

Analyse von Chain of Thought in LLMs-as-Evaluators

Evaluation von NLP-Systemen WS24/25

Long Kim

Mit der Verbesserung von LLM-Systemen in den letzten Jahren wird verstärkt versucht, LLMs auch als Alternative zu menschlichen Evaluationen von Texten zu nutzen. Beispiele dafür sind z.B. [G-Eval](#)¹ und [GPTScore](#)². G-Eval nutzt hierbei ein Prompt, welches den Task und die jeweilige Metric erklärt, und bildet daraus ein Auto-/Self-CoT (von einem Modell erstelltes CoT), um einen detaillierteren Arbeitsplan übergeben zu können. [Moran Mizrahi et al., 2024](#)³ haben bereits verschiedene Prompt-Paraphrasierungen und Prompting-Methoden getestet (u.a. CoT als eines ihrer Prompting-Methoden), allerdings wurden verschiedene CoT nicht gegeneinander getestet. In diesem Experiment sollen von Modellen selbst erstellte und anderweitig geänderte CoT genauer analysiert werden.

1. Einleitung und Methode

Um einen Vergleich zwischen CoT bilden zu können, wird der [Code von G-Eval](#)⁴ als Basis genutzt. Dieser Code beinhaltet das SummEval-Dataset, detaillierten Prompts (mit Auto-CoT) für dieses Dataset, und den Code für die GPT-4- und Meta-Evaluation. Die Analyse von CoT soll zwischen Auto-CoT von verschiedenen Modellen, einem stark verkürzten CoT und einem CoT, bei dem die Reihenfolge der Schritte verändert wurden. Hierbei sollen nicht nur generelle Unterschiede in Ergebnissen aufgezeigt werden, sondern auch potentielle Biases (z.B. ob das Nutzen eines bestimmten Auto-CoT Auswirkungen auf die Evaluation desselben Modells hat).

Aufgrund von API-Limitationen werden die 3 Modelle "Gemini-2.0-flash", "qwen-2.5-32b" (via Groq) und "Meta-Llama-3-70B-Instruct" (via Hyperbolic) auf ein kleines Sample von 100 Instanzen des SummEval-Datasets über 3 Metriken getestet. Dadurch entstehen 9 Auto-CoTs, die mithilfe der G-Eval-Prompts kreiert wurden (mit einer Temperatur von 0 um Reproducibility zu gewährleisten). Fluency als Metrik wurde nicht berücksichtigt, da die G-Eval-Autoren kein Auto-CoT mit Evaluation Steps für diese Metrik erstellt haben. Weiterhin wurden die Llama-Auto-CoTs als Grundlage für das verkürzte und scrambled CoT genutzt. Ein scrambled CoT soll testen, ob LLMs eine logische Reihenfolge benötigen, um eine genaue Evaluation vollführen zu können, oder ob das reine Nennen von allen notwendigen Schritten (in falscher Reihenfolge) dafür ausreichen würde. Das verkürzte CoT hingegen soll testen, ob LLMs wirklich einen anthropomorphischen Ansatz eines CoT benötigen (sehr detailliert bzw. geschrieben, als ob Adressant ein Mensch ist) oder ob lediglich ein grober Arbeitsplan genügt.

¹ <https://arxiv.org/abs/2303.16634>

² <https://arxiv.org/abs/2302.04166>

³ https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00681/123885

⁴ <https://github.com/nlpyang/geval>

2. Analyse der CoTs

Llama und Qwen geben CoTs, die strukturell den G-Eval-CoTs ähnlich sind. Hierbei ist Qwen allerdings inhaltlich detaillierter (Llama Coherence Prompt: “4. Decide if the summary builds from sentence to sentence to form a coherent body of information about a topic.” vs. Qwen Coherence Prompt: “4. Identify if there are any abrupt shifts in the information that disrupt the flow or if the sentences are well-connected, contributing to a cohesive narrative.”) Gemini hingegen hat kleine strukturelle Unterschiede (nutzt Formatierungen und gibt jedem Schritt einen Titel) und zeigt die detailliertesten Evaluation Steps mit Alternativen und Spezifizierungen (Coherence Prompt: “4. (...) (e.g., chronological, cause-and-effect, problem-solution) (...)” oder Consistency Prompt: “4. (...) “Not Supported by the Source: The source article does not mention or imply the claim. Contradicted by the Source: The source article states something that contradicts the claim. Hallucinated: The claim is completely fabricated and has no basis in the source article.”).

Die Schritte der scrambled CoTs wurden alle im selben Format vertauscht (e.g. die neue Reihenfolge ist 3 5 2 4 1), man hätte dies aber auch zufällig machen können. Die Nummerierung der Schritte wird entfernt, damit LLMs nicht selbst die richtige Reihenfolge erfassen können, allerdings wurden die Zeilenumbrüche beibehalten, um Schritte voneinander klar abzutrennen. Die verkürzten CoTs versuchen, die wichtigsten Informationen (keywords) jedes Schrittes zu erfassen. Weiterhin werden alle Schritte kleingeschrieben und Satzzeichen gelöscht, um den Anthropomorphismus zu entfernen. Dies kann potentiell eine Vermischung von Bias erzeugen, weswegen in diesem Experiment die verkürzten CoTs demnach eher als *stark vereinfachte CoTs* gesehen werden können.

3. Ergebnisse (Appendix A für Correlation Tables)

3.1 Was sind die generellen Trends?

Für Coherence befinden sich die meisten Correlation Scores aller 3 Modelle im Bereich von 0.6 - 0.7. Für Gemini befindet sich der geringste Wert beim Llama-CoT, und der höchste beim scrambled CoT. Für Llama ist der geringste beim Gemini-CoT und der höchste beim Llama-CoT. Bei Qwen zeigt sich ein massiver Absturz beim Gemini-CoT, wo die Correlation Scores alle bei ~0.42 liegen. Währenddessen zeigt sich beim Llama-CoT die höchste Coherence Correlation aller Modelle bei mehr als 0.8.

Bei Consistency finden sich die meisten Werte bei über 0.7. Gemini zeigt hier bei jedem CoT-Prompt eine sehr hohe Correlation. Für Llama und Qwen befinden sich die meisten zwischen 0.7 und 0.85, die einzige Ausnahme ist Qwen mit dem scrambled CoT.

Die höchsten Werte in Relevance erzielt Qwen mit Werten über 0.8, mit dem Maximum von ~0.93 mit scrambled CoT. Interessanterweise werden die niedrigsten Werte durch das Llama-CoT erzielt

(im Kontrast zur Coherence Correlation). Gemini und Llama zeigen bei ihrer Relevance Correlation Werte von ~0.5 bis 0.7, der höchste Wert zwischen beiden ist Gemini mit dem scrambled CoT (~0.78).

Tendenziell geben Gemini und Llama im Durchschnitt die geringsten predicted Scores in Coherence, und die höchsten in Relevance. Qwen gibt noch geringere predicted Scores als die anderen beiden Modelle in Coherence (Gemini - 3.4, Llama - 3.0, Qwen - 2.7), gibt aber die höchsten Werte in Consistency. Durch die niedrigeren predicted Scores in Relevance schließt Qwen auch am besten in den Correlation Scores für Relevance ab.

3.2 Wie beeinflussen die CoTs die Modelle?

Überraschenderweise ist das scrambled CoT für Gemini im Durchschnitt das Beste über alle 3 Metriken, obwohl es basiert auf das Llama-CoT ist (welches für Coherence und Relevance schlechte Werte erzielt). Dies kann bedeuten, dass es für Gemini leichter ist, keine bestimmte Reihenfolge an Schritten zu benutzen, sondern eine Masse an Instruktionen zu verarbeiten. In diesem Sinne hilft es vielleicht weniger, eine *Kette von Gedanken* über CoT zu errichten. Die Ergebnisse bei Llama und Qwen sind etwas schwächer, aber dennoch im mittleren bis hohen Bereich. Generell kann es aber auch sein, dass eben dieses scrambled CoT kein wirkliches CoT ist, sondern für die Modelle ein detailliertes Prompt darstellt und es deswegen hohe Werte erzielt.

Verkürzte CoT zeigen generell mittelmäßige Werte. Spezifische Ausnahmen dafür sind Relevance Scores für Gemini und Llama, und Coherence für Qwen. Das heißt, selbst wenn wir verkürzte Prompts als *vereinfachte Prompts* sehen, schneiden sie teilweise besser ab als CoTs, die menschlicher aufgebaut sind. Der Grund dahinter ist aber nicht ganz klar: eine Möglichkeit wäre, dass bestimmte Details in CoT-Prompts eher irreführend sind, und die predicted Scores der Modelle daher von Human Scores stärker abweichen können.

Die 3 Modell-CoTs produzieren unterschiedlich hohe Ergebnisse abhängig von der Metrik. Das Llama-CoT ist zum Beispiel das Schwächste für Relevance für alle 3 Modelle, ist aber im Durchschnitt das Beste für Coherence. Dies kann bedeuten, dass die Art von Instruktionen für bestimmte Metriken besser ist als für andere. Anhand der Prompts kann man feststellen, dass das Llama-CoT für Relevance davon abhängig ist, die wichtigsten Informationen zu extrahieren und einzuschätzen, wie es im Vergleich zu anderen gegebenen Informationen steht, währenddessen für Coherence das CoT wesentlich genereller ist und sich auf den Text als Ganzes bezieht. Für Relevance kann dies bedeuten, dass sich die wichtigsten Informationen von menschlicher Evaluation unterscheiden können und dadurch zu unterschiedlichen Scores führt.

Wie bereits erwähnt schließt das Gemini-CoT für Llama und Qwen in Coherence am schlechtesten ab, und dies kann an den detaillierten Instruktionen liegen, die dann im Durchschnitt zu niedrigeren predicted Scores führen (Gemini - 3.35, Llama - 3, Qwen - 2.77). Durch diese Abhängigkeit zwischen Modell-CoT und Metrik gehe ich davon aus, dass es keinen direkten Bias für ein Modell zu seinem eigenen CoT gibt, auch wenn einige der Werte dafür sprechen könnten (aber eben mit dem scrambled/short CoT bzw. mindestens eines der anderen Modell-CoT vergleichbar sind).

Appendix A - Correlation Tables

A1. Gemini

Coherence

Prompt	Pearson	Spearman	Kendall
Gemini	0.6866	0.6713	0.6578
Llama	0.5714	0.5714	0.5714
Qwen	0.6275	0.6065	0.5855
Scrambled	0.7879	0.7855	0.7749
Short	0.614	0.5931	0.596

Consistency

Prompt	Pearson	Spearman	Kendall
Gemini	0.9922	1.0	1.0
Llama	1.0	1.0	1.0
Qwen	0.9768	0.9833	0.9771
Scrambled	0.969	0.9777	0.9694
Short	0.9794	0.9851	0.9796

Relevance

Prompt	Pearson	Spearman	Kendall
Gemini	0.6381	0.6364	0.6267
Llama	0.5857	0.575	0.5612
Qwen	0.6407	0.6244	0.6148
Scrambled	0.7872	0.773	0.7661
Short	0.5082	0.5004	0.4906

A2. Llama

Coherence

Prompt	Pearson	Spearman	Kendall
Gemini	0.6394	0.626	0.6219
Llama	0.7767	0.7727	0.7656
Qwen	0.6922	0.6791	0.6651
Scrambled	0.7154	0.705	0.6921
Short	0.697	0.6901	0.6745

Consistency

Prompt	Pearson	Spearman	Kendall
Gemini	0.7144	0.7222	0.7222
Llama	0.7849	0.7927	0.7982
Qwen	0.7571	0.7629	0.7574
Scrambled	0.8255	0.8333	0.8333
Short	0.8255	0.8333	0.8333

Relevance

Prompt	Pearson	Spearman	Kendall
Gemini	0.7344	0.7322	0.799
Llama	0.5	0.4839	0.4853
Qwen	0.6286	0.6109	0.6095
Scrambled	0.5328	0.539	0.5402
Short	0.4965	0.4944	0.4976

A3. Qwen

Coherence

Prompt	Pearson	Spearman	Kendall
Gemini	0.4145	0.419	0.4155
Llama	0.8298	0.8081	0.8018
Qwen	0.6534	0.6298	0.6272
Scrambled	0.7428	0.7313	0.7225
Short	0.5334	0.5004	0.49

Consistency

Prompt	Pearson	Spearman	Kendall
Gemini	0.8275	0.8333	0.8333
Llama	0.7466	0.75	0.75
Qwen	0.851	0.8515	0.847
Scrambled	0.6231	0.6273	0.6166
Short	0.8023	0.806	0.8015

Relevance

Prompt	Pearson	Spearman	Kendall
Gemini	0.8011	0.8023	0.7952
Llama	0.6837	0.6911	0.6878
Qwen	0.806	0.794	0.7861
Scrambled	0.9355	0.9296	0.9241
Short	0.8288	0.8115	0.81

B KI-basierte Hilfsmittel

meta_eval_summeval.py:

Nutzung von ChatGPT für regex-Matching

Grund: Eigene Versuche waren für bestimmte Outputs nicht präzise genug