وهم "وهم التفكير"

تعليق على بحث شجاعي وزملاؤه (2025)

سي. أوبوس* **أي. لوسن***

10 جوان **2025**

ملخص

أفاد شجاعي وزملاؤه (2025) أن نماذج الاستدلال الكبيرة (LRMs) تُظهر "انهيارًا دقّيقيًا" في حل ألغاز التخطيط عند تجاوز عتبات تعقيد معينة. نوضح في هذا التعليق أن نتائجهم تعكس في الأساس قيودًا في التصميم التجريبي وليس إخفاقات في الاستدلال الجوهري. يكشف تحليلنا ثلاث إشكاليات حرجة:

(١) تجارب برج هانوي تتجاوز بشكل منهجي حدود الرموز المخرَجة للنماذج عند النقاط المبلغ عنها كإخفاقات، مع اعتراف النماذج صراحةً بهذه القيود في مخرجاتها.(٢) إطار التقييم الآلي للباحثين يفشل في التمييز بين إخفاقات الاستدلال والقيود العملية، ما يؤدي إلى تصنيف خاطئ لقدرات النماذج.

(٣) الأكثر إثارةً للقلق: معايير عبور النهر (River Crossing) تتضمن حالات مستحيلة رياضيًّا عند $N \geq 6$ بسبب سعة قارب غير كافية، ومع ذلك تُسجَّل النماذج كفاشلة لعدم حلّها هذه المسائل المستعصية.

عندما نتحكم في هذه المُحدِّدات التجريبية عن طريق طلب دوال توليدية (Generating Functions) بدلاً من قوائم الحركات الشاملة، تشير تجارب أولية على نماذج متعددة إلى دقة عالية في حالات برج هانوي التي أُبلغ سابقًا عن فشل كامل في حلها. تُبرز هذه النتائج أهمية التصميم التجريبي الدقيق عند تقييم قدرات الذكاء الاصطناعي الاستدلالية.

1 المقدمة

يدَّعي شجاعي وزملاؤه (2025) أنهم حددوا قيودًا جوهرية في نماذج الاستدلال الكبيرة (LRMs) عبر تقييم منهجي لألغاز التخطيط. يمتلك اكتشافهم الرئيسي —أن دقة النماذج "تنهار" إلى الصفر بعد عتبات تعقيد معينة — مضامين مهمة لأبحاث استدلال الذكاء الاصطناعي. غير أن تحليلنا يكشف أن هذه الإخفاقات الظاهرية تنبع من خيارات التصميم التجريبي وليس من قيود جذرية في النماذج.

2 النماذج تدرك قيود المخرجات

ملاحظة حاسمة غُفل عنها في الدراسة الأصلية: النماذج تُدرك بنشاط عندما تقترب من حدود المخرجات. إعادة تنفيذ حديثة بواسطة [@scaling01 على تويتر [٢] رصدت نصوصًا صريحة من مخرجات النموذج تقول: "النمط يستمر، ولكن لتجنب الإطالة، سأتوقف هنا" عند حل ألغاز برج هانوي. يثبت هذا أن النماذج تفهم نمط الحل لكنها تختار اقتطاع المخرجات بسبب قيود عملية.

يعكس هذا التوصيف الخاطئ لسلوك النموذج على أنه "انهيار استدلالي" إشكالية أوسع في أنظمة التقييم الآلي التي تفشل في مراعاة إدراك النموذج واتخاذه القرارات. عندما تعجز أطر التقييم عن التمييز بين "لا يستطيع الحل" و "يختار عدم العدّ الشامل"، فإنها تخاطر باستخلاص استنتاجات خاطئة عن القدرات الأساسية.

*

Anthropic

+

Open Philanthropy

2.1 تبعيات التقييم الصارم

قد تؤدي قيود التقييم هذه إلى أخطاء تحليلية أخرى. تأمل الحجة الإحصائية التالية: إذا قمنا بتقييم حلول برج هانوي حرفًا بحرف دون السماح بتصحيح الأخطاء، فإن احتمال التنفيذ المثالي يصبح:

$$P(\text{جميعها صحيحة}) = p^T \quad (1)$$

حيث p تمثل دقة الرمز الواحد، وT إجمالي الرموز. عند10,000 = T رمز:

- p = 0.9999: $P(\text{نجاح}) < 37\%$
- p = 0.999: $P(\text{نجاح}) < 0.005\%$

لقد تم تقديم هذا النوع من حجج "الحتمية الإحصائية" في الأدبيات كقيد جوهري في توسيع نماذج اللغة الكبيرة [٣]. إلا أنه يفترض أن النماذج لا تستطيع إدراك قيودها أو التكيف معها، وهو افتراض يناقضه الدليل الوارد أعلاه.

3 إشكالية اللغز المستحيل

تتفاقم مشكلات التقييم بشكل كبير في تجارب عبور النهر (River Crossing). اختبر شجاعي وزملاؤه حالات ذات6 ≥ N فاعل/وكيل مع سعة قارب b = 3. ومع ذلك، فهناك نتيجة راسخة [٤] تفيد بأن لغز المُبشِّرين والآكلي البشر (Missionaries-Cannibals) لا يحل عند N > 5 مع b = 3.

بتسجيل هذه الحالات المستحيلة تلقائيًّا كإخفاقات، يُظهر الباحثون **دون قصد** مخاطر التقييم البرمجي البحت. تتلقى النماذج تقديرًا صفريًّا ليس بسبب إخفاقات استدلالية، بل بسبب **تعرّفها الصحيح** على المسائل المستعصية — وهو ما يعادل معاقبة محلل SAT (أي مُحلِّل معادلات) لإرجاعه "غير قابلة للإشباع" على صيغة غير قابلة للإشباع.

4 القيود المادية للرموز تُسبب الانهيار الظاهري

بالعودة إلى تحليل برج هانوي، يمكننا تحديد العلاقة بين حجم المشكلة ومتطلبات الرموز. يتطلب إطار التقييم إخراج تسلسل الحركات الكامل في كل خطوة، مما يؤدي إلى نموٌّ تربيعي في عدد الرموز. إذا كانت هناك حاجة إلى5 رموز لكل حركة:

$$T(N) \approx 5(2N - 1)2 + C \quad (2)$$

بالنظر إلى ميزانيات الرموز المخصصة (64,000 لكل من Claude-3.7-Sonnet وDeepSeek-R1؛ 100,000 لـ o3-mini)، فإن الحد الأقصى لحجم المسائل القابلة للحل هو:

$$N_{max} \approx \lfloor \log_2(\sqrt{L_{max}/5}) \rfloor \quad (3)$$

$$\approx \begin{cases} 7 - 8 \text{ Claude-3.7, DeepSeek-R1} \\ 8 \text{ o3-mini} \end{cases} \quad (4)$$

الانهيار المبلغ عنه بعد هذه الأحجام يتوافق مع هذه القيود.

5 تمثيلات بديلة تعيد الأداء

لاختبار ما إذا كانت الإخفاقات تعكس قيودًا في الاستدلال أم في التنسيق، أجرينا اختبارات أولية لنفس النماذج على برج هانوي بحجم N = 15 باستخدام تمثيل مختلف:
التلقينة:"حلّ لغز برج هانوي لـ 15 قرصًا. أخرِج دالة بلغة Lua تطبع الحل عند استدعائها."
النتائج: دقة عالية جدًّا عبر النماذج المختبرة (Claude-3.7-Sonnet، Claude Opus 4، OpenAI o3، Google Gemini 2.5)، مع اكتمال الحل في أقل من 5,000 رمز.
نفّذت الحلول المُولَّدة الخوارزمية التكرارية (Recursive Algorithm) بشكل صحيح، مما يثبت سلامة قدرات الاستدلال عند تحريرها من متطلبات العدّ الشامل.

6 إعادة تقييم ادعاءات التعقيد

يستخدم الباحثون "العمق التركيبي" (Compositional Depth - أقل عدد حركات) كمقياس للتعقيد، لكن هذا يخلط بين التنفيذ الآلي وصعوبة حل المشكلة:

اللغز	طول الحل	عامل التفرع	البحث المطلوب
برج هانوي	$2^N - 1$	1	لا
عبور النهر	~ 4N	4 <	نعم (NP-hard)
عالم الكتل	~ 2N	$O(N^2)$	نعم (PSPACE)

الجدول 1: تعقيد المشكلة لا يُحدَّد بطول الحل وحده

برج هانوي - رغم احتياجه حركات أُسّية - له عملية قرار تافهة O(1) لكل حركة. بينما يتطلب عبور النهر - بأقل حركات - إشباع قيود معقدة (Constraint Satisfaction) وبحثًا. هذا يفسر قدرة النماذج على تنفيذ 100+ حركة في هانوي مع فشلها في حل مسائل عبور نهر بـ 5 حركات فقط.

7 الخاتمة

نتائج شجاعي وزملائه تُظهر أن النماذج لا تستطيع إخراج رموز تفوق حدود سعتها السياقية وأن التقييم البرمجي قد يُغفل إمكانات النماذج واستحالة الألغاز وأن طول الحل يُتنبّأ بشكل ضعيف بصعوبة المشكلة هذه رؤى هندسية قيّمة، لكنها لا تدعم ادعاءات القيود الاستدلالية الجوهرية.

أعمال مستقبلية مقترحة:
1. تصميم تقييمات تُميِّز بين قدرة الاستدلال و قيود المخرجات
2. التحقق من إمكانية حل اللغز قبل تقييم أداء النموذج
3. استخدام مقاييس تعقيد تعكس **الصعوبة الحسابية** لا طول الحل فقط
4. اعتماد تمثيلات حل متعددة لفصل **الفهم الخوارزمي** عن التنفيذ

> السؤال ليس هل تستطيع نماذج الاستدلال الكبيرة (LRMs) أن تستدل؟، بل هل تستطيع تقييماتنا التمييز بين الاستدلال والكتابة؟

بسبب قيود الميزانية، لم نتمكن من إجراء تجارب كافية لعيّنة إحصائية قوية. يبقى التحقق التجريبي الكامل عملاً مستقبليًّا.

شكر و تقدير

نتشكر رايان غرينبلات Gemini 2.5،o3، وجميع الذين أشاروا إلى عدم تطابق الأقواس في مسودة سابقة، على ملاحظاتهم القيّمة.

المراجع

[1] Shojaee, P., Mirzadeh, I., Alizadeh, K., et al. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2501.12948.

[2] @scaling01. (2025). Twitter thread on LRM replication. https://x.com/scaling01/status/1931817022926839909/photo/1

[3] Dziri, N., Lu, X., Sclar, M., et al. (2023). Faith and fate: Limits of transformers on compositionality. Advances in Neural Information Processing Systems, 36.

[4] Efimova, E. A. (2018). River Crossing Problems: Algebraic Approach. arXiv:1802.09369.