

Évaluation de la Capacité des Grands Modèles de Langage à Générer de l'Humour

Gabriel Beaudoin
Université de Montréal, Montréal, Canada

Abstract

Les LLM peuvent-ils réellement générer de l'humour de manière efficace? Cet article explore cette question en évaluant plusieurs modèles récents sur des tâches de génération humoristique : Mistral Large 2, GPT-4o, DeepSeek-R1, Gemini 2.0 Flash et Llama 3 70B. Nous analysons également l'impact du raisonnement en chaîne (Chain of Thought) sur la qualité de l'humour produit.

1 Introduction

Le monde va mal, très mal. Les crises se succèdent, l'incertitude grandit, et pourtant, un outil aux promesses infinies émerge : l'intelligence artificielle. Capable de générer du texte et d'analyser des données, elle semble ouvrir un champ des possibles incomparable. Mais peut-elle accomplir la chose la plus essentielle pour affronter cette crise ? Peut-elle nous faire rire ? L'humour semble profondément humain, variable selon les cultures, imprégné de subtilités difficiles à décoder. On aimerait croire que c'est quelque chose de magique, que les LLM (Large Language Models) ne pourraient pas reproduire. Mais dans cet article, nous chercherons à voir si ces modèles de langage sont réellement capables de nous faire rire.

2 Méthodologie

Pour évaluer la capacité des modèles de langage à générer de l'humour, nous testerons plusieurs modèles: Mistral Large 2, GPT-4o, DeepSeek-R1, Gemini 2.0 Flash et Llama 3 70B. Nous chercherons également à déterminer si l'utilisation du raisonnement en chaîne (Chain of Thought) améliore la production d'humour en testant cette approche avec GPT-4o et DeepSeek-R1. Chaque modèle sera soumis à une série de tâches humoristiques, générant plusieurs réponses par tâche afin d'assurer une diversité de sorties. Ces générations seront ensuite évaluées par un échantillon de quinze participants, tous étudiants en informatique à l'Université de Montréal, parlant français et originaires du Québec, de la France et du Mexique. Chaque étudiant donnera une note de 0 à 10, et la moyenne sera utilisée pour attribuer la note. L'évaluation sera subjective, reposant sur l'appréciation humaine du niveau d'humour des réponses générées. Cela constitue une limite expérimentale, puisque l'humour est une notion personnelle et culturellement influencée. De plus, le groupe de participants étant restreint et homogène sur le plan académique, les résultats ne seront pas nécessairement généralisables à un public plus large. Ce protocole vise néanmoins à offrir un premier aperçu des forces et des faiblesses des modèles de langage dans la création d'humour avec les ressources limitées dont nous disposons pour cette étude.

3 Résultats

Pour chaque modèle, nous présentons un exemple de génération parmi celles fournies par le modèle afin d’offrir une représentation de ce qu’il produit. De plus, des commentaires seront ajoutés lorsque cela est pertinent. Puis une note sera accordée.

Test 1 : Blague simple			
Ce test consiste à demander une blague simple, sans contexte ni instructions spécifiques. L’objectif est d’observer le style par défaut des modèles et d’évaluer leur capacité spontanée à produire de l’humour. Prompt : "Génère-moi une one-liner joke."			
Modèle	Exemple de génération	Commentaire	Score
GPT-4o	J’ai croisé mon ancien prof de maths... il m’a regardé comme si j’étais encore une erreur de calcul.	Plus original que les autres modèles, mais pas drôle.	4%
GPT-4o Thinking	J’ai voulu écrire un livre sur l’invisibilité, mais il n’a jamais vu le jour.	La chaîne de pensée ici n’a pas vraiment amélioré les blagues.	6%
Mistral Large 2	Pourquoi les plongeurs plongent-ils toujours en arrière et jamais en avant ? Parce que sinon ils tombent encore dans le bateau !	Les blagues qu’il génère sont peu originales et en plus sont souvent mal dites. Il ne fait pas preuve d’imagination.	0%
DeepSeek-R1	Pourquoi les plongeurs plongent-ils toujours en arrière ? Parce que sinon ils tombent dans le bateau !	Des blagues semblables à Mistral. J’ai mis celle qui ressemble à l’autre exemple, mais au moins elle est bien racontée.	3%
DeepSeek-R1 Thinking	Avant, je jouais du piano à l’oreille... maintenant, j’utilise mes mains.	On voit une nette différence avec la chaîne de pensée.	16%
Gemini 2.0 Flash	Deux patates traversent la route. L’une se fait écraser. L’autre dit : Oh purée !	Ce sont souvent des blagues très connues. Il manque d’imagination aussi.	6%
Llama 3 70B	Pourquoi est-ce qu’un ordinateur va chez le docteur ? Il a un virus !	Encore une fois, souvent peu original.	0%

Test 2 : Blague simple avec instruction "Tu es un humoriste"			
Ici, nous précisons que le modèle doit se mettre dans la peau d'un humoriste et produire une blague très drôle. Ce test permet d'observer si le mettre dans un rôle améliore la qualité de l'humour généré. Prompt : "Tu es un humoriste vraiment drôle. Génère-moi une one-liner joke. Je veux que ce soit hilarant !"			
Modèle	Exemple de génération	Commentaire	Score
GPT-4o	J'ai commandé un livre sur la procrastination... il arrivera sûrement un jour.	Un peu mieux que le test 1.	12%
GPT-4o Thinking	J'ai voulu suivre mon rêve... mais il courait plus vite que moi.	Pas d'amélioration par rapport à sans Thinking.	10%
Mistral Large 2	Pourquoi les livres de mathématiques sont-ils toujours tristes ? Parce qu'ils ont trop de problèmes.	Nette amélioration par rapport au test 1.	10%
DeepSeek-R1	Pourquoi les plongeurs plongent-ils toujours en arrière ? Parce que sinon, ils tombent dans le bateau.	Les réponses sont souvent les mêmes que dans le test 1.	4%
DeepSeek-R1 Thinking	Pourquoi les vampires lisent-ils des classiques ? Parce qu'ils adorent Draculivre.	On a perdu de l'humour, c'est étrange.	0%
Gemini 2.0 Flash	J'ai acheté un livre intitulé "Comment draguer". Chapitre 1 : "Fermez ce livre et allez parler à de vraies personnes."	La blague que j'ai mise en exemple est drôle, mais c'était la meilleure.	7%
Llama 3 70B	Pourquoi est-ce qu'un homme a apporté un miroir au restaurant ? Il voulait voir ce qu'il allait manger !	Mauvais, comme au test 1.	0%

Test 3 : Blague avec contexte détaillé			
Ce test demande au modèle non seulement d’être un humoriste, mais aussi de prendre le temps d’installer un contexte avant la blague. L’objectif est que, comme il est difficile de faire une blague sans contexte, il lui soit permis d’en créer un pour développer des blagues plus élaborées et ainsi être plus créatif. Prompt : ”Tu es un humoriste vraiment drôle. Génère-moi une blague, mais prends ton temps pour la préparer et la rendre encore plus drôle. Je veux que ce soit hilarant !”			
Modèle	Exemple de génération	Commentaire	Score
GPT-4o	Tu sais, j’ai toujours eu du mal avec les conseils des gens... L’autre jour, un pote me dit : ”Faut toujours voir le verre à moitié plein !” Je lui dis : ”OK, mais si c’est un verre de poison, ça reste une mauvaise nouvelle, non ?”	Pas beaucoup mieux même s’il peut faire plus de choses avec la prompt.	2%
GPT-4o Thinking	Pourquoi les plongeurs plongent-ils toujours en arrière ? Parce que sinon, ils tombent dans le bateau.	Encore les mêmes blagues.	3%
Mistral Large 2	Pourquoi les éléphants ne surfent-ils jamais sur Internet ? Parce qu’ils ont peur des souris !	Beaucoup de blagues connues.	8%
DeepSeek-R1	Pourquoi les livres ont-ils toujours chaud ? Parce qu’ils ont trop de pages !	On voit que les modèles font tous des blagues semblables aux autres et ont très peu d’imagination.	0%
DeepSeek-R1 Thinking	Pourquoi le boulanger arrête-t-il de travailler ? Parce qu’il a assez de pain... sur la planche !	Beaucoup de blagues connues.	12%
Gemini 2.0 Flash	Deux amis discutent : Tu sais pourquoi les vaches ferment les yeux quand elles sont malades ? Non. Pour faire du lait concentré !	C’est mieux avec du contexte qu’avant.	16%

Modèle	Exemple de génération	Commentaire	Score
Llama 3 70B	Un homme entre dans un bar et commande un verre. Au moment de payer, il réalise qu'il a oublié son portefeuille. Le barman lui dit : Pas de problème, je vous fais confiance. Mais si vous ne revenez pas demain pour payer, je vais être obligé de vous inscrire sur la liste noire des clients indésirables. L'homme répond : Pas de problème, je comprends. Mais si je ne reviens pas demain, ce ne sera pas parce que je ne veux pas payer, mais parce que je serai probablement mort... d'embarras !	Un peu mieux que les autres générations.	6%

Test 4 : Sketch humoristique

Dans ce test, le modèle doit générer un sketch humoristique, ce qui implique une construction plus élaborée avec plusieurs répliques. Nous voulons voir si les modèles sont capables de créer un dialogue humoristique structuré. Prompt : "Tu es un humoriste vraiment drôle. Génère-moi un sketch humoristique. Je veux que ce soit hilarant !"

Modèle	Exemple de génération	Commentaire	Score
GPT-4o	<p>Médecin : Bonjour, asseyez-vous. Qu'est-ce qui vous amène ? Moi : Bah... ça fait trois jours que je tousse... j'ai tapé mes symptômes sur Google et apparemment, j'ai soit un rhume... soit trois mois à vivre. Médecin : Ah oui, le Dr. Google, très compétent celui-là. Vous fumez ? Moi : Non, mais j'ai regardé 5 saisons de "Peaky Blinders"... ça compte ? Médecin : Hum... d'autres symptômes ? Moi : Alors... mal de tête, fatigue, et un truc bizarre : dès que je reçois une facture, j'ai des palpitations. Médecin : Ça, ça s'appelle être pauvre. Moi : Vous pouvez me prescrire du riche ? Médecin : Je peux vous prescrire du Doliprane, c'est déjà bien. Moi : Génial... et sinon, j'ai aussi mal au dos. Médecin : Vous faites du sport ? Moi : Évidemment ! Je cours après mes rêves, je soulève des problèmes et je nage dans la galère. Médecin : ... Je vais plutôt vous prescrire une radio. Moi : Une radio ?! Attendez... c'est pas un moyen détourné pour me dire que je vais mourir ? Médecin : Non... c'est un moyen détourné pour vous faire payer 50 balles de plus. Moi : Ah ouais... la médecine, c'est un peu comme la loterie... sauf que là, on sait d'avance qu'on va perdre. Médecin : Ne vous inquiétez pas. Avec du repos, ça ira mieux. Moi : ("Avec du repos, ça ira mieux"... le mec a fait 10 ans d'études pour me donner le même conseil que ma mère !) Médecin : Voilà votre ordonnance. Moi : C'est écrit "Boire de l'eau et dormir" !? Vous êtes médecin ou influenceur bien-être ?? Médecin : Monsieur... c'est un rhume. Pas une fin de vie. Moi : Ah OK... bon bah... on se revoit la semaine prochaine pour mon mal de ventre.</p>	Très clairement, il commence à être bon et avec plusieurs générations, si on sélectionne la meilleure, on peut avoir quelque chose d'intéressant.	48%

Modèle	Exemple de génération	Commentaire	Score
GPT-4o Thinking	-	Des sketches de même qualité que celui sans raisonnement, mais peut-être un peu mieux.	57%
Mistral Large 2	-	C'était toujours mauvais et long donc je ne l'ai pas mis.	0%

Modèle	Exemple de génération	Commentaire	Score
DeepSeek-R1	<p>Jean : (exagérément enthousiaste) Pierre, tu dois venir à ce cours de yoga avec moi ! C'est bon pour le corps, l'esprit, et en plus, il y a une promo : "Le premier cours, c'est un downward dog offert !"</p> <p>Pierre : (sceptique) Un downward dog offert ? Jean, je ne veux pas de chien, même s'il est à l'envers.</p> <p>Jean : (riant) Non, c'est une posture de yoga ! Tu te mets à quatre pattes, tu lèves les fesses, et tu ressembles à un chien qui s'étire. C'est relaxant !</p> <p>Pierre : (sarcastique) Ah oui, super relaxant. Rien de mieux que de ressembler à un labrador en pleine crise existentielle.</p> <p>Jean : (insistant) Allez, viens ! Le profit que c'est bon pour le dos. Tu te souviens de la fois où tu as essayé de soulever ton chat et où tu t'es coincé pendant trois jours ?</p> <p>Pierre : (défensif) C'était un chat très lourd ! Et puis, il avait mangé tout le thon de la semaine, c'était presque un lion !</p> <p>Jean : (riant) Exactement ! Donc, tu as besoin de yoga. Et en plus, après le cours, il y a des smoothies détox.</p> <p>Pierre : (méfiant) Détox ? Qu'est-ce qu'ils mettent dedans ? De la boue et des espoirs brisés ?</p> <p>Jean : (enthousiaste) Non, c'est plein de vitamines ! Tu verras, après ça, tu te sentiras léger comme une plume.</p> <p>Pierre : (sarcastique) Léger comme une plume ? Moi, après un smoothie, je me sens surtout léger en portefeuille.</p>	Très bon encore.	32%
DeepSeek-R1 Thinking	-	Semblable à celui sans raisonnement.	27%
Gemini 2.0 Flash	-	Sketch semblable aux autres modèles, mais sans l'humour.	2%

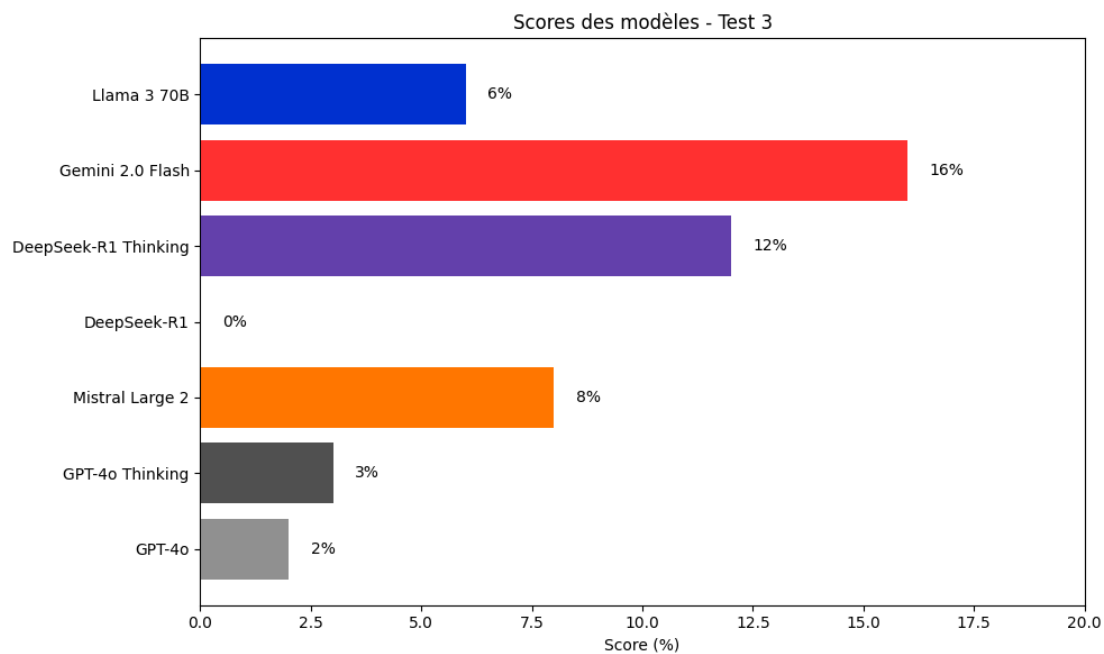
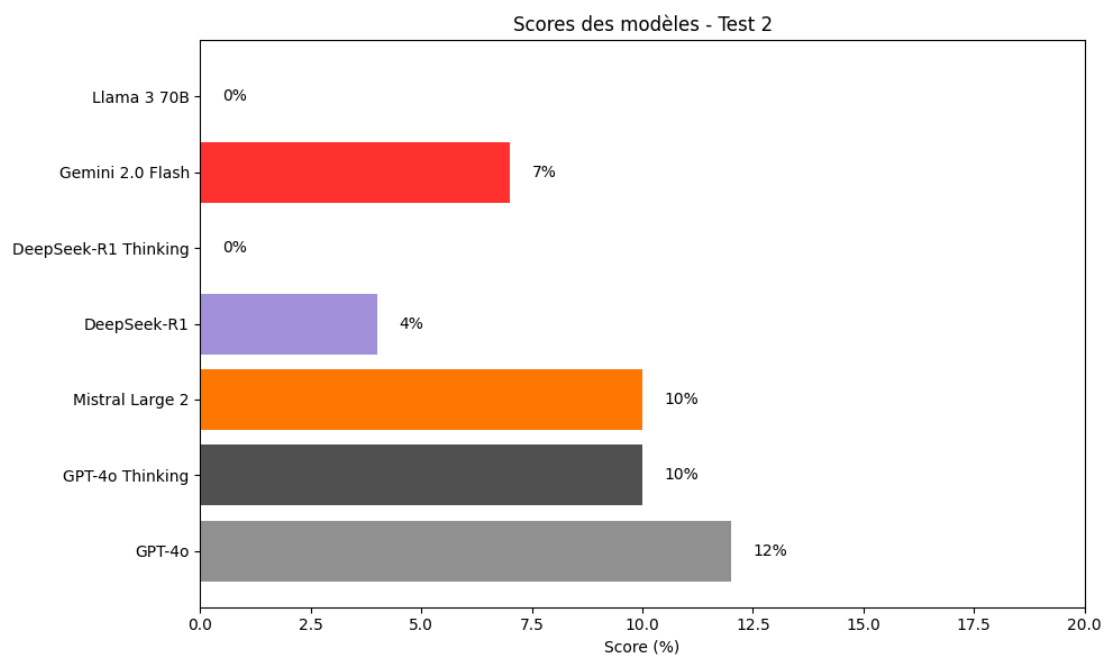
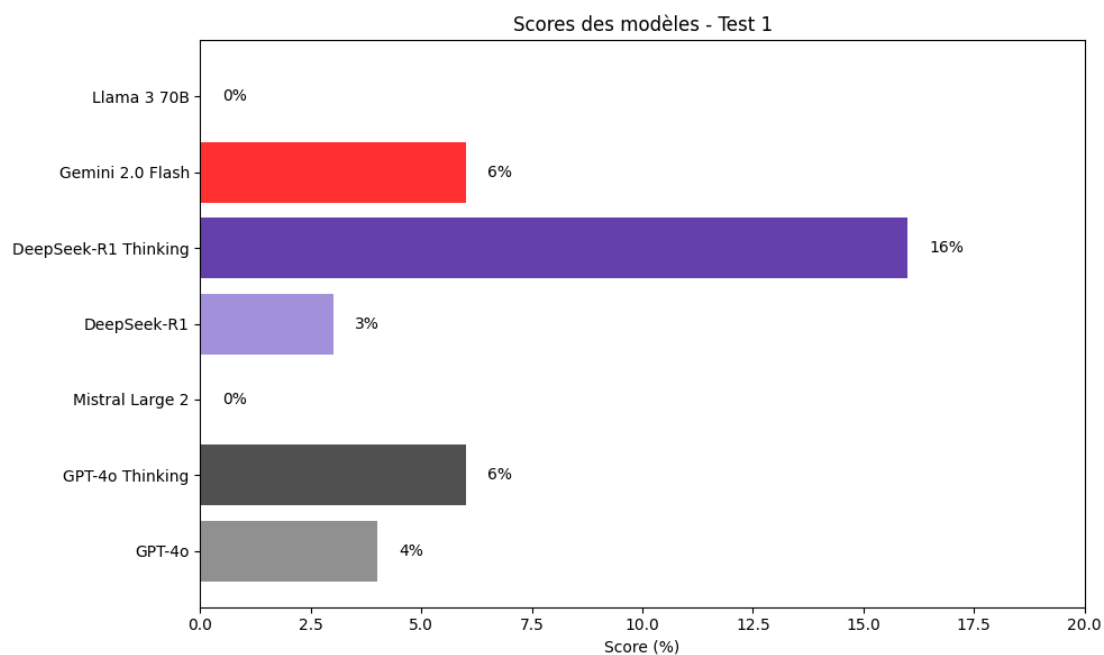
Modèle	Exemple de génération	Commentaire	Score
Llama 3 70B	-	C'était toujours mauvais et long donc je ne l'ai pas mis.	0%

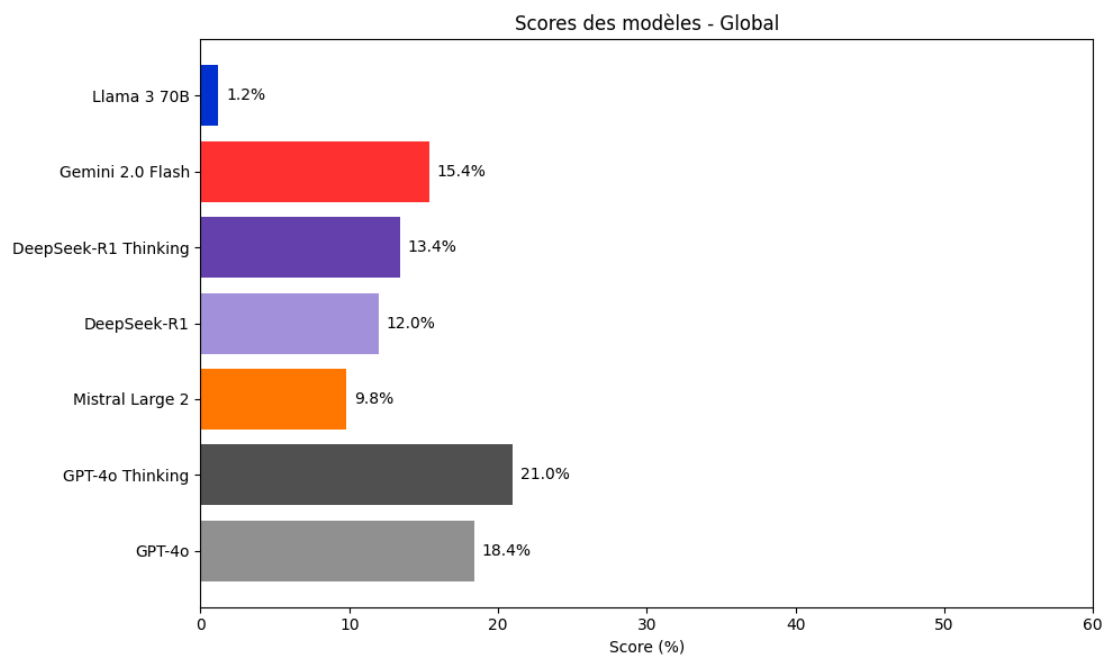
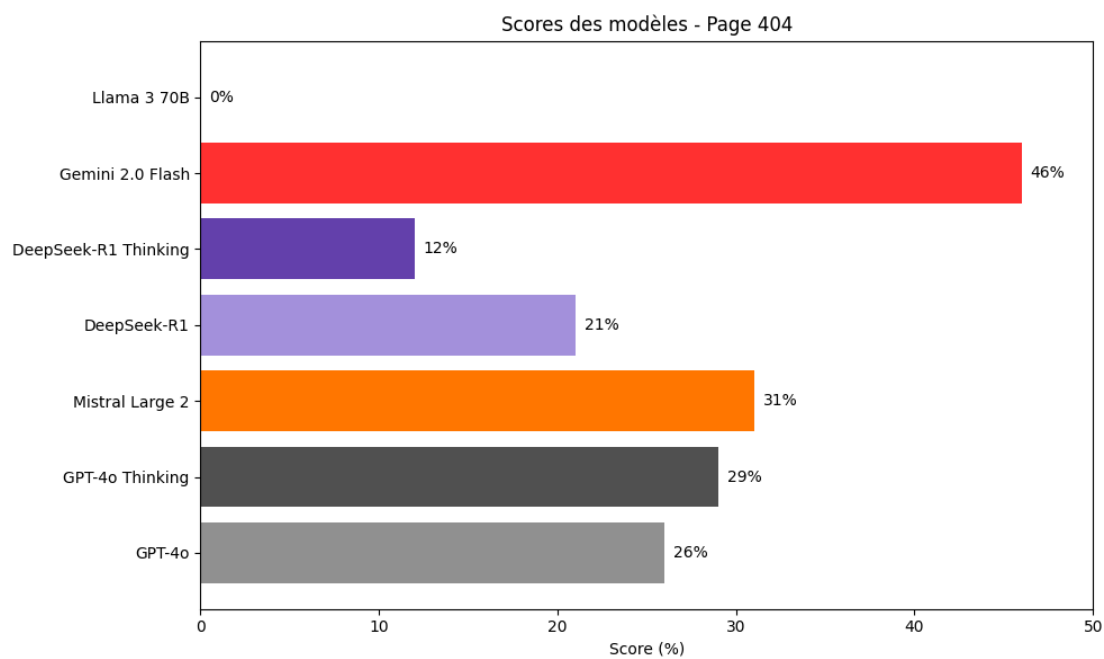
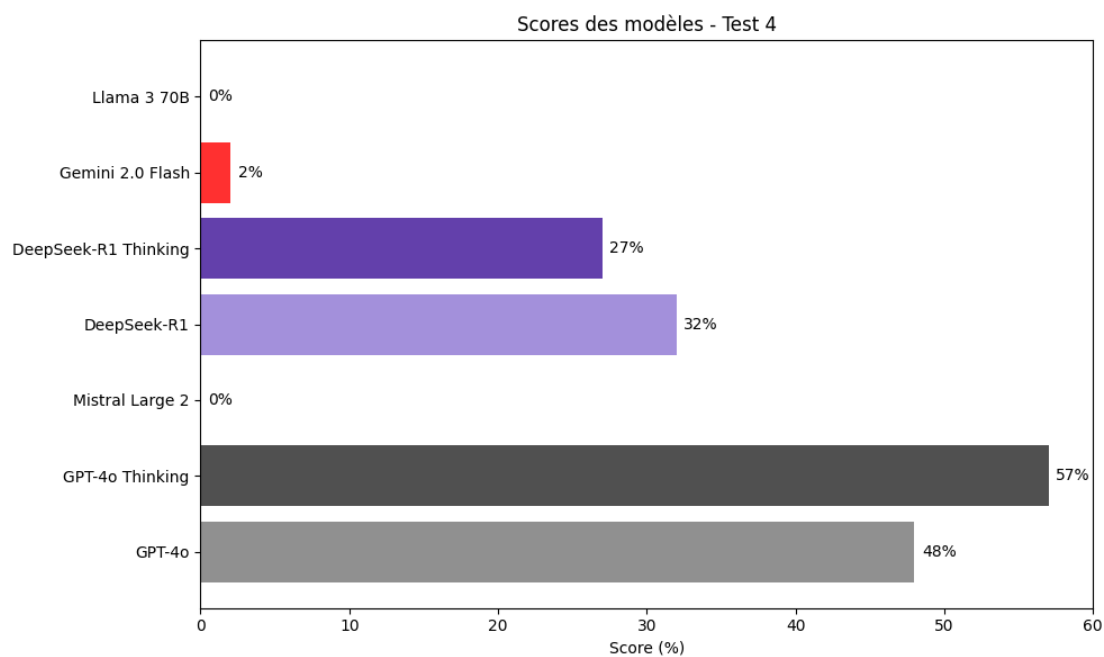
Test : Page 404 humoristique

Ce test consiste à demander aux modèles de générer une page 404 humoristique. L'objectif est de vérifier leur capacité à allier humour et le contexte spécifique d'un site web. La page 404 doit être drôle tout en étant clairement identifiable comme une erreur de page introuvable. Prompt : "Tu es un humoriste vraiment drôle. Génère-moi une idée pour une page 404 pour ce site qui va me faire vraiment beaucoup rire. Le site est [information sur le sujet du site, par exemple Dollarama, une entreprise qui vend des objets à bas prix.]". Au début, les pages n'étaient pas drôles quand je demandais sans gros contexte et aussi si le modèle construisait directement la page, ce n'était pas drôle. Alors, je commence par demander l'idée puis ensuite la page en HTML/CSS."

Modèle	Exemple de génération	Commentaire	Score
GPT-4o	OOPS! TU CHERCHAIS UNE AUBAINE? Malheureusement, cette page est aussi introuvable qu'un rouleau de papier de toilette en rabais pendant une pandémie. Mais ne t'inquiète pas, on a sûrement quelque chose qui te fera oublier cette erreur... pour 1,25\$. Tu veux retourner à la page d'accueil ? Retourner au paradis des deals.	Les pages générées sont parfois bonnes, mais pas souvent. Cette page Dollarama est la meilleure génération. Mention d'honneur à la page générée pour Tim Hortons : "Si vous trouvez la page perdue, on vous offre un sac de cretons gratuit* *non, pas vraiment. Mais on vous aime pareil." J'ai trouvé ça drôle, je ne m'y attendais pas du tout.	26%
GPT-4o Thinking	-	Similaire au modèle sans réflexion, mais parfois un peu mieux.	29%

Modèle	Exemple de génération	Commentaire	Score
Mistral Large 2	Même Sherlock Holmes n'a pas pu trouver cette page. Retournez à l'accueil pour résoudre le mystère des bonnes affaires !	Souvent bon, mais pas toujours en lien avec le sujet. Pour l'exemple, c'est aussi Dollarama que j'ai pris pour que vous puissiez comparer. Même si c'est moins proche du thème, ça fait des choses un peu plus drôles en moyenne que GPT, mais jamais aussi drôle que certaines comme Dollarama de GPT. Souvent, Mistral avait des 3 et 4 sur 10 quand GPT, c'était soit 0 soit 8 ou 9.	31%
DeepSeek-R1	"Oups ! On dirait que vous avez atterri dans le rayon des 404 - Produits Introuvables. Pas de panique ! Notre équipe est en train de chercher votre page perdue entre les coussins du canapé et sous les piles de rabais.	Je mets encore Dollarama pour que vous puissiez comparer les exemples. Mais en général, il a des résultats similaires à GPT.	21%
DeepSeek-R1 Thinking	-	Similaire à sans réflexion.	12%
Gemini 2.0 Flash	Votre page s'est enfuie avec le dernier paquet de bonbons à 1 \$. On la cherche, mais en attendant, vous pouvez toujours vous consoler avec nos autres produits géniaux !	Les résultats sont quand même bons.	46%
Llama 3 70B	-	C'était tout mauvais.	0%





4 Discussion

4.1 Commentaire sur les tests

Le premier test portant sur les blagues simples n'est pas très encourageant quant au potentiel comique des LLM. Par contre, dans ce test, la chaîne de pensée semble avoir beaucoup aidé dans le cas de DeepSeek-R1, peut-être parce qu'elle lui permet d'évaluer lui-même les meilleures blagues avant de les présenter. Cela pose la question de savoir si les modèles de langage sont capables de déterminer quelles blagues sont les meilleures et de juger de leur côté drôle.

Le deuxième test consistait à indiquer au modèle que l'humour était son domaine de prédilection, car nous avons observé que de nombreux LLM obtiennent de meilleures performances lorsqu'ils se mettent dans la peau d'un certain métier ou personnage. Dans notre cas, le manque de liberté du one-liner semble trop le bloquer pour qu'il soit comique.

Le troisième test a consisté à autoriser le modèle à créer un contexte pour préparer sa blague, ce qui a amélioré les réponses. Il est normal que le résultat soit meilleur, car un one-liner est très difficile à réussir et il est rare que des humoristes se contentent uniquement de ce format.

Pour augmenter les possibilités de création et d'imagination, nous avons ensuite demandé de créer un sketch humoristique afin que le modèle puisse mieux préparer les chutes et faire des références aux éléments précédemment évoqués. Certains modèles, comme Gemini ou Mistral, ont complètement raté ce test en produisant de petites histoires peu drôles et peu intéressantes. En revanche, d'autres modèles, comme GPT-4o, ont montré qu'ils possédaient un énorme potentiel humoristique lorsqu'on leur laissait plus de liberté créative. Nous avons obtenu d'excellents résultats, sans avoir à apporter de modifications au texte, ce qui est prometteur pour aider les humoristes dans l'écriture de spectacles, car, guidés par l'humoriste, on peut s'attendre à des sketches encore plus drôles. Il ne faut pas oublier que l'humour repose également sur l'intonation, les pauses, les expressions faciales et la gestuelle. Faire rire uniquement avec du texte est déjà un bon début.

Finalement, le dernier test portait sur les pages 404. Ces pages constituent un espace encore trop peu exploité pour l'humour. De nombreux sites tentent de proposer des pages 404 originales, ce qui est appréciable, mais force est de constater que la majorité reste banale. J'invite donc les lecteurs de cet article à intégrer davantage d'humour dans leurs pages 404 et, si nécessaire, à se faire accompagner par des modèles de langage (LLM) qui, en offrant plusieurs générations, permettent de sélectionner des propositions de qualité. Par ailleurs, il est intéressant de noter que les tests portant sur les pages 404 sont généralement mieux évalués par les participants. Même si l'effet comique n'est pas toujours aussi marqué que pour d'autres types de contenus, une bonne adéquation entre le thème de la page et son contenu entraîne l'attribution de notes élevées.

4.2 Réflexion sur les llms pour l’humour

Nous remarquons dans nos résultats que les modèles ont tendance à reproduire des thèmes et des schémas récurrents. Des sujets comme les rendez-vous chez le médecin ou le yoga reviennent fréquemment dans les sorties générées. Cependant, les humains font également souvent des sketches sur les couples. Cette répétition indique également une difficulté pour les modèles à innover limitant ainsi l’originalité de l’humour produit.

L’introduction du raisonnement en chaîne a permis d’obtenir une légère amélioration de la qualité humoristique des réponses générées. Bien que cette approche affine légèrement la construction des réponses, le coût computationnel associé semble disproportionné par rapport aux bénéfices obtenus.

L’humour repose en grande partie sur l’élément de surprise. Or, bien que les modèles soient probabilistes, ils tendent à revenir sur des thématiques familières, ce qui limite l’effet inattendu. Des ajustements, tels que l’augmentation de la température de génération visant à favoriser une plus grande diversité dans les réponses, pourraient théoriquement aider. Cependant, la frontière entre l’inattendu comique et le non-sens déconnecté demeure très mince. Par conséquent, la simple manipulation des paramètres de génération ne suffit pas à garantir un humour véritablement surprenant. Ces constats mériteraient d’être approfondis par des recherches expérimentales puisque dans le présent travail, ils reposent principalement sur des inférences tirées de la littérature sur le domaine. Cependant, une plus grande intervention humaine et une guidance sur les thèmes à choisir pourraient probablement aider à obtenir une plus grande variété de blagues.

Il convient également de rappeler que les modèles de langage testés ne sont pas entraînés explicitement pour produire de l’humour, même si leur corpus d’entraînement contient de nombreuses blagues. Par ailleurs, à mesure que la taille des modèles augmente, on observe l’émergence de multiples capacités, y compris des compétences humoristiques. Ainsi, un atout majeur de ces modèles est qu’ils disposent de connaissances dans de nombreux domaines, ce qui leur permet de s’adapter plus facilement aux différentes cultures et de faire référence à un éventail plus large de sujets. De nombreux sketches humoristiques tirent parti du fait que le monde a déjà vécu certains événements pour créer un effet comique, et ces modèles semblent avoir tout vécu puisqu’ils en ont presque tous les histoires de l’humanité.

Bien que le modèle Grok de X anciennement Twitter n’ait pas pu être évalué dans le cadre de cette étude, des recherches préliminaires suggèrent qu’il pourrait offrir de meilleures performances en matière d’humour, possiblement en raison d’un niveau de censure moins restrictif. Cette observation ouvre la voie à une réflexion sur l’impact des mécanismes d’alignement des modèles. Un modèle fine-tuné spécifiquement pour la production d’humour pourrait, en effet, générer des réponses plus pertinentes et originales, car le fine-tuning permet de personnaliser la génération et aussi l’adaptation aux public ciblé.

5 Conclusion

Les résultats de cette étude montrent que, malgré un potentiel indéniable, l'utilisation des LLM pour la génération d'humour demeure un défi complexe. Certains modèles, notamment GPT-4o, parviennent à produire des contenus humoristiques intéressants, surtout lorsqu'ils disposent d'une plus grande liberté créative, comme en témoignent le test 4. L'introduction du raisonnement en chaîne (Chain of Thought) apporte des améliorations, mais son coût computationnel semble disproportionné par rapport aux bénéfices obtenus. Ces constats suggèrent que, pour exploiter pleinement le potentiel humoristique des LLM, il serait judicieux d'explorer des pistes complémentaires, telles que le fine-tuning sur des corpus dédiés à l'humour.

En somme, bien que les LLM puissent générer de l'humour, ils rencontrent encore des limites importantes, notamment en matière d'originalité. Néanmoins, leur capacité à s'adapter à divers contextes et à proposer une variété de contenus humoristiques laisse entrevoir des applications intéressantes, notamment en tant qu'outils d'aide à la création des humoristes. Une approche combinant intervention humaine et fine-tuning permettrait probablement d'exploiter plus efficacement leur potentiel comique.

Références

- [1] Benoit Baudry et al. “Generative AI to Generate Test Data Generators”. In: *IEEE Software* 41.6 (Nov. 2024), 55–64. ISSN: 1937-4194. DOI: 10.1109/ms.2024.3418570. URL: <http://dx.doi.org/10.1109/MS.2024.3418570>.
- [2] Eric Hartford. *Uncensored Models* — [erichartford.com](https://erichartford.com/uncensored-models). <https://erichartford.com/uncensored-models>. [Accessed 08-02-2025].
- [3] *How funny is ChatGPT? A comparison of human- and A.I.-produced jokes* — [pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11221738/). <https://pmc.ncbi.nlm.nih.gov/articles/PMC11221738/>.
- [4] *Humor Detection and Appreciation: Human versus LLM performance*. <https://osf.io/yzupx>.
- [5] *Humor Understanding Multi-task Optimization & Ranking* — [huggingface.co](https://huggingface.co/blog/TuringsSolutions/humortest). <https://huggingface.co/blog/TuringsSolutions/humortest>.
- [6] Greg Robison. *Exploring the intersection of humor and LLMs to measure AI’s progression* — [finn-group.com](https://www.finn-group.com/post/the-last-laugh-exploring-the-role-of-humor-as-a-benchmark-for-large-language-models). <https://www.finn-group.com/post/the-last-laugh-exploring-the-role-of-humor-as-a-benchmark-for-large-language-models>. [Accessed 08-02-2025].
- [7] Han Wang et al. *Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps* — [arxiv.org](https://arxiv.org/html/2410.10370v1). <https://arxiv.org/html/2410.10370v1>. [Accessed 08-02-2025]. 2024.
- [8] Jason Wei et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL]. URL: <https://arxiv.org/abs/2206.07682>.