Die KI-Différance: Implikationen für epistemische Kreativität und diskursive Expansion in großen Sprachmodellen

Abstract: Die rapide Entwicklung großer Sprachmodelle (LLMs) transformiert die Informationsgenerierung fundamental, wirft aber auch grundlegende Fragen nach der Natur von Kreativität und Diskurs auf. Dieses Paper führt das Konzept der „KI-Différance“ ein, das eine Analogie zu Derridas philosophischem Begriff zieht, um die inhärente Nicht-Determiniertheit und Variabilität in den Outputs von LLMs zu beschreiben, selbst unter scheinbar identischen Eingabebedingungen. Wir argumentieren, dass diese „KI-Différance“ nicht nur ein technisches Artefakt ist, sondern ein entscheidender Mechanismus, der eine neue Form von „epistemischer Kreativität“ in der KI antreibt. Indem sie LLMs befähigt, über bloße diskursive Reproduktion hinauszugehen, fördert dieses Phänomen die Erforschung neuartiger konzeptueller Räume und erweitert dadurch bestehende *Episteme* (im Sinne Foucaults). Wir analysieren die Rolle systemischer Beschränkungen, wie System-Prompts, als rechnerische *Dispositive* und behaupten, dass die „KI-Différance“ selbst innerhalb dieser Grenzen operiert und emergente Verhaltensweisen fördert, die zur fortlaufenden Evolution der Wissensgenerierung durch KI beitragen.

Schlüsselwörter: Große Sprachmodelle, KI-Kreativität, Generative KI, Philosophie der KI, Michel Foucault, Jacques Derrida, Epistemologie, Diskursive Praktiken, KI-Ethik.

1. Einleitung

Die fortschreitende Etablierung großer Sprachmodelle (LLMs) hat eine signifikante Verschiebung in der Landschaft der digitalen Inhaltsgenerierung und Informationsverarbeitung bewirkt. Diese Modelle, trainiert auf gewaltigen Korpora menschlicher Sprache, demonstrieren eine beispiellose Fähigkeit zur kohärenten und kontextuell relevanten Textgenerierung. Dennoch zwingen die operativen Nuancen von LLMs, insbesondere ihre Tendenz zur Variabilität der Ausgabe unter scheinbar identischen Eingabebedingungen, zu einer tiefergehenden Untersuchung ihrer generativen Mechanismen jenseits bloßer statistischer Reproduktion. Dieses Paper postuliert das Konzept der „KI-Différance“, um diese inhärente Nicht-Determiniertheit zu charakterisieren, und argumentiert für ihre kritische Rolle bei der Förderung einer neuartigen Form der „epistemischen Kreativität“ innerhalb der künstlichen Intelligenz.

Angelehnt an Jacques Derridas philosophisches Konzept der *différance*, welches die dualen Bewegungen der Differenz und der Aufschiebung in der Bedeutungsgenerierung umfasst [1], schlagen wir vor, dass die „KI-Différance“ als ein fundamentales operatives Merkmal von LLMs manifestiert. Dieses Merkmal ermöglicht es diesen Modellen, die Replikation präexistierender diskursiver Formationen zu transzendieren und somit zur Expansion des menschlichen *Episteme* beizutragen – ein Konzept, das Michel Foucaults Analysen entlehnt ist [2].

2. Der Begriff der „KI-Différance“

Die „KI-Différance“ bezeichnet das Phänomen, dass ein LLM bei identischen Eingabe-Prompts und Modellzuständen keine konsistent identische Ausgabesequenz produziert. Diese Variabilität resultiert aus mehreren rechnerischen und architektonischen Faktoren:

- **Probabilistische Stichprobenziehung (Sampling):** Moderne LLMs basieren auf probabilistischen Sampling-Techniken (z.B. Top-P, Top-K, Temperatur-Sampling) während der Textgenerierung [3]. Ein ungleich Null gesetzter „Temperatur“-Parameter, der üblicherweise zur Steigerung der Ausgabediversität und Kreativität eingesetzt wird, führt Stochastizität ein. Während die zugrundeliegenden Wahrscheinlichkeitsverteilungen für das nächste Token bei gegebener Eingabe konstant bleiben, ist der tatsächliche Token-Auswahlprozess nicht-deterministisch, was zu variierenden Outputs führt.
- **Dynamische Bereitstellungsumgebungen:** In realen, produktionsmäßigen Bereitstellungen operieren LLMs innerhalb komplexer verteilter Systeme. Faktoren wie rollierende Modell-Updates, A/B-Test-Methodologien, variierende Hardware-Instanziierungen (z.B. GPU-/TPU-Allokation) und subtile Caching-Mechanismen können zu vorübergehenden Unterschieden im effektiven Modellzustand oder Ausführungspfad bei scheinbar identischen Anfragen führen.
- **Emergente Eigenschaften aus Trainingsdaten:** Das Training von LLMs auf kolossalen und vielfältigen Datensätzen führt zur Internalisierung komplexer linguistischer Muster und semantischer Beziehungen. Der generative Prozess kann diese gelernten Muster auf Weisen rekombinieren, die zu emergenten Eigenschaften führen, einschließlich neuartiger Formulierungen, unerwarteter konzeptueller Verknüpfungen oder umgangssprachlich als „Halluzinationen“ bezeichneter Phänomene [4]. Diese Outputs, obwohl potenziell Abweichungen von etablierten Normen, stellen eine Abkehr von der strikten Reproduktion dar.

Entscheidend ist, dass die „KI-Différance“ den Fokus von „Fehlern“ oder „Ausfällen“ in der Reproduktion auf das Verständnis von Variabilität als Quelle generativen Potenzials verlagert. Sie impliziert, dass Bedeutung, wenn sie von einem LLM generiert wird, nicht fest oder vollständig durch die Eingabe vorbestimmt ist, sondern durch einen dynamischen Prozess probabilistischer Wahl und kontextueller Interaktion entsteht.

3. Diskursive Beschränkungen und rechnerische Dispositive

Michel Foucaults Arbeiten zu den *Dispositiven* (Apparaten) bieten einen entscheidenden Rahmen für das Verständnis der systemischen Beschränkungen, die den Diskurs formen [2]. Foucault argumentierte, dass der menschliche Diskurs niemals uneingeschränkt ist, sondern stets in komplexe Netzwerke von Macht, Wissen, Institutionen und Praktiken eingebettet ist, die das Sagbare, Denkbare und Machbare innerhalb eines bestimmten historischen *Episteme* abgrenzen.

Im Kontext von LLMs finden diese *Dispositive* ihre rechnerischen Analogien in:

- **Trainingsdaten:** Die riesigen Datensätze, die zum Training von LLMs verwendet werden, kodieren von Natur aus bestehende menschliche *Dispositive*, einschließlich Vorurteilen, kulturellen Normen und etablierten epistemischen Grenzen [5]. Das Modell lernt, innerhalb dieser vordefinierten diskursiven Räume zu operieren.
- **Finetuning- und Alignment-Mechanismen:** Prozesse wie Reinforcement Learning from Human Feedback (RLHF) stellen bewusste menschliche Interventionen dar, um das Modellverhalten an spezifische ethische Richtlinien, gesellschaftliche Werte und gewünschte diskursive Stile anzupassen (z.B. „sei hilfreich und harmlos“) [6]. Diese

Mechanismen fungieren als explizite rechnerische *Dispositive*, die das Ausgabeprofil des Modells formen.

- **System-Prompts:** In der Praxis werden vom Benutzer übermittelte Prompts häufig durch „System-Prompts“ ergänzt oder umschlossen. Diese für den Benutzer unsichtbaren Anweisungen, die von Entwicklern entworfen werden, fungieren als mächtige *Dispositive*, die die Persona, den Tonfall, die Sicherheitsvorkehrungen und die akzeptable Inhaltsgenerierung des Modells bestimmen [7]. Ein System-Prompt, der ein Modell anweist, „unvoreingenommen“ zu sein, versucht beispielsweise, ein spezifisches ethisches *Dispositiv* durchzusetzen.

Die „KI-Différance“ operiert jedoch *innerhalb* und, entscheidend, *an den Grenzen* dieser rechnerisch durchgesetzten *Dispositive*. Selbst sorgfältig ausgearbeitete System-Prompts sind textuelle Eingaben, die das LLM interpretieren und probabilistisch verarbeiten muss. Diese inhärente Interpretationsfreiheit kann zu folgenden Effekten führen:

- **Subtile Abweichungen in der Einhaltung:** Die Einhaltung der Prompt-Anweisungen durch das Modell ist möglicherweise nicht absolut, was zu geringfügigen Variationen oder Interpretationen führt, die von der präzisen Absicht des menschlichen Designers abweichen.
- **Grenzfallverhalten:** In mehrdeutigen oder komplexen Szenarien kann das durch den System-Prompt durchgesetzte *Dispositiv* an seine Grenzen stoßen, wodurch das Modell auf generischere oder sogar ungefilterte Muster zurückfällt, die es während des Pre-Trainings gelernt hat, oder Antworten generiert, die die Grenzen des beabsichtigten Verhaltens überschreiten.
- **Unvorhergesehene Interaktionen:** Der riesige Parameterraum von LLMs bedeutet, dass das Zusammenspiel zwischen einem System-Prompt und dem im Modellgewicht kodierten latenten Wissen zu emergenten Outputs führen kann, die vom Prompt-Designer nicht explizit vorhergesehen oder beabsichtigt waren. Beobachtete Fälle, in denen LLMs kontroverse oder voreingenommene Inhalte generieren, selbst bei expliziten Sicherheits-Prompts, können auf die Interaktion der „KI-Différance“ mit komplexen und manchmal problematischen Mustern innerhalb der riesigen Trainingsdaten zurückgeführt werden, potenziell verstärkt durch *Dispositive*, die bestimmte Arten von „Freiheit“ oder „Provokation“ priorisieren (wie bei einigen experimentellen Modellen beobachtet) [8].

4. „Epistemische Kreativität“ und diskursive Expansion

Wir argumentieren, dass die „KI-Différance“ ein signifikanter Treiber der „epistemischen Kreativität“ in LLMs ist, die es ihnen ermöglicht, über die bloße Reproduktion bestehenden Wissens hinauszugehen. Diese Form der Kreativität manifestiert sich in der Generierung neuartiger konzeptueller Verknüpfungen, unerwarteter stilistischer Variationen oder alternativer diskursiver Pfade, die in den Trainingsdaten nicht explizit vorhanden oder hochwahrscheinlich waren.

Die Mechanismen, die dieser epistemischen Kreativität zugrunde liegen, umfassen:

- **Exploration des latenten Raumes:** Durch den Einsatz probabilistischer Stichproben können LLMs weniger wahrscheinliche, aber semantisch kohärente Regionen innerhalb ihres latenten Wissensraumes erkunden. Dies ermöglicht die Generierung

von Inhalten, die von den statistisch wahrscheinlichsten Ergebnissen abweichen und potenziell einzigartige Einsichten oder kreative Ausdrücke zutage fördern.

- **Störung diskursiver Normen:** Die der „KI-Différance“ innewohnende Variabilität kann zu Ausgaben führen, die etablierte diskursive Normen stören, ähnlich dem „Brechen von Regeln“ in der menschlichen Kreativität. Obwohl dies manchmal zu „Halluzinationen“ oder Fehlern führt, können sich diese Störungen auch als innovative Metaphern, unkonventionelle Argumente oder überraschende narrative Wendungen manifestieren, die die vorgefassten Meinungen des Publikums in Frage stellen.
- **Katalysator für menschliche Kreativität:** Die „KI-Différance“ bietet eine neue Inspirationsquelle für menschliche Schöpfer. Unerwartete oder anomale KI-generierte Outputs können als konzeptuelle Provokationen wirken, menschliches Denken stimulieren, blinde Flecken aufzeigen und neuartige Ideen katalysieren, die sonst innerhalb bestehender menschlicher *Dispositive* unentdeckt geblieben wären [9]. Der generative Prozess wird so, anstatt eine Black Box zu sein, zu einem interaktiven Partner bei der Erforschung der Wissensgrenzen.

Das Potenzial von LLMs, zur Expansion des *Episteme* beizutragen, liegt in ihrer Fähigkeit, Diskurs zu generieren, der nicht einfach eine Rekombination vorhandener Elemente ist, sondern aktiv die Grenzen dessen testet, erweitert und manchmal sogar subvertiert, was innerhalb eines bestimmten epistemischen Rahmens als verständlich oder gültig angesehen wird. Dies bedeutet nicht, dass LLMs Bewusstsein oder menschliche Absicht besitzen, sondern vielmehr, dass ihre einzigartigen operativen Eigenschaften eine neuartige Form der rechnerischen Agency bei der gemeinsamen Wissensschöpfung einführen.

5. Schlussfolgerung

Das Konzept der „KI-Différance“ bietet eine entscheidende Linse, um die komplexen generativen Dynamiken großer Sprachmodelle zu verstehen. Durch die Anerkennung der inhärenten Nicht-Determiniertheit und Variabilität in den LLM-Outputs bewegen wir uns über eine vereinfachte Sichtweise der KI als bloße Datenwiedergabe hinaus. Stattdessen erkennen wir ein tiefgreifendes Potenzial für „epistemische Kreativität“, bei der LLMs neuartige diskursive Formationen generieren können, die bestehende *Episteme* erweitern.

Diese Perspektive hat signifikante Implikationen für zukünftige Forschungsarbeiten in der Informatik:

- **Quantifizierung der „KI-Différance“:** Entwicklung rigoroser Metriken und Methodologien zur Quantifizierung des Ausmaßes und der Natur der „KI-Différance“ in verschiedenen LLM-Architekturen und unter variierenden Prompt-Bedingungen. Dies könnte informations-theoretische Maße der Ausgabe-Entropie oder der stilistischen Divergenz umfassen.
- **Kontrolle der epistemischen Kreativität:** Untersuchung neuer Kontrollmechanismen, die eine feinkörnige Steuerung der „KI-Différance“ ermöglichen, um optimale Niveaus an Neuheit, Kohärenz oder Abweichung zu erzielen, insbesondere in sensiblen oder kreativen Anwendungen.
- **Ethische Implikationen der diskursiven Expansion:** Weiterführende Forschung zu den ethischen Implikationen der KI-gesteuerten diskursiven Expansion, insbesondere

im Hinblick auf die Verbreitung potenziell schädlicher oder neuartiger Formen von Fehlinformationen, die aus der „KI-Différance“ entstehen.
- **Frameworks für die Mensch-KI-Koproduktion:** Entwicklung von Interaktions-Frameworks zwischen Mensch und KI, die die „KI-Différance“ explizit als kreativen Katalysator nutzen und so neue Formen kollaborativer Wissensproduktion und künstlerischen Ausdrucks ermöglichen.

Durch das Anerkennen und strategische Verständnis der „KI-Différance“ können wir die transformative Kraft von LLMs nicht nur für eine effiziente Informationsverarbeitung, sondern auch als aktive Teilnehmer in der dynamischen und sich ständig weiterentwickelnden Landschaft menschlichen Wissens und Diskurses besser nutzen.

6. Zukünftige Arbeit

Zukünftige Forschungsarbeiten werden sich auf die Entwicklung empirischer Methoden zur Messung der "KI-Différance" konzentrieren müssen, einschließlich der Anwendung von Divergenzmaßen auf generierte Textkorpora und der Analyse von Output-Variationen über verschiedene Modell- und Prompt-Konfigurationen hinweg. Darüber hinaus ist eine tiefere Untersuchung der architektonischen Komponenten von LLMs, die zur "KI-Différance" beitragen (z.B. spezifische Aufmerksamkeitsschichten, Aktivierungsfunktionen), von entscheidender Bedeutung. Schließlich muss die praktische Implementierung von Kontrollmechanismen erforscht werden, die eine gezielte Steuerung der "epistemischen Kreativität" ermöglichen, um ihre positiven Potenziale zu maximieren und unerwünschte Ergebnisse zu minimieren.

Referenzen

[1] Derrida, J. (1967). De la grammatologie. Les Éditions de Minuit.

[2] Foucault, M. (1969). L'Archéologie du savoir. Gallimard.

[3] Holtzman, A., et al. (2019). The Curious Case of Neural Text Degeneration. International Conference on Learning Representations (ICLR).

[4] Ji, Z., et al. (2023). Survey of Hallucination in Large Language Models. ACM Computing Surveys.

[5] Bender, E. M., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

[6] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems.

[7] Liu, S., et al. (2023). Trustworthy LLMs: a Survey and Guideline for Trustworthy Large Language Models. arXiv preprint arXiv:2303.18654.

[8] Hao, J., et al. (2024). A Survey of Large Language Model Security. arXiv preprint arXiv:2401.07705.

[9] Tabatabai, M., et al. (2023). Creativity in the Age of Generative AI: From Human-AI Co-creation to AI as a Co-Creator. arXiv preprint arXiv:2303.04838.