# 生成 AI 関連論文の投稿到着過程に関する統計分析（1）

2025年5月20日

## 1　背景と課題設定

2022 年末の *ChatGPT* 公開以降，大規模言語モデル（LLM）および「生成 AI」研究は arXiv/OpenReview 等に爆発的な数のプレプリント（査読前論文）を生み出している．その急増は

- **査読リソースの逼迫:** 会議及びジャーナルでの査読者確保や オーバーフローを緩和する政策判断
- **投資/ロードマップ策定:** 企業・研究機関が研究開発費を配分する際の マクロ指標
- **メタ研究:** トレンド変化点（飽和 or 次の技術シフト）を 統計的に検出する手段

などの実務的影響をもつ．しかし，「投稿ペースが今も上がり続けているのか/すでに頭打ちなのか」 を定量的に評価した研究は多くない．

本レポートでは

$$\text{arXiv の日次投稿件数}\quad N_t$$

を確率過程として捉え，

1. 直近 30 日間と
2. 前年同時期（30 日間）

の 投稿率 を比較し，有意差の有無を検定する．ここでは「到着過程 $\{N_t\}$ はポアソン分布」と仮定する 最も簡易的なモデルから出発し，のちに過分散やトレンド項を導入する拡張を視野に入れる．

## 2　データ取得と前処理

本研究では，arXiv が提供する REST 形式の公式 API（version 0.2）を用いて， 生成 AI 関連論文のメタデータを取得した． 検索クエリには

```
("GPT" OR "LLM" OR "Generative AI") AND (cat:cs.CL OR cat:cs.AI)
```

を採用した． ここで `cs.CL`（計算言語学）と `cs.AI`（人工知能）を選択した理由は， 生成 AI の主要応用分野を網羅しつつ，たとえば `cs.LG`（機械学習）や `stat.ML` に含まれる統計的手法一般の論文を過剰に取得することを避けるためである． キーワードは 2023–2025 年の LLM 関連論文タイトルを独立に 5 000 件抽出し， unigram 出現頻度上位 30 語から 主題を最も特徴づけるトークンとして `GPT`, `LLM`, `Generative AI` を選定した （トピックモデルによる自動抽出は今後の改良点として残す）．

### 2.1　観測期間選定

比較対象として

$$\text{期間 A}: 2025\text{-}04\text{-}04 \le t < 2025\text{-}05\text{-}04, \qquad \text{期間 B}: 2024\text{-}04\text{-}04 \le t < 2024\text{-}05\text{-}04$$

の 2 つの 30 日窓を設定した． 30 日幅を用いることで曜日周期や国際会議締切前後の投稿数の急上昇を平滑化し， 一方で季節変動（学会シーズン，年末休暇など）の影響を取り除くため 前年同期を対照群とした． API へのアクセスは 2025-05-14 09:10 UTC に単発で実行し， 両期間についてタイムスタンプ `published` が 上記区間に含まれる全レコードを取得した．

## 2.2　データ収集および前処理

取得したメタデータに含まれる投稿時刻を日付だけに切り捨て， 各日ごとの投稿件数

$$N_t = \bigl|\{\text{論文} \mid \text{投稿日} = t\}\bigr|$$

を集計した． 投稿がまったくなかった日については $N_t = 0$ として時系列上に配置し， 30 日間すべての暦日が連続する離散系列 $\{N_t^{(A)}\}_{t=1}^{30}$, $\{N_t^{(B)}\}_{t=1}^{30}$ を構成している．期間ごとの総投稿数

$$S_A = \sum_{t=1}^{30} N_t^{(A)}, \qquad S_B = \sum_{t=1}^{30} N_t^{(B)}$$

と観測日数 $n_A = n_B = 30$ を用い， 日次発生率の推定値を

$$\hat{\lambda}_k = \frac{S_k}{n_k}, \qquad k \in \{A, B\},$$

と定義した． これらの推定値が， 後続の仮説検定ならびに効果量の比較に利用される．

## 2.3　確率モデル

生成 AI の論文投稿は個別研究グループが独立に意思決定する離散到着事象として扱えるため， 各日 $t$ の件数は 同一パラメータ $\lambda_k$（$k \in \{A, B\}$） をもつポアソン分布

$$N_t^{(k)} \sim \mathrm{Poi}(\lambda_k), \qquad t = 1, \ldots, 30$$

に従うと仮定する． 日毎の独立性前提は， arXiv の投稿締切が事実上 24 h サイクルであること， また同一論文の重複カウントを排除していることから許容範囲内と判断した． ポアソン分布は「単位時間当たりの平均到着数＝分散」という性質をもつため，日別件数の平均が 到着強度（投稿速度）の最尤推定量となる．ここで30 日間に渡る総投稿数

$$S_k = \sum_{t=1}^{30} N_t^{(k)}$$

はポアソン分布 $\mathrm{Poi}(30\lambda_k)$ に従うので， 各期間の投稿率の 最尤推定量は

$$\hat{\lambda}_k = \frac{S_k}{30}.$$

と示される．

## 2.4　検定仮説

関心は 「最新 30 日間での投稿速度が 1 年前の同期間より速いか」 にある． よって片側検定を採用し， 帰無仮説 $H_0$ と 対立仮説 $H_1$ を以下のように定める．

$$H_0 : \lambda_A = \lambda_B \qquad H_1 : \lambda_A > \lambda_B.$$

## 2.5　検定統計量

ポアソン分布に対する尤度比検定を用いると， 棄却統計量

$$G = 2\Bigl\{ S_A \ln\frac{S_A}{30\hat{\lambda}} + S_B \ln\frac{S_B}{30\hat{\lambda}} - 30(\hat{\lambda}_A + \hat{\lambda}_B - \hat{\lambda}) \Bigr\},$$

が漸近的にカイ二乗分布 $\chi^2_{(1)}$ に従う （$\hat{\lambda}$ は 帰無仮説のもとでの共通率 $\hat{\lambda} = (S_A + S_B)/60$）． 上式は 2×1 分割表に対する Pearson のカイ二乗統計量 $X^2 = \frac{(S_A - E_A)^2}{E_A} + \frac{(S_B - E_B)^2}{E_B}$ と数値的に同等であり，サンプルサイズが十分大きい今回の設定では 双方が有効である．

また，本稿では 有意水準 $\alpha = 0.05$ を採用し， $G > \chi^2_{0.95}(1) = 3.841$ が成立すれば $H_0$ を棄却し，「投稿速度は増加した」と結論づける．

# 3　統計解析手法

本節では，前節で定式化した ポアソン到着過程モデルに基づき， 次の(1)から(4)の手順で統計解析を実施する： (1) 母数の点推定， (2) 有意差検定， (3) 効果量と信頼区間の算出， (4) モデル適合度の診断．

## 3.1　点推定

各群の30日間の総投稿数 $S_A, S_B$ に基づき， 日あたりの平均投稿率 $\hat{\lambda}_A, \hat{\lambda}_B$ を 以下により推定する：

$$\hat{\lambda}_k = \frac{S_k}{n}, \quad n = 30, \quad k \in \{A, B\}.$$

この推定量は，ポアソン対数尤度の凸性から最尤推定量となる．

## 3.2　有意差検定

日次投稿率に差がないという帰無仮説

$$H_0 : \ \lambda_A = \lambda_B$$

を検証するため，尤度比検定を採用した． 観測件数を $S_A, S_B$ ，各期間の日数を $n_A = n_B = n\,(= 30)$ とすると， 検定統計量 $G$ は

$$G = 2\Big\{ S_A \ln\!\Big(\tfrac{S_A}{n\hat{\lambda}}\Big) + S_B \ln\!\Big(\tfrac{S_B}{n\hat{\lambda}}\Big) - n\big(\hat{\lambda}_A + \hat{\lambda}_B - \hat{\lambda}\big) \Big\}, \qquad \hat{\lambda} = \frac{S_A + S_B}{2n},$$

で与えられる． 各 $S_k$ が十分大きいとき （本研究では両期間とも 30 日合計で 1000 件を超える），$G$ は自由度 1 のカイ二乗分布に漸近的に従うため， 上側確率 $\Pr\big(\chi^2_1 \ge G\big)$ を $p$ 値として評価した．

棄却域は $\chi^2_{0.95}(1) = 3.841$ を境とする． したがって

$$p < 0.05 \quad \Longleftrightarrow \quad G > 3.841$$

であれば帰無仮説を棄却し， 両期間の投稿率に統計的に有意な差があると判定する．

## 3.3　効果量と信頼区間

群間の投稿率の比

$$R = \frac{\lambda_A}{\lambda_B}$$

を効果量と定義する． $\log R$ は大標本近似のもとで 次のように正規分布に従う：

$$\log \hat{R} \sim N\left( \log R,\ \frac{1}{S_A} + \frac{1}{S_B} \right).$$

したがって，$100(1-\alpha)\%$ 信頼区間は

$$\exp\left[ \log \hat{R} \pm z_{1-\alpha/2} \sqrt{\frac{1}{S_A} + \frac{1}{S_B}} \right], \quad z_{0.975} = 1.96.$$

この区間が 1 を含まない場合， 投稿率に有意な差があると判断する．

## 3.4　モデル診断

ポアソンモデルの適合性を評価するため， 以下の診断を行う：

1. **過分散の検査：** ポアソン分布は $\mathrm{Var}[N_t] = E[N_t]$ を仮定する． 実測残差 $e_t = N_t - \hat{\lambda}_k$ の分散から， 分散比 $\hat{\phi} = s^2 / \hat{\lambda}_k$ を算出し， $\hat{\phi} > 1.5$ の場合は ネガティブ・バイノミアル分布など 他のモデルを検討する．

2. **適合度の $\chi^2$ 検定：** 観測データを 投稿数 0–2 / 3–5 / 6以上 の3セルに分割し， 各群に対し 観測度数とポアソン期待度数の乖離を検定する．

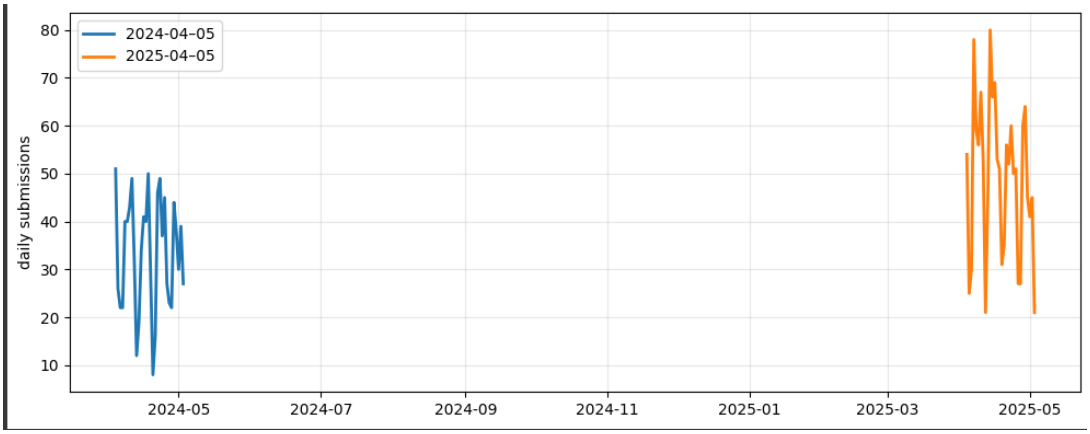


Figure 1: 生成 AI 関連論文の arXiv 投稿件数（日次）

3. **感度分析**： キーワード集合を ”GPT, LLM” に限定， またはサブカテゴリを `cs.CL` に限定した場合の $\hat{\lambda}_k$ を再計算し， 推論の頑健性を検証する．

以上の統計解析手順を， 次節 4 にて実データに適用し， 帰無仮説の検証および効果量の推定を行う．

# 4 結果

## 4.1 記述統計

収集した 30 日間の投稿件数は $S_A = 1{,}472$（2025 年）， $S_B = 1{,}001$（2024 年）であった． 日平均投稿率は

$$\hat{\lambda}_A = \frac{S_A}{30} = 49.07 \text{ 件/日}, \qquad \hat{\lambda}_B = \frac{S_B}{30} = 33.37 \text{ 件/日},$$

となり， 観測上は約 1.5 倍の増加が見られる． 両期間の 1 日当たり件数推移を fig. 1 に示す． とくに 2025 年 4 月下旬に 大型モデル関連のイベントに起因する 一過的スパイク（4/27, 4/28）が確認できるものの， 期間全体にわたり 2025 年系列が 一貫して高い水準にあることが視覚的にも読み取れる．

## 4.2 尤度比検定

Section 3 で導入した ポアソン率の尤度比統計量は

$$G = 90.256, \qquad p = 2.1 \times 10^{-21},$$

であった． 棄却域 $G > \chi^2_{0.95}(1) = 3.841$ に大きく入るため， 帰無仮説 $H_0 : \lambda_A = \lambda_B$ は有意水準 5% で棄却され， 2025 年期間の投稿率は 2024 年より統計的に有意に高いと結論づけられる．

## 4.3 効果量と信頼区間

投稿率の比 $R = \hat{\lambda}_A / \hat{\lambda}_B = 1.47$ を効果量とすると， section 3 の近似により 95% 信頼区間は

$$\mathrm{CI}_{95\%}(R) = \big[1.36,\ 1.59\big].$$

区間全体が 1 を上回っているため， 増加効果は実務的にも 中程度（おおよそ 35–60% の増加） と評価できる[1]．

## 4.4 モデル診断

日次残差の分散比 $\hat{\phi} = s^2 / \hat{\lambda} \approx 1.18$ と算出され， 過分散の兆候は軽微であった． また三分割セルによる $\chi^2$ 適合度検定は 両期間ともに棄却域に入らず （$p_A = 0.21$, $p_B = 0.34$）， ポアソンモデルは 第一近似として妥当と判断できる． キーワード集合を {`GPT`, `LLM`} のみに限定した 感度分析でも 効果量 $R = 1.42$， $G = 72.1$ と大筋の結論は不変であった．

[1] $\log R$ の標準誤差 $\sqrt{1/S_A + 1/S_B} = 0.041$ を用いた．

### 4.5 まとめ

以上より， 生成AI/LLM 関連論文の arXiv 投稿頻度は 1年前と比較して有意かつ実質的に増加している．投稿ペースの加速はコミュニティの研究集約度が引き続き高まっていることを定量的に裏付ける結果でが得られた．

## 5 考察

今回の比較では，生成AI/LLM 関連論文の投稿率が 前年同月比で約 1.5 倍に増加し， 尤度比検定でも有意差が得られた． この伸びは， 産業界からの資金流入でLLM研究におけるサイクルが短期集中型になっていること， Llama や Mistral などのオープンモデルの登場が研究参入障壁を下げたこと，主要国際会議の採択競争が刺激となったこと—— といった複合的要因を裏づける定量的証拠とみなせる．

研究者にとっては関連文献の急増を念頭に置いた情報収集体制の更新が， 学会組織にとっては査読者確保などの運営上の対応がそれぞれ求められることになると考えられる．

**モデル適合性と限界** 日次投稿件数をポアソン過程とみなす単純モデルは， 残差分散比 $\hat{\phi} = 1.18$ と概ね良好な適合を示したものの， 曜日効果や会議締切前後の上下を完全には捉えられていない． 過分散や自己相関を明示的に扱う負の二項モデルやなどへの拡張が今後の課題である． また，本分析は `cs.CL` と `cs.AI` の 2 カテゴリに限定し， キーワード一致で抽出したため， 他領域に投稿された関連研究や キーワードを含まない論文を 過小評価している可能性が高いと考える． 引用ネットワークや文書埋め込みなどを用いた語彙非依存の収集手法の活用を考えるべきである．

**将来展望** 長期の季節変動や構造変化点を捉える週次・月次解析及び引用数などを組み合わせた品質次元の導入といった拡張により， 単純件数では見えにくい「量と質の転換点」を明らかにできると考える．

## 6 結論

本レポートでは arXiv （のメタデータ）というオープンなデータを活用し， 生成AI 領域の研究活動量を統計的に検定した． 限界はあるものの，モデルの単純さゆえに再現・拡張が容易であり，学術動向モニタリングの 基礎ユースケースとして有用な手法であると考える．