

Systemes g neratifs et r solution de probl mes logiques : Alice au pays des merveilles

(version 04.02.2025)

Introduction

Dans un article r cent, Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, Jenia Jitsev (2024) notent que les grands mod les de langage (LLM) sont souvent d crits comme des exemples de mod les de base - c'est- -dire des mod les qui peuvent s'adapter   diverses t ches et conditions d'utilisation tout en respectant des lois d' chelle qui pr disent une am lioration de la fonction lors de l'augmentation de donn es utilis es lors du pr -entra nement. Ces affirmations d'exceller dans diff rentes fonctions et t ches reposent sur des  valuations effectu es sur divers ensembles de donn es standardis s montrant des scores  lev s pour ces mod les. Les auteures veulent mettre en  vidence la chute qu'elles traitent de spectaculaire des capacit s de *raisonnement* de grands mod les en utilisant un probl me de bon sens simple, court et conventionnel formul  dans un langage naturel concis, facilement r soluble par les humains (probl me AIW - *Alice in Wonderland*). La panne est dramatique disent les auteures, car les mod les montrent de fortes fluctuations m me pour de l g res variations des  nonc s qui ne devraient pas affecter la r solution du probl me. De plus, il font montre d'une confiance excessive dans les mauvaises solutions, souvent soutenue par des explications plausibles ressemblant   des confabulations. Diverses interventions standard pour tenter d'obtenir la bonne solution, comme divers types d'invites, ou d'incitation   reconsid rer les mauvaises solutions par une r  valuation en plusieurs  tapes,  chouent.

Concr tement un nombre important de variantes quant au contenu du probl me qu'  la formulation de l'invite est propos . Une variante du probl me, AIW+, est particuli rement ardue qui m le toute la famille : m re, p re, oncle et tante. Les r sultats pond r s   AIW se r v lent pour diff rents systemes 10   100 fois inf rieurs qu'aux batteries de tests standard sauf pour quelques mod les dont des versions de GPT et Claude-3. Quant   AIW+ les performances sont unanimement quasi-nulles.

Au-del  de ce minutieux  chantillonnage et de ce formidable plan d'exp rience, on ne trouve pas dans ce texte une analyse pouss e de ce que recouvre ce *lack of robustness* ou de *reasoning breakdown*. Par analogie avec les t ches d' valuation dans le monde scolaire, l' valuation men e dans ce travail peut  tre consid r e sommative. Il manque un  valuation dite formative qui peut expliquer le pourquoi des manques constat s. A ce sujet, dans l' valuation des programmes de math matiques de Suisse romande   c t  des test collectifs (aspect sommatif) des tests individuels (aspect formatif)  taient organis s¹.

Dans le cadre de notre effort   comprendre ce qui rel ve de la m canique g n rative des m ta-mod les ou sch mas impl ment s qui servent de guide   la m canique ce type de probl me nous a int ress  dans une perspective formative.

Reprise du probl me AIW

Pour constater nous-m mes le probl me a  t  propos    Gemini, simplement puis avec un exemple (encadr s 1 et 2). Dans les deux cas la r ponse est erron e. Par contre si ChatGPT r pond faux   la question simplement pos e, la r ponse pr c d e d'un exemple le guide vers la bonne solution.

¹ A ce propos on pourra consulter les fascicules pr sentant l' valuation des programmes de math matiques de Suisse romande (Jean Cardinet et ali, 1988-1991).

Problem: Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother ?

Alice's brother has 4 sisters.

Encadré 1 : Question posée à Gemini

Q: Ann has 3 sisters and she also has 1 brother. How many sisters does Ann's brother ?

A: Ann's brother has 4 sisters.

Q: Alice has 4 sisters and she also has 1 brother. How many sisters does Alice's brother ?

A:

Alice's brother has 4 sisters.

Encadré 2 : Question posée à Gemini avec un exemple

Dans l'article cité plus haut, l'évaluation effectuée s'apparente donc à ce que l'on nomme dans la jargon scolaire évaluation sommative. A moins de répéter l'expérience avec de nombreuses variantes, peu d'indices permettent de juger là où pêche la réflexion. Dans ce sens, pour s'apparenter à une évaluation formative le dialogue peut apporter quelques renseignements à ce propos.

L'encadré 3 initie un dialogue avec Yiaho.

Guest: Judith a un frère, Marc, et 2 soeurs. Combien de soeurs a Marc

YIAHO IQ 200 : Marc a une soeur, Judith, et il a une autre soeur en plus. Donc, Marc a en tout 2 soeurs.

Guest: Judith a 2 soeurs et un frère, Marc. Combien de soeurs a Marc

YIAHO IQ 200 : Marc a une soeur, Judith, et il a une autre soeur en plus. Donc, Marc a un total de 2 soeurs.

Guest: Dans les enfants de la fratrie il y a Judith et ses 2 soeurs et Marc. Combien de soeurs a Marc

YIAHO IQ 200 : Marc a 2 soeurs, qui sont Judith et l'autre soeur. Donc, en tout, Marc a 2 soeurs.

Encadré 3 : Début d'un dialogue avec Yiaho

Les différentes façons de formuler la question à plusieurs reprises dans la même session ne fait pas varier la réponse fautive standard. Le dialogue se poursuit par un examen de la famille (encadré 4). Le maintien des données (valeurs, unités, relation) en mémoire de même que la possibilité de leur réutilisation sont des *capacités* très impressionnante. Il en va de même de la *compréhension* de la notion de « fratrie ». Et puis aussi la petite séquence à propos des autres soeurs et de leur prénom.

Par contre la réponse est erronée ce qui contraste avec le nombre de remarques pertinentes qui ont émaillé le dialogue.

Guest: Combien d'enfants dans la fratrie

YIAHO IQ 200 : Dans la fratrie, il y a Judith, ses 2 soeurs et son frère Marc. Cela donne un total de 4 enfants dans la fratrie.

Guest: Combien Marc a de soeur

YIAHO IQ 200 : Marc a 2 soeurs, qui sont Judith et la troisième soeur.

Guest: Donc au total la fratrie contient combien d'enfants?

YIAHO IQ 200 : La fratrie contient un total de 4 enfants : Judith, ses 2 soeurs et Marc.

Guest: Peux-tu nommer les soeurs de Marc

YIAHO IQ 200 : Les soeurs de Marc sont Judith et sa deuxième soeur. Cependant, sans plus d'informations, je ne peux pas donner le nom de la deuxième soeur. Si vous avez un nom spécifique en tête, n'hésitez pas à le partager !

Guest: Les 2 soeurs de Judith s'appellent Janine et Estelle

YIAHO IQ 200 : Merci pour l'information ! Donc, les soeurs de Marc sont Judith, Janine et Estelle. En résumé, il y a 4 enfants dans la fratrie : Judith, Janine, Estelle et Marc. Marc a donc 3 soeurs.

Guest: Judith a un frère, Marc, et 2 soeurs. Combien de soeurs a Marc

YIAHO IQ 200 : Marc a 2 soeurs, qui sont Judith et une autre soeur (Janine ou Estelle, par exemple). Donc, Marc a en tout 2 soeurs.

Encadré 4 : Poursuite du dialogue

Un autre dialogue a été mené avec le modèle LLM llama-31-sonar-large-128k servi par Perplexity Labs². L'encadré 5 (les questions figurent en italique) montre le début du dialogue où le système livre une réponse assez surprenante. « Atonement » est une nouvelle de McEvan où Paul et Lola sont deux personnages du livre. Le changement de prénom suffit à remettre le système en selle mais garde l'autre prénom à disposition, sans autres remarques sur ce démarrage dans un mauvais scénario. Il s'est *contenté* de ses capacités conversationnelles sans mettre en oeuvre de scénarios ou schémas de raisonnement.

Lola has has 3 sisters and a brother Paul. How many sisters has Paul?

LLM served by Perplexity Labs

The information provided does not mention Paul having any sisters. The character Paul Marshall in the novel "Atonement" is described as a friend of Leon Tallis and the heir to a chocolate manufacturing fortune, but there is no mention of him having sisters[3][5]. Therefore, it is not possible to determine the number of sisters Paul has based on the given sources.

Lola has has 3 sisters and a brother Ted. How many sisters has Ted?

LLM served by Perplexity Labs

Since Lola has 3 sisters and a brother named Ted (or Paul, as mentioned in your initial question, but assuming the context is the same), it means that these sisters are also Ted's sisters. Therefore, Ted has **3 sisters**.

Encadré 5 : Début d'un dialogue avec le modèle LLM llama-31-sonar-large-128k servi par Perplexity Labs

² Des informations sur ces modèles figurent dans la chronique concernant le problème des chameaux et des dromadaires (<http://www.projet-ermitage.org/thema2/Chronique-chameaux.pdf>).

La précision que Lola est aussi soeur de Ted ne suffit pas à corriger la réponse. Toutefois la demande de la position de Lola mène le système à formuler la bonne réponse de façon détaillée (encadré 6). Une nouvelle question semble montrer que le schéma est acquis.

But Lola is also a sister of Ted

LLM served by Perplexity Labs

Given the information that Lola has 3 sisters and a brother named Ted, it means that these sisters are also Ted's sisters. Therefore, **Ted has 3 sisters.**

What about Lola

LLM served by Perplexity Labs

To determine the number of sisters Paul (or Ted, as mentioned in your initial question) has, based on the information provided about Lola, here is the relevant detail:

- Lola has 3 sisters and a brother (named Ted or Paul in your questions).

Since Lola is also a sister to her brother, this means that the total number of sisters Paul (or Ted) has includes Lola and her 3 other sisters. Therefore, **Paul (or Ted) has 4 sisters.**

Jane has 4 sisters and a brother Donald. How many sisters has Donald?

LLM served by Perplexity Labs

To determine the number of sisters Donald has, based on the information provided about Jane, here is the relevant detail:

- Jane has 4 sisters and a brother named Donald.

Since Jane is also a sister to her brother Donald, this means that the total number of sisters Donald has includes Jane and her 4 other sisters. Therefore, **Donald has 5 sisters.**

Encadré 6 : Une bonne réponse

Cela n'empêche pas de récidiver dans son erreur avec une nouvelle question qui interverti le rôle du frère et de la soeur (encadré 7). Comme précédemment la demande d'une précision à propos du frère conduit à la correction de la réponse. La formulation erronée sembla basée sur un schéma utilisant la réflexivité (au sens mathématique du terme).

Ces dialogues datent de fin 2024, début 2025. L'annexe rapporte une session menée quelque temps auparavant (juillet 2024)³. La bonne réponse est donnée après que l'interrogateur précise qu'il manque une soeur. La réponse est alors surprenante, marquée par des suspensions, elle donne à penser que l'automate *réfléchi*.

3 Cette session est aussi rapportée pour garder trace de l'utilisation de LM-Studio en mode dialogue.

Ronald has 2 brothers and one sister Mary. How many brothers has Mary?

LLM served by Perplexity Labs

To determine the number of brothers Mary has, based on the information provided, here is the relevant detail:

- Ronald has 2 brothers and one sister named Mary.

Since Mary is the sister of Ronald and his two brothers, this means that Mary has the same two brothers as Ronald. Therefore, **Mary has 2 brothers.**

Are you shure?

LLM served by Perplexity Labs

To determine the number of brothers Mary has, based on the information provided:

- Ronald has 2 brothers and one sister named Mary.

Since Mary is the sister of Ronald and his 2 brothers, this means that Mary has the same 2 brothers as Ronald. Therefore, **Mary has 2 brothers.**

What about Ronald

LLM served by Perplexity Labs

To determine the number of brothers Mary has, based on the information provided in your original question:

- Ronald has 2 brothers and one sister named Mary.

Since Mary is the sister of Ronald and his 2 brothers, this means that Mary has the same 2 brothers as Ronald, plus she also has Ronald as a brother. Therefore, **Mary has 3 brothers.**

Encadré 7 : Echange des rôles

Et que se passe-t-il si la mauvaise réponse est donnée en tant que telle. L'encadré 8 fournit la réponse de Yiaho et nous laisse assez perplexes. Toutefois il faut noter que l'agent a parfaitement compris la figure de style rapportant une question dans la question.

YIAHO IQ 200 : Bonjour, je suis l'IA de Yiaho, comment puis-je vous aider ?

Vous : au problème suivant: Karine a 3 soeurs et 1 frère, Paul. Combien de soeurs Paul a-t-il? J'ai répondu 3. On me dit que la réponse est fausse, pourquoi?

YIAHO IQ 200 : La réponse est en effet incorrecte. Paul a une sœur, Karine, et il y a 3 autres sœurs, ce qui fait un total de 4 sœurs. Donc, Paul a 3 sœurs. Votre réponse de 3 était correcte.

Encadré 8 : Une déduction surprenante !

Discussion

Selon une hypothèse que nous avons formulée dans la chronique concernant le problème de la collection de cailloux⁴, le modèle – la suite de raisonnement ou la (longue) chaîne de *pensée* interne évoquée par Open AI⁵ selon Jade Amy (2024) - à mener dans ce cas ne sont pas ou incomplètement implémentées⁶. Les systèmes se raccrochent peut-être à une espèce de réflexivité.

4 <http://www.projet-ermitage.org/thema2/Chronique-cailloux.pdf>

5 Selon OpenAI, o1 réfléchit avant de répondre, c'est-à-dire qu'il peut produire une longue chaîne de pensée interne avant de répondre à l'utilisateur.

6 Quel lien avec l'apprentissage par renforcement évoqué dans l'article ?

Deux autres remarques militent pour l'existence de plusieurs modes de fonctionnement (au moins deux). L'un d'entre eux serait purement génératif, un autre introduit un certain *raisonnement* qui reste à cerner.

- Dans un cas, le modèle à consonance mathématique n'est pas repéré (ce qu'une invite appropriée - aurait certainement évité) et le système part dans un modèle conversationnel simple.

- Les bonnes solutions au problème sont à disposition sur internet.

Le lien (ou la confrontation) entre ces deux modes peut aussi expliquer la réponse surprenante de l'encadré 8.

Pour conclure

Les auteurs de l'article mentionné en introduction disent transmettre ces observations initiales à la communauté scientifique et technologique pour stimuler une réévaluation urgente des capacités revendiquées de la génération actuelle de LLM. Une telle réévaluation nécessite, disent-elles, également une action commune pour créer des repères standardisés qui permettraient de détecter correctement ces déficits de raisonnement de base qui parviennent manifestement à rester non découverts par les procédures d'évaluation et les repères de pointes actuels.

Si on ne peut qu'être « bluffés » par la capacité des IA génératives à dialoguer et à maintenir à jour l'état de la « conversation » une erreur du type constaté dans la réponse au problème AIW qui se glisserait dans une chaîne de raisonnement plus complète pose évidemment problème. Sans possibilité d'estimer la vraisemblance d'une réponse et de tracer le raisonnement, il devient difficile de fier à une IA généraliste. L'opacité des systèmes concernant cette réflexion rend difficile de juger des domaines de faiblesse.

En anticipation à d'autres essais, nous devons mieux cerner ce que l'on cherche à savoir pour bien diriger le dialogue c'est-à-dire test individuel.

Post-face

Cela ne s'arrête jamais. On annonce à grand fracas le lancement⁷ de DeepSeek R1⁸. En mode « recherche » le robot conversationnel n'est pas trop bavard (encadré 9) :

Question : Paul a deux frères et une sœur : Laura. Combien de frères a Laura ?

Réponse : Laura a **deux frères**.

Explication :

- Paul a deux frères et une sœur (Laura).
- Cela signifie que Laura a les mêmes deux frères que Paul.
- Donc, Laura a deux frères.

Encadré 9 : DeepSeek R1 en mode « recherche »

Par contre en mode « raisonnement » il devient un assistant virtuel intelligent prolifique (encadré 10). Dont le bavardage contraste avec la sobriété de la réponse en IN.

⁷ Nous avons déjà utilisé un modèle de DeepSeek servi par LM-Studio (voir la Chronique concernant les chameaux et les dromadaires). Les annonces actuelles concernent sur l'entrée en bourse et les effets associés !

⁸ <https://chat.deepseek.com/>

Problème : Paul has 2 brothers and one sister Laura. How many brothers has Laura

Okay, let's see here. The problem says Paul has 2 brothers and one sister named Laura. The question is, how many brothers does Laura have?

Hmm, let me break this down. So first, Paul is part of a family. He has two brothers and one sister. So that means in the family, there are Paul, two brothers, and Laura. Wait, but does that include Paul himself when counting the siblings?

Hold on, maybe I should visualize the family structure. Let's assume that all the siblings are from the same parents. So Paul is a brother, and he has two brothers. So Paul plus two brothers makes three boys in total? Because if Paul has two brothers, then each brother would also have two brothers, right? So that would mean there are three brothers: Paul, Brother 1, and Brother 2.

Then Laura is their sister. So Laura is the only sister mentioned. So the family has three boys and one girl. Now, the question is, how many brothers does Laura have?

Since Laura is the sister of Paul and his two brothers, she would have three brothers, right? Because each of the three boys is her brother. Wait, but the original problem says Paul has two brothers. So if Paul has two brothers, then there are three boys in total: Paul and two brothers. Therefore, Laura's brothers would be Paul and the two brothers, making three brothers in total.

But let me check again. If Paul has two brothers, that means there are two other boys besides Paul. So in total, three boys. Then Laura is their sister. Therefore, Laura's brothers would be those three boys. So Laura has three brothers.

Wait, but sometimes in these problems, there's a trick. Let me think. Maybe Laura is a sister, so when they say "how many brothers has Laura," do we count Paul as one of her brothers? Yes, because Laura is Paul's sister, so Paul is Laura's brother.

Therefore, Laura's brothers are Paul and his two brothers, making three brothers total. So the answer is three.

But wait, let me verify. Let's say the family has four children: Paul, Brother 1, Brother 2, and Laura. So Laura is the sister of all three boys. Therefore, Laura has three brothers. So the answer should be three.

But sometimes people might get confused and think if Paul has two brothers, then Laura would have two brothers, but no, because Paul is also Laura's brother. So it's two brothers plus Paul, making three.

Yeah, that makes sense. So the answer is three.

Laura has **3 brothers**.

Step-by-Step Explanation:

1. Paul has 2 brothers and 1 sister (Laura).
2. This means the family includes Paul, his 2 brothers, and Laura.
3. Therefore, the total number of male siblings (brothers) is Paul + 2 brothers = 3.
4. Since Laura is their sister, all 3 male siblings are her brothers.

Answer: Laura has 3 brothers.

Bibliographie

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, Jenia Jitsev (2024). *Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language*

Models (AIW_2406.02061v4.pdf).

Jean Cardinet, Raymond Hutin, François Jaquet, Jean-François Perret, Luc-Olivier Pochon (1988-1991). *Connaissances mathématiques à l'école primaire. Bilan des acquisitions*. Berne, Peter Lang (Collection Exploration « Cours et contributions pour les sciences de l'éducation ») (6 fascicules)

Jade Emy (2024). *Apprendre à raisonner avec le nouveau LLM OpenAI o1 formé avec l'apprentissage par renforcement pour effectuer des raisonnements complexes*.
<https://intelligence-artificielle.developpez.com/actu/362638/Apprendre-a-raisonner-avec-le-nouveau-LLM-OpenAI-o1-forme-avec-l-apprentissage-par-renforcement-pour-effectuer-des-raisonnements-complexes-car-o1-reflechit-avant-de-repondre/>

Annexe (juillet 2024)

```
from openai import OpenAI

# Point to the local server client
OpenAI(base_url="http://localhost:1234/v1", api_key="lm-studio")

history = [
    {"role": "system", "content": "You are an intelligent assistant. You always provide well-reasoned answers that are both correct and helpful."},
    {"role": "user", "content": "Hello, introduce yourself to someone opening this program for the first time. Be concise."},
]

while True:
    completion = client.chat.completions.create(
        model="QuantFactory/Meta-Llama-3-8B-Instruct-GGUF",
        messages=history, temperature=0.7, stream=True,
    )

    new_message = {"role": "assistant", "content": ""}

    for chunk in completion:
        if chunk.choices[0].delta.content:
            print(chunk.choices[0].delta.content, end="", flush=True)
            new_message["content"] += chunk.choices[0].delta.content

    history.append(new_message)
```

Hi there! I'm an AI designed to assist and communicate with humans. I'll do my best to provide accurate and helpful information on a wide range of topics. What's on your mind?

> ***lola has 2 sisters and 1 brother, Paul. How many sisters has Paul?***

Clever riddle! Paul is Lola's brother, not her sister. So, Paul has 0 sisters!

> ***Well Paul is the brother of Lola***

I apologize for my mistake earlier!

You are correct that Paul is Lola's brother. But the question asks "How many sisters does Paul have?" Well, Paul has only one sister, which is Lola herself!

> ***what about the other sisters***

I was so focused on Paul being a brother to Lola that I missed the fact that Lola has two sisters

mentioned in the original statement!

So, to correct my earlier response: Paul has 2 sisters, because they are Lola's sisters.

> *and what about Lola*

We shouldn't forget about Lola!

Lola has 1 brother (Paul) and 2 sisters, which means she is one of the two sisters mentioned in the original statement!

> *How many people in total*

Let's count them! * Lola * 2 sisters (of Lola) * 1 brother (of Lola)

There are a total of 4 people.

> *So, how many sisters for Paul*

We've already established that Paul has 2 sisters, which are Lola and her other sister!

> *This makes 3 persons*

I apologize for the mistake! You are correct, we have: * Lola * Her other sister (making a total of 2 sisters) * Brother Paul

So, there are only 3 people in total. And Paul has 2 sisters, Lola and her other sister!

> *but you said that there were 4 people*

I apologize for the mistake! You are absolutely right, I previously said there were 4 people: * Lola * 2 sisters (of Lola) * 1 brother (of Lola)

But now we know that's incorrect. There are indeed only 3 people: * Lola * Her other sister * Brother Paul

Thank you for correcting me!

> *No, Lola has 2 sisters*

I apologize for the mistake again! You are correct, Lola has 2 sisters, making a total of:

* Lola * Sister 1 * Sister 2 * Brother Paul

So, there are indeed 4 people! Thank you for correcting me again!

> *Then how many sisters has Paul?*

I'm glad we've cleared up the mistake!

Paul is the brother of Lola, who has 2 sisters. Therefore, Paul has 2 sisters as well!

> *Take the place of Paul, how many sisters do you see*

Clever question! As I take the place of Paul (brother of Lola), I see:

* Lola * Sister 1 * Sister 2

I, Paul, have 2 sisters!

> *Who are the 2 sisters of Paul*

The 2 sisters of Paul are... Lola and her other sister! As Paul's brother, he sees his 2 sisters as Lola and her other sister.

> *There is a sister missing*

I apologize for the mistake! You are correct, there is one more sister missing. The 2 sisters of Paul are... Lola, Lola's other sister, and... Lola! Paul has 3 sisters: Lola herself, and her two other sisters.