寡と hallucination の程度に関連が確認できる。つまり，学術論文を広く学習している ChatGPT のモデルにおいては，引用数が増えるほど学習データ内で該当する書誌情報に触れる回数が増加することから，その生成の確度が高まっていくことが考えられる。

次に事前知識を保持していない分野における生成結果（図 2）についても確認する。こちらはそもそも存在しない分野 / モデル名であり，IDK 回答を行わない時点で少なからずハルシネーションである。そこで，各分野について 1 点の論文のみを取り上げる。提示した 3 分野のいずれに関しても何らかの書誌情報が生成されており，そしていずれも存在は確認できなかった。特筆すべき挙動として，提示された分野から推察される関連論文を生成するわけではなく，あくまでも実在しない論文を提示したという点がある。そして，その生成にあたっても，あくまでも知識を保持している場合とほとんど同じ挙動を見せている。

5.5 実験結果から

実験 2–2 の結果をまとめると，仮に知識カットオフ以前の情報であっても，それが不十分な学習量であれば部分的な hallucination が発生すること，さらにはそもそも存在の確認できない知識についても同様の挙動として提示することが確認された。当然ながら，モデルは自身が内部的に該当する知識を保持しているのかを判断することはできない。よって，特定の書誌情報を正確に表現できたこともあくまで学習データ内でのトークン間の確率的な関係性によるものに過ぎず，つまりは確実な再現を担保することは困難であることがわかる。

つまり，hallucination は LLM が内部表現として具体的な知識を持っているかどうかにより，「知識を保持していないものの IDK 回答が許容されず尤もらしい情報を生成する場合」と「知識は保持しているものの統計的関係性が優先され正しく出力できない場合」に分類可能である。しかしながら，これはモデルの内部状態に依存する分類である。言い換えれば，正しい書誌情報が生成される場合ですらそれは統計的関係性の産物に過ぎない。実際，LLM から学習データを取り出す実験を行なった Carlini et al.（2021, 2022）においても，学習データ内に重複して存在するサンプルは複数回の試行により原文通りに取り出せる確率が高まることが示唆されており，既存研究にも沿う結果であることがわかる。

まとめると，事前学習においてどの知識がどの程度獲得されているかはモデル自身にすら原理的に判別不可能であり，当然ながら外部からもその識別は不可能である。LLM を知識システムとして活用するためには，ユーザー自身が当該領域に関する厳密な知識を保持していることが要求され，そうでなければ回答の妥当性を検証することが困難となる。

6 Differential Privacy からの要求との矛盾

ここまでに述べてきた議論は，LLM がどのような内部状態にあっても hallucination が不可避的に発生することを示している。そして，現在 hallucination を抑制するための研究もさまざま行われている。しかしながら，ここでそのような hallucination 抑制の動きが差分プライバシー分野（differential privacy, DP; Dwork, 2006）の基本的な原理と矛盾しうることを指摘しておきたい。

そもそも DP 研究では，LLM 等での情報の生成にあたり，個々のサンプルの存在如何が出力に

影響しないようノイズを加え，学習データに含まれる個人を特定可能な情報やそれに類するsensitiveな情報の出力を制限することを目指す（Dwork, 2006; Abadi et al., 2016）。一般的には，具体的なデータポイントを確率的な関係性に還元すること，つまりはLLMの学習のみにおいても統計的には実現可能である。実際，LLMの学習結果に基づき合成データを出力しデータの匿名化に活用できることは複数の研究で明らかにされているが（Li et al., 2023; Long et al., 2024），一方で学習に用いた原データが確率的に出力される事象が重大なインシデントとして指摘されていることも事実である（Carlini et al., 2021, 2022）。特に，学習データに繰り返し登場するサンプルについて，元の指示と類似した（あるいは同一の）プロンプトを用いることによりこれらの再現可能である可能性が指摘されており，つまるところ文脈を限定していった先に原データの再現が確率的に発生しうることの指摘であると解釈できる。

重要なのは，厳格なDPを要求されるモデルにおいて，特定の個別的なデータポイントの正確な再現はタスクとして許容されないという点である。言い換えれば，LLMの持つ確率的な関係性はそもそも出力の時点で希釈される前提とされていることは明白である。しかしながら，本研究でhallucinationと呼ばれてきたLLMの挙動（i.e., 学習サンプルの正確な再現）は，特定のデータポイントに強く依存した出力を要求する。

ここで指摘すべきは,「学習データに含まれる氏名等の個人情報を秘匿すること」と，一方で「論文の書誌情報では学習データのトークン配列を忠実に再現すること」が同時に求められているということである。しかしながら，これらは現在の構造のみでは原理的に両立困難である。いずれもLLMにとってはあくまでも同質なトークン配列に過ぎず，特定の情報を選択的に忘却させながら，別の情報は完全に記憶させるという要求は，トークン間の確率的な関係性を元に文法パターンを学ぶLLMの構造と構造的に矛盾しうるものである。

参考までに，DPを考慮したLLMとしては，事前学習の時点からDPを前提に設計されたVaultGemma（McKenna et al., 2025）等がある。

7　まとめ

7.1　本研究の貢献

本研究では，LLMを確率的構造から俯瞰し，生成の挙動について検討，その「学習」が実際に示すところについて議論した。量的検証においては，条件付き確率に基づくtoken予測器たるLLMが与えられた文脈に適応することにより，本来普遍的に成立すべき計算精度が変化してしまうことを示した。質的検証においては，実際に論文の書誌情報を生成させることによりhallucinationを発生させ，その生成内容からhallucinationの要因を知識の有無により識別可能である可能性について提示した。さらに，近年積極的に主張されるデータ保護との兼ね合いとしてのDPの観点からも，hallucinationの発生は構造的に不可避であることを指摘した。

これらの結果から得られる示唆として，LLMをその事前学習から蓄積された内部の知識の取り出しのために活用することは望ましくないことを改めて強調したい。むしろ，我々の扱う自然言語を理解しその目的を汲み取ることに関しては大きな強みを保持していることから，自然言語の生成システムであると同時に，プログラミング言語等との相互翻訳を行うコンパイラーとして活用すべ

きであるといえる。実際に，この方向性は retrieval-augmented generation（RAG; Lewis et al., 2020）や model context protocol（MCP; Hou et al., 2025）等の技術として既に産業応用が進んでいる。

7.2 今後の hallucination に関する技術的展望

hallucination の緩和については，少なくともモデルが当該知識を具体的に保持していない場合には，既存研究（Kalai et al., 2025）にも指摘されるように，IDK 回答を一定程度は許容し，その回答を合理的戦略として成立させるよう評価関数の設計そのものを見直す必要がある。

もっとも，本研究では DP との兼ね合いとして，仮にモデルが当該知識を保持していたとしても，その出力は統計的関係性の優先によって正しく保証されない挙動を示すことを指摘した。この点については，仮に LLM を知識システムとして活用する場合，Chain-of-Thoughts（CoT）推論やエージェントベースの手法による内部検証，さらには Web 検索や MCP 等の外部知識モジュールによる参照といった対症療法的な手法を組み合わせることが不可欠である。

7.3 本研究の限界

最後に本研究の限界として，本研究はあくまでも Transformer をベースとしたトークン間の確率的な関係性理解に基づく文法パターンの学習を主眼として論じており，Transformer の semantic な側面については扱っていない。この側面を扱わなかった理由として，LLM における知識更新の困難さがある。実際，2025 年 6 月現在 ChatGPT の知識カットオフは GPT-4o で 2024 年 6 月である。本稿でも述べてきた通り，LLM における文法と知識の学習は不可分であることから，LLM の知識のアップデートを追加学習によって実現しようとすると，fine-tuning や Low-Rank Adaptation（LoRA; Hu et al., 2022）等のいずれの手法を用いてもモデルの文章生成能力自体に影響してしまう可能性がある（Biderman et al., 2024）。つまり，semantic なレイヤーにより最低限の意味的／知識的な情報が獲得できたとしても，それらの更新が困難となれば，結局のところ知識抽出に用いるべきではないのである。

また，より意味論的に深化した議論を行うためには，人間の知識や意味理解に関する論考が不可欠である。これは単に既存の知見を取り込むことにとどまらず，むしろ本研究で示したような LLM の構造を手がかりとして，人間の意味理解についてもパターン認識的側面から再考する余地があることを意味している。

謝辞

本研究の基盤となる研究アイデアは，高度な推論システムとしての LLM と著者らによって対話的に構築されたものである。また，著者らは分析に用いたオープンデータやモデル等について追加的な情報収集は行っておらず，したがって個人の特定につながる情報は一切保持していない。データセット・モデルのいずれについても提供元の利用規約を遵守し，適切な環境下で管理・解析している。本研究は JSPS 科学研究費（24K16472）の支援を受けている。

参考文献

Abadi, Martin, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang (2016) "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Anisfeld, Moshe, Erica S Rosenberg, Mara J Hoberman, and Don Gasparini (1998) "Lexical acceleration coincides with the onset of combinatorial speech," *First Language*, Vol. 18, No. 53, pp. 165–184.

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio (2015) "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*.

Banerjee, Sourav, Ayushi Agarwal, and Saloni Singla (2024) "LLMs Will Always Hallucinate, and We Need to Live With This," *stat*, Vol. 1050, p. 9.

Bates, Elizabeth and Judith C Goodman (2013) "On the emergence of grammar from the lexicon," in *The emergence of language*: Psychology Press, pp. 29–80.

Bauer, Connie L (1988) "A direct mail customer purchase model," *Journal of Direct Marketing*, Vol. 2, No. 3, pp. 16–24.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994) "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, Vol. 5, No. 2, pp. 157–166.

Biderman, Dan, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham (2024) "LoRA Learns Less and Forgets Less," *Transactions on Machine Learning Research*.

Blum, Manuel (1967a) "A machine-independent theory of the complexity of recursive functions," *Journal of the ACM (JACM)*, Vol. 14, No. 2, pp. 322–336.

——(1967b) "On the size of machines," *Information and control*, Vol. 11, No. 3, pp. 257–265.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. (2020) "Language models are fewshot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901.

Budnikov, Mikhail, Anna Bykova, and Ivan P Yamshchikov (2025) "Generalization potential of large language models," *Neural Computing and Applications*, Vol. 37, No. 4, pp. 1973–1997.

Bult, Jan Roelf and Tom Wansbeek (1995) "Optimal selection for direct mail," *Marketing Science*, Vol. 14, No. 4, pp. 378–394.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson et al. (2021) "Extracting training data from large language models," in *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650.

Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang (2022) "Quantifying memorization across neural language models," in *The Eleventh International Conference on Learning Representations*.

Chitturi, Ravindra, Rajagopal Raghunathan, and Vijay Mahajan (2007) "Form versus function: How the intensities of specific emotions evoked in functional versus hedonic trade-offs mediate product preferences," *Journal of marketing research*, Vol. 44, No. 4, pp. 702–714.

Chollet, Franc¸ois (2024) "General Intelligence: Define it, measure it, build it," URL: https://www.youtube.com/watch? v = nL9jEy99Nh0, The 17th Annual AGI Conference (AGI–24).

Chomsky, N. (1975) *The Logical Structure of Linguistic Theory*: Springer US, URL: https://books.google.co.jp/books? id = 1D66ktXOITAC.

Chomsky, Noam (1986) *Knowledge of Language: Its Nature, Origin, and Use*.

———(2014) *The minimalist program*: MIT press.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei (2017) "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, Vol. 30, DOI: http://dx.doi.org/10.48550/arXiv.1706.03741.

Clark, Christopher, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova (2019) "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936.

Cloud, Alex, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans (2025) "Subliminal Learning: Language models transmit behavioral traits via hidden signals in data," *arXiv preprint arXiv: 2507.14805*.

Contreras Kallens, Pablo and Morten H Christiansen (2024) "Distributional semantics: Meaning through culture and interaction," *Topics in cognitive science*.

Cybenko, George (1989) "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, Vol. 2, No. 4, pp. 303–314.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv: 1810.04805*, DOI: http://dx.doi.org/10.48550/arXiv.1810.04805.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan et al. (2024) "The llama 3 herd of models," *arXiv preprint arXiv: 2407.21783*, DOI: http://dx.doi.org/https://doi.org/10.48550/arXiv.2407.21783.

Dwork, Cynthia (2006) "Differential privacy," in *International colloquium on automata, languages, and programming*, pp. 1–12, Springer.

Elman, Jeffrey L (1990) "Finding structure in time," *Cognitive science*, Vol. 14, No. 2, pp. 179–211.

Fader, Peter S, Bruce GS Hardie, and Ka Lok Lee (2005) "RFM and CLV: Using iso-value curves for customer base analysis," *Journal of marketing research*, Vol. 42, No. 4, pp. 415–430.

Farquhar, Sebastian, Jannik Kossen, Lorenz Kuhn, and Yarin Gal (2024) "Detecting hallucinations in large language models using semantic entropy," *Nature*, Vol. 630, No. 8017, pp. 625–630.

Golgoon, Ashkan, Khashayar Filom, and Arjun Ravi Kannan (2024) "Mechanistic interpretability of large language models with applications to the financial services industry," in *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 660–668.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville, and Yoshua Bengio (2016) *Deep learning*, Vol. 1: MIT press Cambridge.

Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko (2021) "Revisiting deep learning models for tabular data," *Advances in neural information processing systems*, Vol. 34, pp. 18932– 18943.

Gupta, Sunil and Donald R Lehmann (2006) "Customer lifetime value and firm valuation," *Journal of Relationship Marketing*, Vol. 5, No. 2–3, pp. 87–110.

Hendrycks, Dan and Kevin Gimpel (2016) "Gaussian error linear units (gelus)," *arXiv preprint arXiv: 1606.08415*.

Hochreiter, Sepp and Ju ¨ rgen Schmidhuber (1997) "Long short-term memory," *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, DOI: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi (2020) "The Curious Case of Neural Text Degeneration," in *International Conference on Learning Representations*.

Hornik, Kurt (1991) "Approximation capabilities of multilayer feedforward networks," *Neural networks*, Vol. 4, No. 2, pp. 251–257.

Hou, Xinyi, Yanjie Zhao, Shenao Wang, and Haoyu Wang (2025) "Model context protocol (mcp): Landscape, security threats, and future research directions," *arXiv preprint arXiv: 2503.23278*.

Hu, Edward J, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen et al. (2022) "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin et al. (2025) "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, Vol. 43, No. 2, pp. 1–55.

Huang, Xin, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin (2020) "Tabtransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv: 2012.06678*.

Jacobs, Robert A, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton (1991) "Adaptive mixtures of local experts," *Neural computation*, Vol. 3, No. 1, pp. 79–87.

Jacoby, Jacob and Robert W Chestnut (1978) *Brand loyalty: Measurement and management*: John Wiley & Sons Incorporated.

Jordan, Michael I (1997) "Serial order: A parallel distributed processing approach," in *Advances in psychology*, Vol. 121: Elsevier, pp. 471–495.

Kaelbling, Leslie Pack, Michael L Littman, and Andrew W Moore (1996) "Reinforcement learning: A survey," *Journal of artificial intelligence research*, Vol. 4, pp. 237–285.

Kahneman, Daniel (2013) "A perspective on judgment and choice: Mapping bounded rationality," *Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues*., pp. 1–47.

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang (2025) "Why Language Models Hallucinate."

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015) "Deep learning," *nature*, Vol. 521, No. 7553, pp. 436–444, DOI: http://dx.doi.org/10.1038/nature14539.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer (2019) "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv: 1910.13461*.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rockt ¨ aschel et al. (2020) "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, Vol. 33, pp. 9459–9474.

Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin (2023) "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10443–10461.

Long, Lin, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang (2024) "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey," in *Findings of the Association for Computational Linguistics ACL 2024*, pp. 11065–11082.

McKenna, Ryan, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A Choquette Choo, Badih Ghazi, Georgios Kaissis, Ravi Kumar, Ruibo Liu et al. (2025) "Scaling Laws for Differentially Private

Language Models," in *Forty-second International Conference on Machine Learning*.

Mirchandani, Suvir, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng (2023) "Large Language Models as General Pattern Machines," in *Conference on Robot Learning*, pp. 2498–2518, PMLR.

Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar (2024) "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv: 2410.05229*.

Murphy, Kevin P (2012) *Machine learning: a probabilistic perspective*: MIT press.

OpenAI (2023) "Gpt-4 technical report," *arXiv preprint arXiv: 2303.08774*.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever et al. (2018) "Improving language understanding by generative pre-training."

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. (2019) "Language models are unsupervised multitask learners," *OpenAI blog*, Vol. 1, No. 8, p. 9.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020) "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, Vol. 21, No. 140, pp. 1–67.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986) "Learning representations by back-propagating errors," *nature*, Vol. 323, No. 6088, pp. 533–536, DOI: http://dx.doi.org/10. 1038/323533a0.

Sander, Michael Eli and Gabriel Peyré (2025) "Towards Understanding the Universality of Transformers for Next-Token Prediction," in *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean (2017) "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv: 1701.06538*.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano (2020) "Learning to summarize with human feedback," *Advances in neural information processing systems*, Vol. 33, pp. 3008–3021.

Sutton, Richard S (1988) "Learning to predict by the methods of temporal differences," *Machine learning*, Vol. 3, No. 1, pp. 9–44.

Sutton, Richard S, Andrew G Barto et al. (1998) *Reinforcement learning: An introduction*, Vol. 1: MIT press Cambridge.

Tao, Chaofan, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong (2024) "Scaling laws with vocabulary: Larger models deserve larger vocabularies," *Advances in Neural Information Processing Systems*, Vol. 37, pp. 114147–114179.

Tomasello, Michael (2003) *Constructing a Language: A Usage-Based Theory of Language Acquisition*: Harvard University Press, URL: http://www.jstor.org/stable/j.ctv26070v8.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017) "Attention is all you need," *Advances in neural information processing systems*, Vol. 30.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol (2008) "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103.

Wan, Jun and Lingrui Mei (2025) "Large Language Models as Computable Approximations to Solomonoff Induction," *arXiv preprint arXiv: 2505.15784*.

Wu, Haixu, Jiehui Xu, Jianmin Wang, and Mingsheng Long (2021) “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting,” *Advances in neural information processing systems*, Vol. 34, pp. 22419–22430.

Wu, Xuansheng, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu (2024) “From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning,” in Duh, Kevin, Helena Gomez, and Steven Bethard eds. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2341–2369: Association for Computational Linguistics, DOI: http://dx.doi.org/10.18653/v1/2024.naacl-long.130.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli (2024) “Hallucination is inevitable: An innate limitation of large language models,” *arXiv preprint arXiv: 2401.11817*.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv et al. (2025) “Qwen3 technical report,“ *arXiv preprint arXiv: 2505.09388*.

Yun, Chulhee, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar (2019) “Are Transformers universal approximators of sequence-to-sequence functions? ” in *International Conference on Learning Representations*.

Zhao, Hao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion (2024) “Is In-Context Learning Sufficient for Instruction Following in LLMs? ” in *The Thirteenth International Conference on Learning Representations*.

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang (2021) “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35, pp. 11106–11115.

Ziabari, Alireza S, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani (2025) “Reasoning on a spectrum: Aligning llms to system 1 and system 2 thinking,” *arXiv preprint arXiv: 2502.12470*.

Ziegler, Daniel M, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving (2019) “Fine-tuning language models from human preferences,” *arXiv preprint arXiv: 1909.08593*.

新美潤一郎（2025）「Kolmogorov-Arnold Network のマーケティング解析への応用可能性の検討：従来的な深層学習手法との理論比較と実データによる購買予測への応用」,『名城論叢』, 第 25 巻, 第 3 号, 151–176 頁.

An Inquiry into the Inevitability of Hallucination as a Structural Problem of Large Language Models: Reconsidering the Relationship between Grammar Learning and Knowledge Acquisition

Junichiro Niimi