

DOI: 10.53104/xdkxsts.2026.02.01.003

訓練語料特徵對大型語言模型湧現行為的理論探討

李忻程¹

1. 江蘇師範大學，江蘇 徐州 221116

摘要：大型語言模型在自然語言處理任務中表現出超出訓練目標的能力，這些能力通常被稱為湧現行為。本文從理論角度探討訓練語料特徵對湧現行為的影響。通過分析語料的資料量、品質、多樣性和結構，討論這些因素如何在模型學習過程中作用，並可能促成模型在未見任務中展現新能力。本文指出，湧現行為不僅受模型規模和結構影響，也受到語料特徵的共同作用。訓練語料設計的量、質、類型和組織方式可能決定模型能力形成的效率和複雜程度。本研究為理解大型語言模型湧現行為提供了語料視角的理論分析，並提出未來可以從語料特徵角度進一步研究模型能力形成的方向。

關鍵字：大型語言模型；訓練語料；湧現行為；理論分析；資料特徵

A Theoretical Study on the Influence of Training Corpus Features on Emergent Behaviors of Large Language Models

LI Xin-cheng¹

1. Jiangsu Normal University, Xuzhou 221116, P. R. China

Correspondence to: LI Xin-cheng; Email: li184463@163.com

Abstract: Large language models exhibit capabilities beyond their direct training objectives, often referred to as emergent behaviors. This paper presents a theoretical study on how training corpus features influence these emergent behaviors. It analyzes corpus characteristics including size, quality, diversity, and structure, and discusses how these factors interact during model learning to enable new abilities on unseen tasks. The study highlights that emergent behaviors are influenced not only by model scale and architecture but also by the combined effects of corpus features. The design of training corpora in terms of quantity, quality, type, and organization may determine the efficiency and complexity of model capability formation. This work provides a theoretical perspective based on training data and suggests directions for future research on how corpus features shape model abilities.

Key words: large language models; training corpus; emergent behaviors; theoretical analysis; data features

引言

大型語言模型近年來在自然語言處理領域取得了顯著進展。這類模型可以生成連貫文本、

回答問題，並展示某些超出訓練目標的能力。這些能力通常被稱為湧現行為^[1]。湧現行為說明模型可以處理未見過的任務，並展現一定的推理或創造能力。研究者對這種現象很關注，他

通信作者：li184463@163.com

引用格式：李忻程. 訓練語料特徵對大型語言模型湧現行為的理論探討[J]. 現代科學探索, 2(1): 15-19.

們想理解它的產生機制和影響因素。訓練語料在模型能力形成中很重要。模型通過在大量文本上學習統計模式和語言規律來獲得知識。語料的量、品質和多樣性會影響模型理解語言、生成文本和應對複雜任務的能力。語料的特徵決定模型能記住什麼，也會影響它在新任務中展現的能力。本文關注的問題是訓練語料特徵如何影響大型語言模型的湧現行為。我們分析語料量、語料多樣性、語料品質和語料結構，探討它們在理論上可能對模型能力產生的作用。討論這些因素可以為理解湧現行為提供理論視角，也可以為訓練資料設計提供參考。

1 湧現行為概述

湧現行為是大型語言模型研究中的一個重要概念。它指模型在規模達到一定程度後，會表現出一些較小模型不明顯具備的能力。這些能力不是訓練目標中直接規定的結果，也不一定來自某一條具體語料。它們更像是模型在大量文本學習後形成的綜合表現。模型學習了詞語關係、句子結構、知識表達和上下文聯繫。這些學習結果在複雜任務中組合時，就可能出現新的能力。

在大型語言模型中，湧現行為可以表現為多種形式。常見的是推理能力。模型可以複述已有知識，也可能根據問題條件進行簡單判斷，完成多步分析，或者在不同資訊之間建立聯繫。語言生成品質也是一種重要表現。模型在規模較小時，生成內容容易重複、斷裂或偏離問題。模型規模提高後，文本可能變得更連貫，語義更完整。少樣本學習能力也是湧現行為的重要類型。模型只看到少量示例，甚至沒有明確訓練過某類任務，也可能根據提示理解任務要求，並給出合適回答。這些現象說明，湧現行為不是單一能力的簡單增加。它涉及語言理解、知識組織、上下文識別和任務遷移等多個層面。模型不能像人一樣真正理解世界，但它可以從語料中學習大量模式。當這些模式在新問題中重新組合時，模型就可能表現出看似“新出現”的能力。這也使湧現行為常常給人一種突然出現的感覺。較小模型難以完成某些任務，較大模型在相似條件下卻能完成，這種差異使湧現行為具有明顯的理論討論價值。

湧現行為的產生機制仍然存在難點。一個難點是很難判斷某種能力究竟來自模型規模，還是來自訓練語料。模型參數增加會提高表達能力，但如果訓練語料缺少足夠豐富的語言材料，模型也難以形成穩定能力。訓練語料中如果包含大量推理文本、問答文本和解釋性文本，模型可能更容易表現出推理和任務遷移能力。由此可見，湧現行為不能只從模型結構本身解釋，也需要從訓練語料角度分析。另一個難點是湧現行為具有不穩定性。同一個模型在不同任務、不同提示和不同語境中，可能表現出不同水準的能力。模型有時可以給出清楚的推理過程，有時也會出現錯誤判斷。這個現象說明，湧現行為不等於模型已經真正掌握某種能力。它更像是一種受條件影響的表現。訓練語料、提示方式、任務難度和上下文安排都會影響這種表現。由於這種不穩定性，湧現行為很難被簡單定義，也很難被完全預測。湧現行為還存在解釋上的問題。模型訓練主要依靠大量文本和統計學習，但模型最後表現出的能力，可能超出單個語料片段能夠解釋的範圍。研究者可以觀察模型輸出結果，但很難完全說明模型內部如何組織這些資訊。訓練語料中的知識、表達方式和任務形式被模型吸收後，會進入參數空間，並以複雜方式保存。模型回答問題時調用的不是某條固定文本，而是大量模式共同作用的結果。這讓湧現行為的理論解釋變得複雜，也讓訓練語料特徵成為理解這一現象的重要入口。

2 訓練語料特徵的維度

訓練語料是大型語言模型能力形成的基礎。語料的不同特徵可能對模型的學習和湧現行為產生影響。語料特徵可以從幾個方面理解。

資料量指訓練文本的總規模。文本學習詞語關係和句子模式^[2]。資料量大時，模型可以接觸到更多語言現象和表達形式，從而積累豐富的語言知識。資料量不足可能限制模型掌握複雜模式的能力，也可能影響其在新任務中的表現。

資料多樣性指語料在主題、風格和語言上的覆蓋範圍。多樣化語料可以讓模型學習不同場景中的語言使用方法。語料覆蓋的主題和文

本類型越多，模型在未見任務中的適應能力就可能越強。風格和語言形式越豐富，模型生成文本時也可能更自然，並減少固定化和重複化表達。

資料品質指語料是否準確、清楚和一致。雜訊、錯誤標注和不一致內容會干擾模型學習，也可能導致模型理解偏差或生成錯誤資訊。高品質語料可以提供穩定的模式和可靠的語言規律，讓模型更容易形成有效能力。

資料結構包括文本順序、上下文長度和文檔組織方式。合理的結構可以說明模型學習長序列關係和上下文關係。上下文長度會影響模型在生成或推理任務中保持資訊的能力。文檔組織方式會影響模型捕捉跨段落或跨文檔語言規律的能力，也會影響複雜任務中的湧現行為。

3 理論機制分析

資料量對模型能力有顯著影響。訓練語料越大，模型接觸的語言模式和表達形式就越豐富。大量文本讓模型能夠學習更多詞語關係和句子結構，並積累更多潛在的組合方式。通過這樣的學習，模型在面對未見任務時可能展現出新的能力。例如，在推理任務中，大模型能夠利用之前學到的模式完成多步推理，而小模型可能無法做到。同樣，在生成任務中，大規模語料提供的多樣例子讓模型生成文本更自然、流暢，也更符合上下文要求。如果語料量不足，模型掌握的模式有限，學習能力受限，難以形成複雜的知識關聯，也難以在新場景中表現出湧現行為^[3]。

資料多樣性影響模型的泛化能力和適應能力。語料中包含不同主題、風格和語言形式時，模型能夠理解在不同情境下如何使用語言^[4]。豐富的主題可以讓模型處理各種類型的問題，例如科學、文學或日常對話等。風格多樣性讓模型在生成文本時能夠呈現不同表達方式，不會一味重複同一種語氣或結構。語言形式的多樣性，比如口語、書面語或不同語言變體，有助於模型在跨領域任務中更靈活地應對。缺乏多樣性的語料容易讓模型生成內容單一，限制新能力的發揮，也可能降低模型在新任務上的表現水準。

資料品質決定模型學習模式的準確性和穩定性。高品質語料提供一致、可靠的語言規律，使模型能夠正確捕捉詞語、句子和上下文之間的關係。低品質語料中存在雜訊、標注錯誤或表達不一致的內容，可能讓模型學習到偏差模式，影響推理能力和創造力^[5]。例如，如果訓練文本中大量出現不合理邏輯或混亂敘述，模型在推理或生成複雜文本時可能出錯，降低其湧現行為的可能性。高品質語料還可以說明模型形成更穩定的知識表示，提高在少樣本任務中的適應能力。

資料結構對模型理解上下文和處理長序列資訊具有重要作用。訓練文本的順序、段落組織和上下文長度影響模型在生成或推理時的表現。合理組織的文檔和較長的上下文可以說明模型捕捉跨句或跨段落的語言關係，從而支援複雜任務。長序列資訊讓模型能夠在生成回答時保持邏輯一致性，避免斷裂或重複。文檔結構的設計也可能影響模型對任務提示的敏感性和理解深度。例如，當模型在結構化的問答文本或解釋性文本上訓練時，它可能更容易形成多步推理能力和任務遷移能力，這些都是湧現行為的基礎條件。

資料量、資料多樣性、資料品質和資料結構之間存在相互作用。大量、豐富、可靠且結構合理的語料可以為模型提供充分的學習條件，使其在複雜任務中表現出新的能力。語料特徵的優化可以提高模型對新任務的適應性，並增加湧現行為出現的可能性。理論上，這些因素共同作用，通過提供多樣化的語言模式和資訊組合方式，使模型形成超越單條語料片段的綜合能力。

4 訓練語料特徵與湧現行為的交互

訓練語料的不同特徵可能相互作用，共同影響模型的湧現行為。資料量大可以讓模型學習更多模式，但如果語料品質低，這些模式可能包含錯誤或不一致資訊。高品質語料可以在一定程度上彌補資料量不足，但當量太少時，模型掌握的模式仍然有限。資料多樣性提供不同情境和表達形式，有助於模型形成靈活的語言使用能力。多樣性不足可能限制模型在新任務中的表現，即使語料量很大或品質很高，也

難以產生複雜湧現能力。資料結構決定模型處理上下文的方式，良好的結構能夠讓模型在長序列和跨段落任務中保持資訊連續。結構不合理可能降低模型利用大量、多樣、優質語料的效果。

理論上，湧現行為可能出現在特定語料條件組合下。例如，大量、高品質、多樣且結構合理的語料有助於模型形成複雜的模式組合。這種模式組合使模型能夠在未見任務中產生新的能力。缺少任一特徵都可能降低湧現行為出現的可能性。語料特徵之間的相互作用決定了模型學習的效率和知識的組織方式。

模型架構對語料特徵的敏感性也會影響湧現行為。相同的語料在不同規模或結構的模型中可能產生不同表現。較大模型能夠更好地捕捉長序列和複雜模式，因此在豐富語料上更容易表現出湧現能力。較小模型可能無法充分利用語料中的資訊，即使語料品質高或多樣性豐富，也難以產生新的能力。模型架構和語料特徵共同作用，決定了湧現行為的強度和範圍。

5 理論啟示與研究展望

訓練語料設計不能只看規模。資料量很重要，但語料的品質、多樣性和結構也會影響模型能力。大量低品質文本可能讓模型學到混亂的表達和錯誤的知識。內容過於單一的語料也可能限制模型在新任務中的適應能力。訓練語料設計需要關注文本來源、主題覆蓋、語言風格和內容可靠性。把這些因素一起考慮，模型才更可能獲得穩定的語言能力和處理複雜任務的能力。

訓練語料的多樣性是湧現行為的重要條件。大型語言模型需要從不同文本中學習不同的表達方式和知識組織方式。新聞文本、學術文本、問答文本、說明文本和對話文本都可以為模型提供不同的學習材料。模型接觸的任務形式越多，就越容易理解新的任務提示。模型接觸的知識場景越廣，也越容易在不同問題之間建立聯繫。這種聯繫可能是推理、遷移和複雜生成能力出現的基礎。

語料品質對湧現行為的影響也很重要。低品質語料可能不會直接阻止模型形成能力，但

會增加模型輸出錯誤的風險。錯誤資訊、重複內容和邏輯混亂的文本會干擾模型學習。模型可能記住錯誤模式，也可能在生成內容時放大這些問題。高品質語料可以為模型提供更穩定的語言樣本，使模型更容易形成清晰的表達方式和可靠的知識聯繫。

語料結構提供理解湧現行為的另一個角度。模型不僅學習詞語和句子，也學習文本內部的資訊關係。段落之間如何銜接，問題和答案如何對應，解釋如何展開，都影響模型處理複雜任務的能力。結構清楚的語料可以說明模型學習問題分析、資訊組織和答案生成的方式。長文本語料還可以說明模型處理跨段落資訊，使模型在長上下文任務中保持連貫性。

從理論上看，湧現行為不應只被理解為模型規模擴大的結果。模型規模提供了能力空間，訓練語料則提供了學習內容。沒有足夠合適的語料，模型規模很難轉化為穩定能力。訓練語料中的知識、表達形式和任務結構會進入模型參數，並在新任務中被重新組合。湧現行為可以被理解為這些模式組合後的表現，而不是某一種單獨因素造成的結果。

這一視角可以改變對湧現行為的單一解釋。過去討論湧現行為時，研究者常把注意力放在參數規模、計算資源和模型結構上。訓練語料視角提醒我們，模型能力的形成也受到資料環境的影響。不同語料可能讓同一類模型形成不同能力。相同規模的模型，如果訓練語料不同，最後表現也可能不同。語料特徵可以解釋部分模型能力差異，也可以幫助理解某些能力為什麼只在特定條件下出現。

後續理論研究可以繼續討論訓練語料和模型能力之間的關係。一個重要方向是分析不同語料類型對湧現行為的影響。比如，推理類文本是否更容易促進推理能力，解釋類文本是否更有助於複雜回答生成，對話類文本是否更能提高任務理解能力。這些問題還需要實證研究進一步驗證，但在理論上已經有討論價值。另一個方向是討論語料特徵之間的組合關係。資料量、品質、多樣性和結構不是孤立因素。它們會一起影響模型學習過程。較大的語料規模如果缺少品質控制，可能不會帶來更強能力。高品質語料如果類型過窄，也可能限制模型的遷

移表現。結構清楚的語料如果數量太少，模型也難以形成穩定模式。研究這些因素的組合關係，可以更準確地解釋湧現行為出現的條件。

訓練語料也涉及價值和安全問題。模型從語料中學習語言，也會學習語料中的偏見、錯誤和不平衡表達^[6]。如果訓練語料中有大量偏見內容，模型可能在回答中重複類似傾向。如果語料中缺少某些群體、地區或語言的表達，模型在相關任務中的表現也可能較弱。湧現行為不只代表能力提升，也可能帶來難以預料的錯

誤和風險。訓練語料設計需要考慮能力形成，也需要考慮模型輸出是否可靠，以及它可能產生的社會影響。未來對大型語言模型湧現行為的理解，可以更多回到訓練語料本身。語料不是模型訓練前的簡單材料，而是模型能力形成的基本條件。語料的選擇、清洗、組織和分佈都會影響模型最終表現。把訓練語料作為理論分析對象，可以使湧現行為研究更加具體，也能避免只用模型規模解釋所有問題。

參考文獻：

- [1] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J/OL]. arXiv preprint, 2022: arXiv:2206.07682. [2026-06-09]. <https://arxiv.org/abs/2206.07682>
- [2] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J/OL]. arXiv preprint, 2020: arXiv:2001.08361[2026-06-09]. <https://arxiv.org/abs/2001.08361>
- [3] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[C]. Advances in Neural Information Processing Systems. 2022, 35: 30016-30030.
- [4] Gao L, Biderman S, Black S, et al. The Pile: an 800GB dataset of diverse text for language modeling[J/OL]. arXiv preprint, 2020: arXiv:2101.00027[2026-06-09]. <https://arxiv.org/abs/2101.00027>
- [5] Dodge J, Sap M, Marasović A, et al. Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 1286-1305.
- [6] Bender E M, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?[C]. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2021: 610-623.

版權聲明

© 2026 作者版權所有。本文依據“知識共用署名 4.0 國際授權合約”（CC BY 4.0）以開放獲取方式發佈。該許可允許使用者在任何媒介中自由使用、複製、傳播與改編文章（含商業用途），惟須明確署名原作者及出處，並注明所作修改（如有）。完整協議詳見：<https://creativecommons.org/licenses/by/4.0/deed.zh-hans>

出版聲明

所有出版物中的陳述、觀點及資料僅代表作者及供稿者個人立場，與 Brilliance Publishing Limited 及/或編輯人員無關。Brilliance Publishing Limited 及/或編輯人員對因內容所提及的任何理念、方法、說明或產品所導致的人身或財產損害概不負責。