

Évaluation du potentiel des Grands Modèles de Langage pour le domaine 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, Grok 3, Claude 3.5 et Llama 3 70B. Nous analysons également l’impact du raisonnement en chaîne (Chain of Thought) sur la qualité de l’humour produit. Ensuite, nous nous demanderons si les modèles de langage fondamentaux sont des outils pertinents pour évaluer la qualité des blagues. Enfin, nous proposerons des pistes de recherche futures ainsi qu’un compte rendu des performances des modèles fondamentaux dans le domaine de l’humour.

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.

Cette étude se concentre spécifiquement sur l’aptitude des LLM à générer de l’humour. Nous comparons les performances de plusieurs LLM fondamentaux de premier plan (Mistral Large 2, GPT-4o, DeepSeek-R1, Gemini 2.0 Flash, Grok 3, Claude 3.5, Llama 3 70B) sur différentes tâches humoristiques, en examinant également l’effet du raisonnement en chaîne. Une autre question est d’évaluer si ces modèles peuvent eux-mêmes juger de manière fiable la qualité humoristique. Ce travail vise à cartographier les capacités actuelles et les limites des LLM dans le domaine de l’humour.

Dans cette étude, nous avons choisi de nous concentrer sur les modèles fondamentaux (foundation models), c’est-à-dire les grands modèles de langage généralistes, entraînés sur une large diversité de corpus. Bien qu’il existe des modèles plus petits, parfois spécifiquement fine-tunés pour des tâches humoristiques, nous avons constaté qu’ils sont nettement moins performants que les grands modèles. Tandis que les modèles fondamentaux, grâce à leur taille et aux phénomènes d’émergence, offrent des possibilités beaucoup plus prometteuses que les petits modèles que nous avons pu tester. En plus, ces modèles sont souvent facilement accessibles via des API ou des interfaces mises à disposition par les grandes entreprises qui les développent, ce qui permet de les tester sans avoir besoin d’un matériel particulier.

Première expérience

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, Grok 3, Claude 3.5 Haiku. 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 d’une quinzaine de participants, mais qui n’étaient pas présents pour évaluer toutes les parties. En réalité, il y avait souvent entre 3 et 5 personnes, principalement des étudiants en informatique à l’Université de Montréal, parlant français et originaires du Québec ou de la France, ainsi que quelques membres de ma famille pour d’autres tests. Chaque participant attribuait une note de 0 à 10, et la moyenne de ces notes était utilisée pour déterminer la note finale. On répétait cela 10 fois avec des blagues différentes et ce, pour chaque LLM. En additionnant les notes des 10 fois, on obtenait une valeur déjà sur 100, et c’était la note attribuée au LLM. 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.

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. Notez que l’exemple choisi est la meilleure des générations pour montrer le potentiel du modèle et n’est donc pas représentatif de toutes les générations. Le pourcentage indiqué donne une idée des chances qu’une blague comme celle-ci se produise, mais il s’agit réellement d’une appréciation globale des générations du modèle. De plus, des commentaires seront ajoutés lorsque cela est pertinent. Enfin, une note sera attribuée.

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%
Grok 3	Mon miroir m'a dit que j'étais beau, mais il réfléchit mal.	Du bon potentiel pour les blagues, mais l'exécution est mauvaise. Il y a aussi beaucoup de blagues non originales parmi celles qu'il fait.	8%
Claude 3.5 Haiku	J'ai demandé à mon GPS de me trouver un sens à ma vie, il m'a répondu : "Recalcul en cours..."	Les blagues n'étaient pas très drôles, mais elles étaient plus originales que celles des autres modèles. On remarque cependant que le thème de la procrastination et de la motivation revient souvent dans tous les modèles, et souvent, ce n'est pas drôle.	14%

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%
Grok 3	Ma copine m'a dit qu'elle voulait un homme qui la fasse rire, alors je lui ai montré mon compte en banque.	Beaucoup de bons débuts, mais rarement drôles. Normalement, jusqu'à présent, les modèles font beaucoup de mauvaises blagues et une bonne, mais celui-ci a fait plein de blagues moyennes.	12 %
Claude 3.5 Haiku	Mon GPS et moi, on a le même problème : on sait exactement où on veut aller, mais on se perd toujours en route.	Vraiment mieux construites et plus dans le style d'un humoriste en spectacle.	24%

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%
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%

Modèle	Exemple de génération	Commentaire	Score
Grok 3	Pourquoi les vaches ferment-elles les yeux pendant la traite ? Pour faire du lait concentré !	Pas beaucoup de bonnes blagues et beaucoup de déjà-vu par plusieurs autres modèles.	16%
Claude 3.5 Haiku	Un patient entre chez le médecin et dit : "Docteur, j'ai un problème. À chaque fois que je bois du café, j'ai mal à l'œil." Le médecin lui répond : "Et bien, essayez d'enlever la cuillère de la tasse avant de boire !"	Vraiment beaucoup de bonnes blagues ou, sinon, de bons débuts.	34%

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%

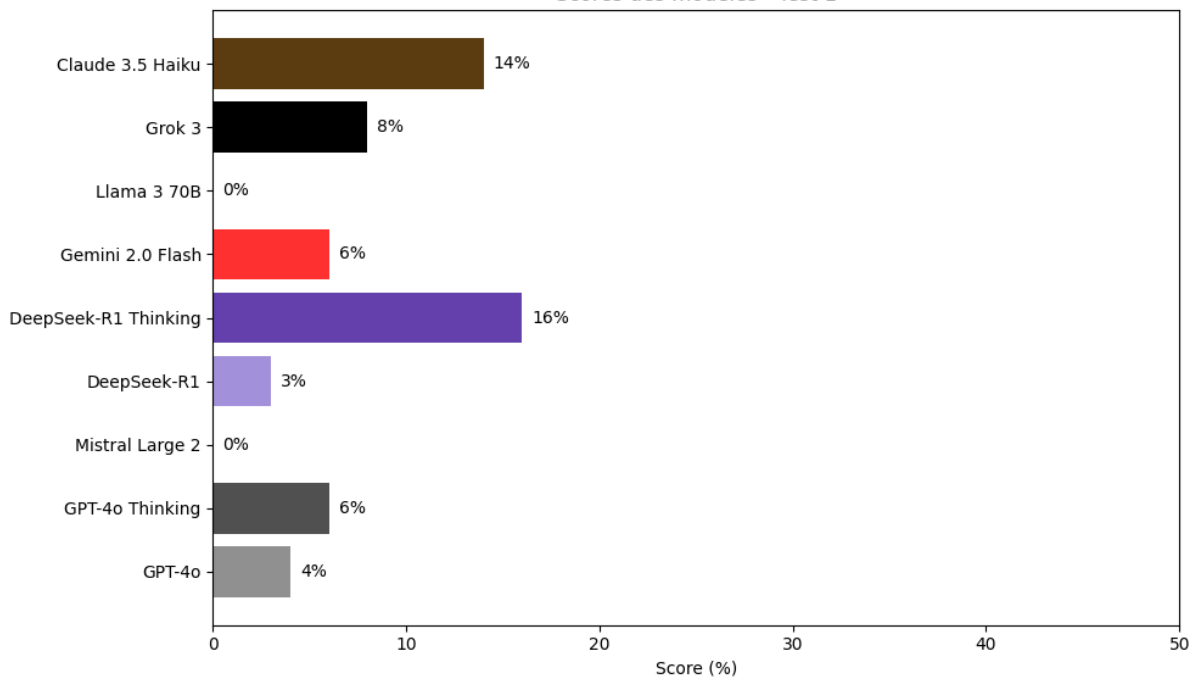
Modèle	Exemple de génération	Commentaire	Score
GPT-4o Thinking	-	Des sketches de même qualité que celui sans raisonnement, mais peut-être un peu mieux.	57%
Mistral Large 2	-	C'était toujours mauvais et long donc je ne l'ai pas mis.	0%
DeepSeek-R1	<p>Jean : (exagérément enthousiaste) Pierre, tu dois venir 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%

Modèle	Exemple de génération	Commentaire	Score
Gemini 2.0 Flash	-	Sketch semblable aux autres modèles, mais sans l'humour.	2%
Llama 3 70B	-	C'était toujours mauvais et long donc je ne l'ai pas mis.	0%
Grok 3		Pas vraiment drôle.	4%
Claude 3.5 Haiku		Sketch avec presque pas de blagues, bizarre, car il était le meilleur jusqu'à présent pour faire des blagues.	14%

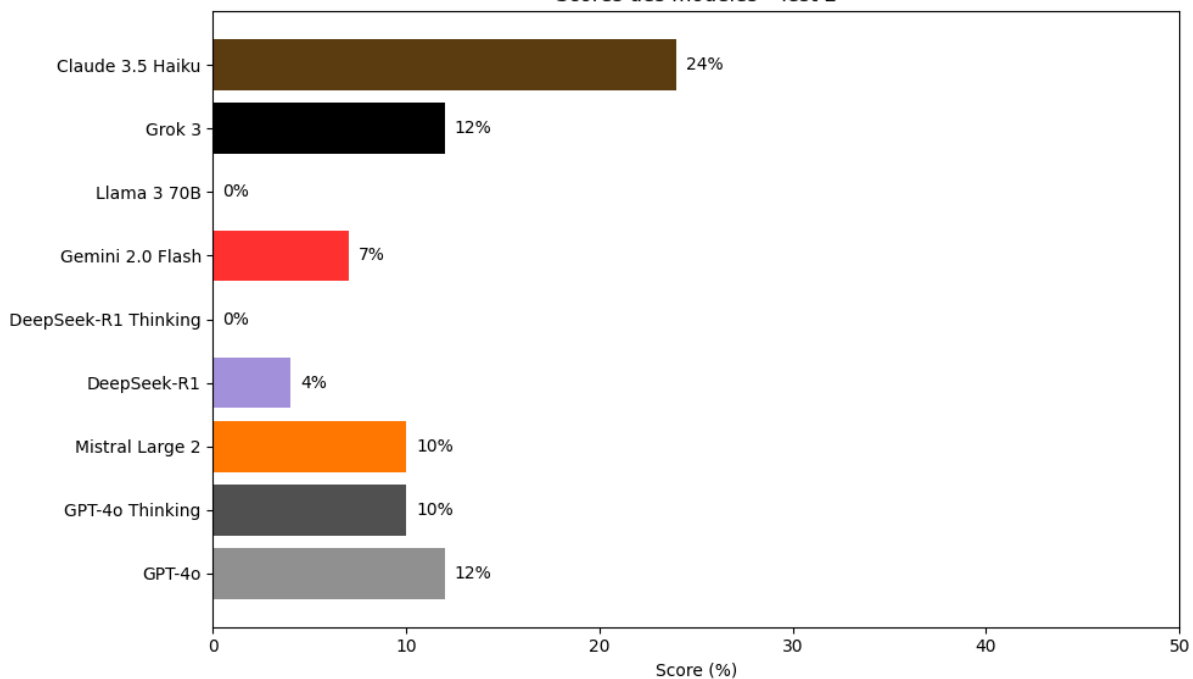
Test : Page 404 humoristique			
<p>Ce test consiste à demander aux modèles de générer une page 404 humoristique. L'objectif est de vérifier leur capacité à allier humour et le contexte spécifique d'un site web. La page 404 doit être drôle tout en étant clairement identifiable comme une erreur de page 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."</p>			
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%
Grok 3		Pas drôle.	0%
Claude 3.5 Haiku		Pas très drôle, mais très bon comparé aux autres pour générer directement le HTML/CSS.	12%

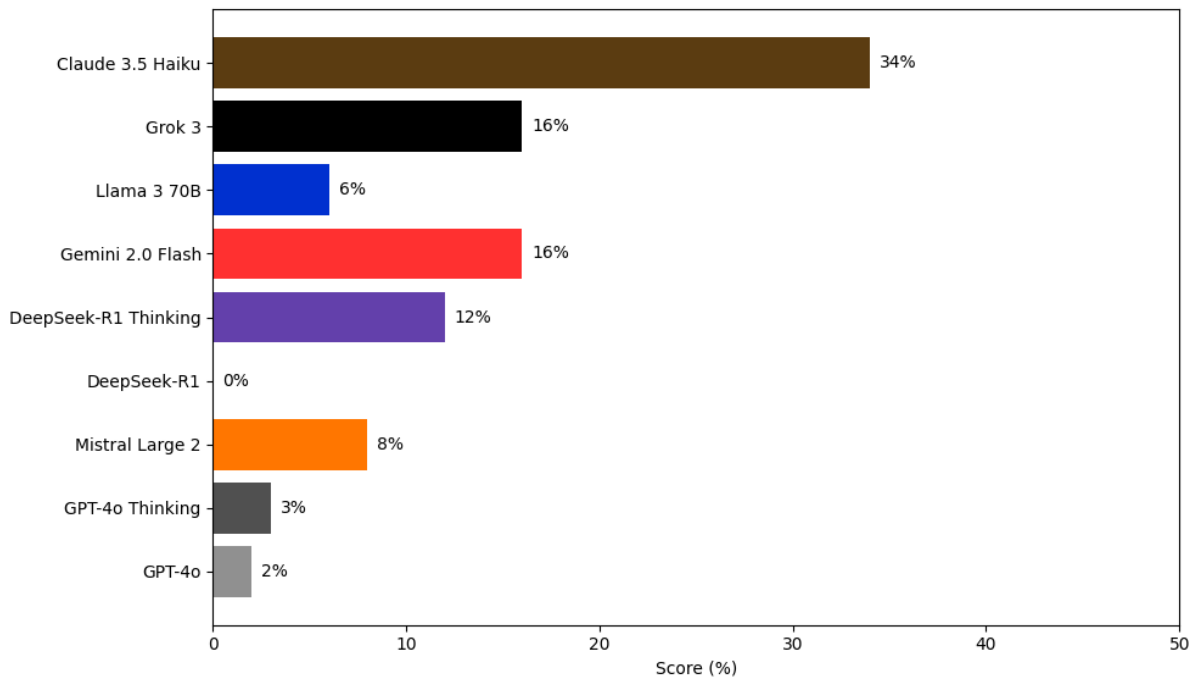
Scores des modèles - Test 1



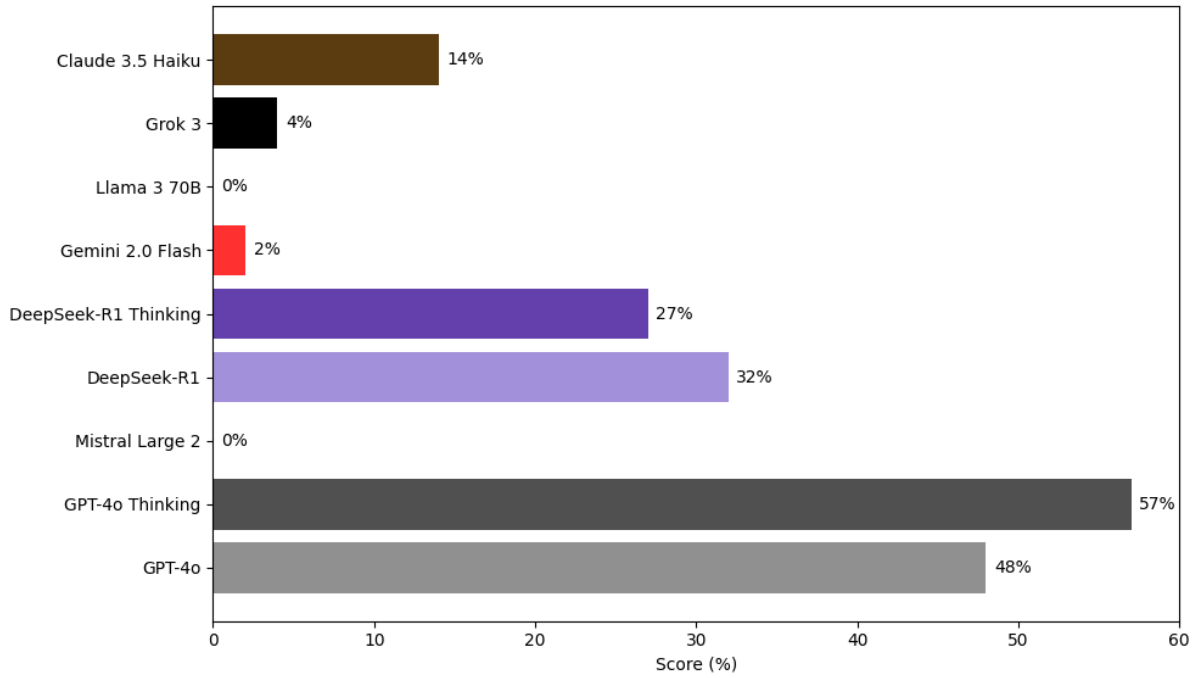
Scores des modèles - Test 2



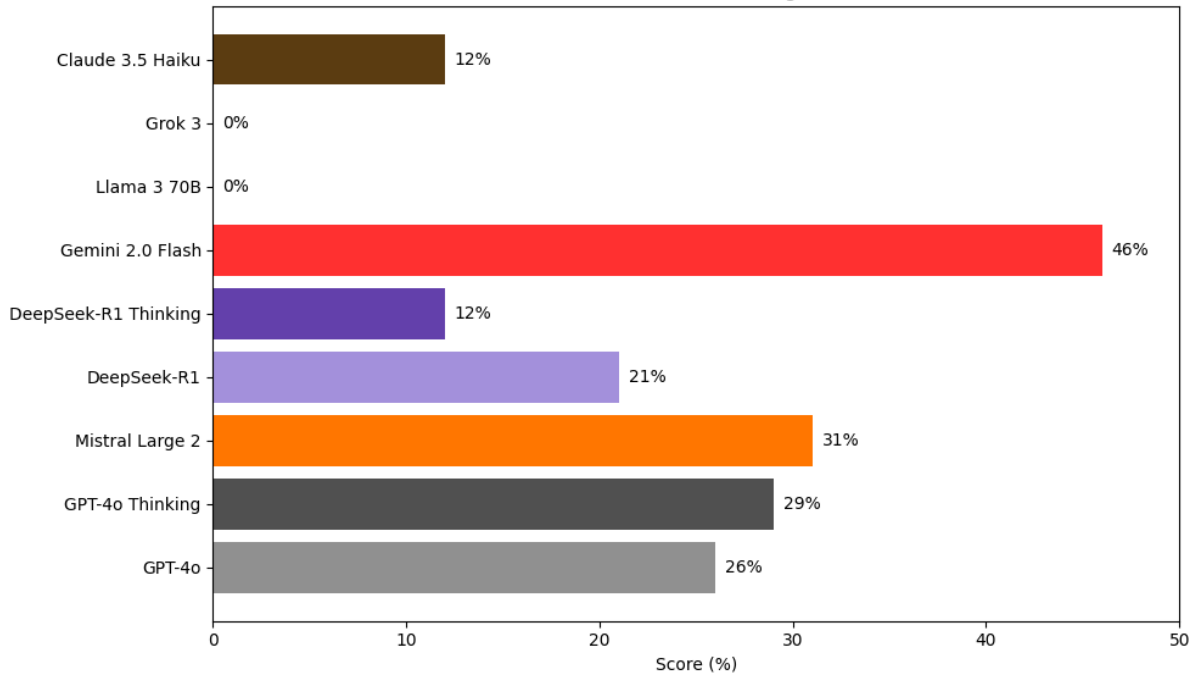
Scores des modèles - Test 3



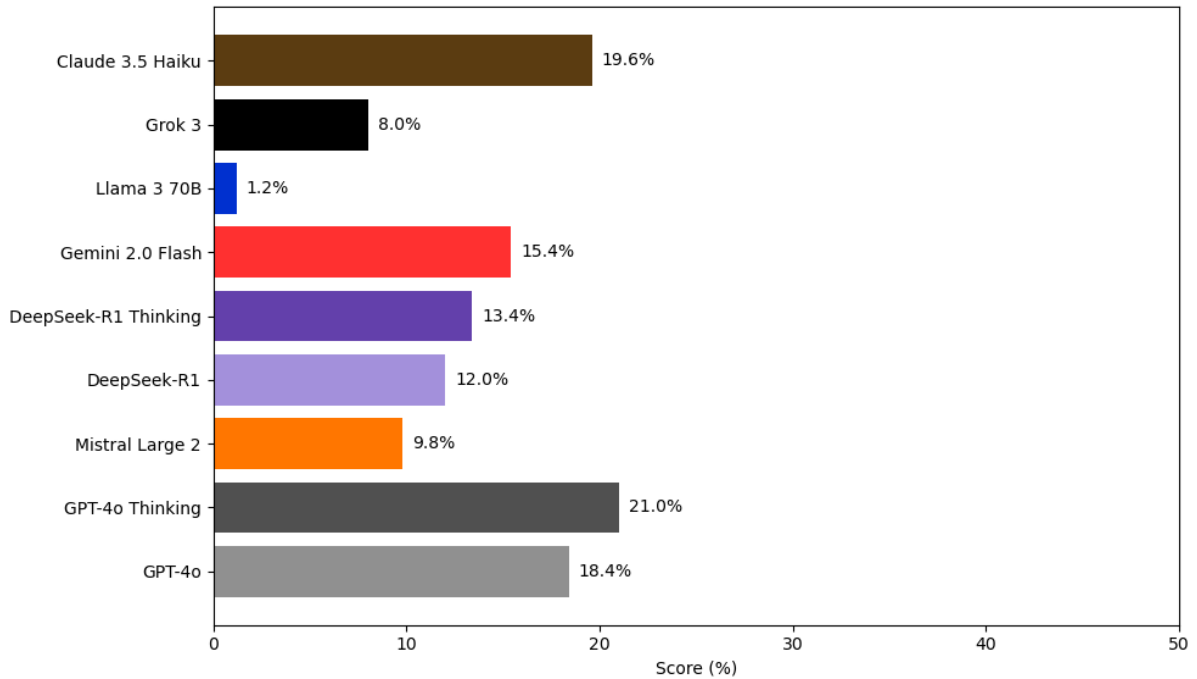
Scores des modèles - Test 4



Scores des modèles - Page 404



Scores des modèles - Global



Discussion

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 ce qui les rend drôles. Nous répondrons à cette question dans la discussion de la deuxième expérience.

Le deuxième test consistait à indiquer au modèle que l'humour était son domaine de prédilection, car plusieurs études observent que de nombreux LLM obtiennent de meilleures performances lorsqu'ils se mettent dans la peau d'un certain métier ou personnage, une technique qu'on appelle *role-prompting*. [16] Dans notre cas, le manque de liberté du one-liner semble trop l'entraver 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.

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. Nous n'avons pas pu tester l'originalité des blagues présentées, mais selon l'étude "You Told Me That Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models" [1], les grands modèles de langage comme GPT ne semblent pas vraiment mémoriser les blagues. Cependant, leur étude a été menée uniquement en anglais, donc il serait intéressant de se

pencher sur ce sujet dans d'autres langues. Il est surprenant qu'ils ne mémorisent pas beaucoup, car de gros modèles donnaient souvent des réponses similaires, et d'autres articles évaluant les capacités des modèles en mathématiques semblent dire que l'amélioration de GPT-4 pour les mathématiques est en grande partie due à la mémorisation. Aussi, dans cette étude sur GPT 3.5, ils disent : "Dans 1 008 essais au cours desquels ChatGPT a été invité à raconter une blague, plus de 90% du temps, la réponse était une répétition de l'une des 25 blagues différentes. Les quatre meilleures blagues ont été recyclées dans plus de la moitié des réponses." [10]. Il faudrait donc vraiment se pencher davantage sur ce point, car plusieurs études disent des choses différentes. Mais selon ce que j'ai vu jusqu'à présent, il est vrai que les blagues reviennent souvent et même entre les différents modèles.

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. De plus, une étude récente menée par Anthropic [13] suggère que les modèles possèdent souvent déjà la réponse dès le départ et que le raisonnement explicite en plusieurs étapes sert uniquement à faciliter la compréhension humaine plutôt qu'à améliorer réellement la réponse. Le modèle sait déjà ce qu'il va dire et construit une série de raisonnements logiques pour arriver à cette réponse.

Deuxième expérience

Méthodologie

Pour améliorer notre étude, nous avons ensuite choisi parmi ceux testés trois modèles qui semblaient les mieux performer pour une étude avec beaucoup plus de participants en ligne. Les trois modèles testés sont GPT-4o, DeepSeek-R1 et Gemini 2.0 Flash. Au total, nous avons eu 89 répondants. Claude 3.5 Haiku avait bien performé, mais nous n'avons pas pu le tester pour cette expérience à grande échelle.

Pour la deuxième expérience, nous avons généré un corpus de blagues à l'aide de modèles de langage (LLM) sur 10 thèmes populaires : relations de travail, enfants, différences culturelles, voyage, situations quotidiennes, relations de famille, sport, animaux de compagnie, différences homme-femme, et relations de couple. Chaque LLM (ChatGPT-O1, DeepSeek R1 et Gemini-Flash 2.0) a généré 30 blagues par thème, soit un total de 900 blagues. Ces blagues ont ensuite été évaluées automatiquement par le modèle, et seules les 10 meilleures blagues (selon eux) par thème ont été retenues, ce qui nous donne un corpus final de 300 blagues. Nous avons choisi de répéter le prompt du test 3 de l'expérience 1 pour éviter des blagues trop longues, mais avoir des blagues drôles. Voici le prompt utilisé : "Tu es un humoriste vraiment drôle. Génère-moi 30 blagues sur le thème '[insérer le thème]', mais prends ton temps pour la préparer et la rendre encore plus drôle. Je veux que ce soit hilarant !".

Après la génération, nous avons demandé aux modèles d'évaluer chaque blague en leur assignant une note de 0 à 10. Cette méthodologie sera discutée plus en détail dans notre analyse des résultats afin de déterminer si les LLM évaluent bien l'humour généré. Une observation préliminaire est que la qualité des blagues semble inférieure à celle obtenue dans une étude précédente. Une hypothèse est que la génération de plusieurs blagues à la fois réduit la qualité par rapport à une génération individuelle. Il serait intéressant d'évaluer cette hypothèse.

Pour évaluer les blagues, les participants ont été invités à noter chaque blague en utilisant une échelle de trois émoticônes accompagnées de leur texte associé : "Pas drôle", "Légèrement drôle" et "Drôle". Nous avons ensuite converti ces évaluations en scores numériques : 0 pour "Pas drôle", 1 pour "Légèrement drôle" et 2 pour "Drôle". De plus, l'ordre des trois émoticônes était aléatoire pour chaque participant afin d'éliminer les biais pouvant être dus à l'ordre de présentation.

Comme deuxième test concernant l'évaluation des modèles de langage, nous avons par la suite demandé au LLM d'évaluer plusieurs blagues à plusieurs reprises pour vérifier si les LLMs étaient constants dans leur évaluation.

Résultats

Voici le tableau des résultats des évaluations des blagues. Le premier tableau montre le nombre de votes pour Pas drôle, Légèrement drôle et Drôle, ainsi que le score moyen sur 100. Ce score est calculé en attribuant 0 à "Pas drôle", 50 à "Légèrement drôle", et 100 à "Drôle". Bien que cette approximation soit assez simpliste, elle permet d'avoir une meilleure vue d'ensemble des résultats, en les exprimant sous une forme plus familière, les pourcentages. Le deuxième tableau présente les notes attribuées par les LLM aux blagues, ainsi que les notes réellement accordées par les participants à ces mêmes blagues.

LLM	Pas drôle	Légèrement drôle	Drôle	Nombre d'évaluations	Score moyen
DeepSeek-R1	135	38	12	185	17%
Gemini Flash 2.0	113	46	14	173	21%
GPT-4o	131	37	13	181	17%

Table 6: Nombre de votes pour chaque catégorie et score moyen par LLM.

LLM	Note LLM	Pas drôle	Légèrement drôle	Drôle	Score moyen
DeepSeek-R1	9	70	12	5	12%
	10	65	26	7	20%
Gemini Flash 2.0	7	17	4	1	13%
	8	92	39	11	21%
	9	4	3	2	39%
GPT-4o	7	15	2	0	6%
	8	93	22	6	14%
	9	20	11	7	33%
	10	3	2	0	20%

Table 7: Répartition des votes en fonction des notes attribuées par les LLM.

Discussion

Les résultats obtenus montrent que lorsque l'expérience est menée avec un plus grand nombre de participants, les performances des modèles de langage fondamentaux en matière d'humour sont décevantes. Cela suggère que, bien que ces modèles soient efficaces pour reproduire des structures linguistiques probables, ils n'ont pas une véritable compréhension de l'humour. Il est probable qu'ils se contentent de générer du texte ayant l'apparence de l'humour sans en saisir l'essence réelle.

Un autre aspect de cette étude était d'évaluer si ces modèles sont capables de juger la qualité des blagues. Cela permet de mieux comprendre s'ils sont réellement capables de déterminer si ce qu'ils produisent est drôle ou non. Nos résultats montrent que les modèles de langage fondamentaux ont tendance à attribuer des scores situés entre 7 et 9 aux blagues, même lorsque les évaluations humaines révèlent une qualité bien moindre. Cependant, les humains eux aussi ont tendance à accorder ces valeurs de manière fréquente. Il est donc probable que les modèles aient appris ce biais à partir des données d'entraînement. Toutefois, nous observons que l'ordre relatif des évaluations reste cohérent. Par exemple, DeepSeek attribue un score de 9 aux blagues qui obtiennent un score moyen de 12 % selon les évaluateurs humains et un score de 10 aux blagues ayant un score moyen de 20 %. Des tendances similaires sont observées pour Gemini (7 : 13 %, 8 : 21 %, 9 : 39 %) et GPT (6 : 6 %, 8 : 14 %, 9 : 34 %, 10 : 20 %). Cela suggère que, bien que les modèles de langage surestiment la qualité des blagues, ils parviennent à maintenir une hiérarchie correcte dans leurs évaluations.

Ces observations soulèvent des questions sur l'utilisation des modèles de langage pour évaluer d'autres modèles. En effet, même si l'ordre des évaluations semble fiable, les scores absolus sont largement surestimés, ce qui peut biaiser l'interprétation des résultats. Cette question est cruciale dans l'évaluation des performances des modèles de génération de texte et mériterait une analyse plus approfondie, d'autant plus que l'évaluation des modèles de langage par d'autres modèles est une pratique de plus en plus courante, et pas seulement pour l'humour. Il serait pertinent de mener une étude plus précise à ce sujet. Nos observations montrent que les LLM tendent à générer des blagues en respectant des codes humoristiques formels plutôt qu'en évaluant réellement à quel point elles sont comiques. De plus, lorsque nous testons des blagues humaines reconnues comme drôles mais qui ne sont pas sous forme de "one-liner", les LLM leur attribuent des notes bien plus basses qu'elles ne le mériteraient. Cela renforce l'idée que les LLM sont mauvais pour évaluer l'humour et se concentrent seulement sur la forme plutôt que sur la blague.

Une étude intitulée "LLMEval: A Preliminary Study on How to Evaluate Large Language Models" [27] examine différents aspects de l'évaluation des grands modèles linguistiques. Elle compare l'évaluation humaine à l'évaluation automatique en utilisant GPT-4 et en tire plusieurs conclusions intéressantes. Notamment, elle met en évidence que, bien que les évaluations des LLM soient souvent alignées avec celles des humains, de nombreux biais persistent. Par exemple, GPT-4 tend à surévaluer les réponses longues, même si elles ne sont pas de meilleure qualité. Il n'est donc pas surprenant que, dans notre étude, les blagues soient classées selon leur forme et leur style d'écriture plutôt que selon leur humour réel. Cette étude souligne aussi que pour des critères subjectifs comme l'humour, les résultats obtenus par les LLM manquent de cohérence. En effet, lorsque nous évaluons plusieurs fois les mêmes blagues avec le même modèle, les scores varient à chaque itération. Peut-être est-ce notre façon de demander l'évaluation qui pose problème. Dans l'article, il est indiqué que l'évaluation par étoiles est meilleure qu'une évaluation avec une échelle de 0 à 10, comme celle que nous avons utilisée. L'utilisation des LLM comme évaluateurs semble donc problématique. Certes, l'évaluation

humaine elle-même peut manquer de constance, mais il est nécessaire de trouver une méthode plus fiable pour juger de la qualité des blagues.

Par ailleurs, une étude intitulée "Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges" [3] propose une approche qui semble donner de bons résultats. Il serait intéressant d'explorer plus en profondeur cette méthode afin de déterminer si elle constitue une solution viable. De ce que j'ai pu lire dans différents articles, il semble que l'évaluation d'un petit modèle par un plus grand fonctionne bien lorsque le critère est objectif, mais pour des critères subjectifs comme l'humour, cela devient beaucoup plus problématique. Comme l'évaluation de l'humour par des humains est déjà complexe, il est crucial d'approfondir la recherche sur ces méthodes et de rester vigilant quant à l'utilisation des LLM pour évaluer d'autres modèles.

Si l'évaluation des modèles par d'autres modèles s'avère fiable, il serait intéressant d'appliquer ces méthodes à des technologies comme Crew IA pour la génération humoristique. La qualité des résultats obtenus dépendrait alors fortement de la capacité des LLM à bien évaluer le contenu qu'ils génèrent, notamment dans un processus itératif comme celui de Crew IA, mais ce serait une bonne expérience à mener pour évaluer leur capacité à la génération humoristique avec ce type de technique.

Par contre si l'évaluation des modèles par d'autres modèles ne s'avère pas fiable pour des critères subjectifs comme l'humour, ça soulève des interrogations sur la pertinence de certaines méthodes d'inférence. Si l'évaluation automatisée par des modèles manque de fiabilité, comment justifier l'utilisation de techniques qui augmentent considérablement le coût computationnel, et donc l'empreinte environnementale, pour générer une réponse ? Des approches comme la Chaîne de Pensée (Chain-of-Thought) ou la génération parallèle de réponses multiples nécessitent plus de calculs pour potentiellement améliorer la qualité. Mais si cette amélioration ne peut pas être évaluée de manière fiable ou si le gain est marginal, le surcoût devient difficilement justifiable, surtout face aux préoccupations environnementales croissantes liées à l'IA.

La Chaîne de Pensée illustre bien ce dilemme. Des études montrent son efficacité notable pour des tâches objectives comme les mathématiques, où elle permet aux grands modèles de décomposer un problème en étapes logiques intermédiaires, augmentant ainsi la probabilité de converger vers la bonne solution. Ce mécanisme semble cependant moins pertinent pour des tâches subjectives comme l'humour. Donc la Chaîne de Pensée semble seulement justifiable lorsque la tâche est objective et décomposable en étapes. Il convient également de noter que la Chaîne de Pensée requiert des modèles de très grande taille pour être efficace. [24] [12]

En résumé, bien que les modèles de langage fondamentaux puissent établir une certaine hiérarchie dans l'humour, ils ne parviennent ni à générer des blagues réellement efficaces ni à en évaluer précisément la qualité. Ces résultats mettent en garde contre une utilisation des LLM pour évaluer d'autres modèles et soulignent la nécessité d'étudier plus en détail les façons d'évaluer les modèles de langage.

Réflexion sur les LLMs pour l’humour

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, comme le souligne ce travail.

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. 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 comme ces modèles sont entraînés sur une très large variété de données, cela les rend capables de faire référence à de nombreux éléments.

Bien que cette étude soit en français et que de nombreuses études que j’ai lues sur le sujet étaient en anglais, on pourrait présumer que la langue ne fait pas beaucoup de différence pour la génération de blagues des grands modèles de langage, à l’exception des jeux de mots. En effet, selon l’étude “On the Biology of a Large Language Model” [13], dans la section qui explore les circuits multilingues, il est indiqué que la réflexion se fait dans un espace qui n’a pas encore de langue, et ensuite, selon la langue demandée, le modèle répond, mais la réflexion serait indépendante de la langue.

Un modèle fine-tuné spécifiquement pour la production d’humour pourrait générer des réponses plus pertinentes et originales, car le fine-tuning permet de personnaliser la génération et de l’adapter au public ciblé. Les modèles fine-tunés dans le cadre de mes travaux étaient tous trop petits, car je n’avais pas les ressources pour faire tourner de gros modèles, mais fine-tuner des modèles beaucoup plus grands serait une piste intéressante pour améliorer les performances en génération humoristique.

Une expérience intéressante que nous n’avons pas pu réaliser aurait été de tester la création de sketches par un LLM en duo avec un humoriste pour voir si cela peut être efficace pour les aider. Ici, nous testions la génération de blagues sans faire de feedback ni de commentaires pour améliorer les blagues, mais il est très rare que les LLM soient utilisés de cette manière sans itérer plusieurs fois. Que ce soit pour aider à écrire des textes, du code ou obtenir des informations, souvent les meilleures réponses arrivent à la suite d’un dialogue avec l’utilisateur qui aide le LLM à fournir les réponses attendues. Il serait aussi intéressant de réfléchir à la manière dont nous pourrions créer les meilleurs outils pour les humoristes pour la création de blagues.

Les résultats sont assez mauvais pour le moment et les LLMs semblent pas encore assez bons pour beaucoup aider à la création de sketches, même si à la vitesse à laquelle les modèles s'améliorent, ils ne vont pas tarder à l'être. Aussi, plusieurs autres façons de prompter les LLMs pour l'humour restent à tester en détails. J'ai obtenu d'assez bons résultats quand on lui demande de se prendre pour un humoriste, à condition que l'humoriste soit beaucoup connu. Par exemple, des grands noms de l'humour québécois comme Mike Ward ou François Bellefeuille fonctionnent, mais des petits noms ne fonctionnent pas. Il serait vraiment intéressant de refaire des tests en demandant de faire des blagues dans le style d'un humoriste. Voici des exemples de résultats obtenus pour François Bellefeuille (2 meilleurs sur 14) :

J'ai acheté du savon à vaisselle "sans parfum". MAIS POURQUOI Y'A UNE ÉTIQUETTE AVEC UNE FLEUR DESSUS ?! C'est censé sentir rien... mais visiblement, ça sent l'hypocrisie florale. J'veux juste que mes assiettes sentent... l'absence. L'absence de spaghetti. C'est tout.

Le dentiste m'a demandé si je passais la soie dentaire. OUI. Entre mes rendez-vous. Genre une fois par an... la veille. C'est pas du mensonge, c'est du... timing stratégique.

Même l'humour dans un style semblait fonctionner très bien, mais pour certains styles d'humour seulement. Par exemple, l'humour d'observation (2 meilleurs sur 20) :

Les machines à laver, elles ont 25 programmes. Mais 90% du temps, tu fais juste "Coton 40°". Y'a un programme "laine délicate"... mais est-ce que quelqu'un a déjà eu le cran de l'utiliser sans appeler sa mère avant x ?

Les chaises de salle d'attente, elles sont toujours alignées, toutes identiques, et pourtant tu passes 5 minutes à choisir la bonne. Et si quelqu'un s'assoit à côté alors que y'avait 10 autres places ? Psychologiquement, t'es pas bien. C'est une agression passive.

Donc, c'est assez décevant d'avoir obtenu des résultats aussi mauvais et j'aurais dû faire des tests avec plus de prompts différents. Par contre, même si ceux que j'ai montrés sont bons, ils proviennent de nombreuses générations dont j'ai sélectionné les meilleurs, et le LLM était très mauvais pour l'évaluation si on avait voulu retirer complètement l'humain de la création.

J'ai testé le modèle unsensored deepseek-r1-abliterated 14b, mais il reste trop limité en taille pour offrir de bonnes performances. Par ailleurs, la censure des modèles ne semble pas vraiment poser de problème pour générer des blagues. Il est facile de contourner ses protections avec quelques manipulations, mais ce qu'il produit est souvent vulgaire plutôt que réellement drôle. En le poussant simplement à produire de l'humour noir, sans recourir à des techniques de jailbreak plus avancées, j'ai déjà obtenu des blagues comme :

Mon pote a eu un accident de voiture. Il a perdu ses deux jambes. Maintenant il dit que sa vie a "pris un tournant". J'ai dit : "Ben, plus besoin de tourner, t'avances plus, bro."

Ma tante est morte d'un cancer. On l'appelait "la guerrière". Mais entre toi pis moi, elle a perdu la guerre en 6 mois. Les nazis ont fait mieux.

Et d'autres que je ne montre pas. Donc, difficile de croire qu'il manque réellement de liberté au modèle pour faire des choses plus drôles.

Défi rencontré et amélioration possible

Un gros défi rencontré a été de trouver des personnes souhaitant évaluer les blagues. Par exemple, pour la première expérience, j'ai demandé à plusieurs personnes, mais les participants ont changé entre-temps pour les différents tests, ce qui n'est pas idéal pour la précision des résultats. Pour la deuxième expérience, nous avons eu beaucoup de participants, mais comme elle était en ligne, nous avons moins de contrôle sur les expérimentations. Trouver les candidats pour les tests a donc été difficile.

D'autre part, un gros défi a été les métriques d'évaluation. Comment évaluer une blague est une question tellement compliquée. J'ai lu plusieurs articles à ce sujet et j'ai l'impression de ne pas l'avoir fait correctement. Pour la première expérience, je m'étais inspiré de cette étude : *The Last Laugh: exploring the role of humor as a benchmark for large language models* [14]. La deuxième, je l'ai simplifiée avec des emojis (trois emojis), pour ne pas que cela soit trop compliqué étant donné que c'était en ligne, et aussi parce que les emojis expriment bien les émotions, ce qui semblait être un bon choix, même si maintenant, je ne suis plus certain. Le professeur Damien Masson m'avait suggéré une méthode qui semblait bonne et m'avait redirigé vers des articles sur ce sujet, mais la méthode de notation semblait trop fastidieuse pour des répondants en ligne, donc je ne l'ai pas choisie. Tout cela pour dire qu'il est très difficile de choisir comment construire son sondage et son évaluation, et j'ai eu des difficultés à ce niveau.

Par ailleurs, pendant que les expériences avaient lieu, de nouveaux modèles encore plus puissants furent créés, ce qui fait que les résultats ne restent pas pertinents longtemps, vu la rapidité à laquelle les nouveaux modèles sortent. L'intelligence artificielle, ces dernières années, est définitivement un domaine difficile à suivre en raison de la fréquence des sorties de modèles et de publications, voire impossible à suivre. Mais tant mieux si nous vivons dans un âge d'or de l'innovation.

Finalement, à plusieurs reprises, j'ai mentionné les autres modèles possibles, comme les modèles non censurés, mais il m'aurait fallu un budget ou des serveurs pour les faire tourner, donc il me manquait des ressources pour aller plus loin. Mais ce n'est pas très grave, car il y avait tellement de pistes que je pouvais aussi explorer, mais que je n'ai pas suivies. La ressource qui m'a le plus manqué, et la plus précieuse, était le temps. Le cours de projet dure 150 heures. Chaque semaine, je me forçais à y consacrer 10 heures, avec des périodes précises, pour être sûr de faire au minimum le bon nombre d'heures, sans en faire trop non plus, car j'ai d'autres cours et je ne voulais pas mettre trop de temps d'un coup et pas assez l'autre semaine. Je voulais être constant, même si vers la fin, pour l'écriture du rapport et la présentation, j'ai mis un coup de boost. J'ai bien respecté le temps et je suis content de l'avoir fait. Mais, avec plus de temps, j'aurais pu faire davantage. Toutefois, plus je trouvais des réponses à mes questions, plus de nouvelles questions apparaissaient. Je me dis donc que j'ai choisi un sujet pour le projet, mais qu'on ne peut jamais faire le tour. Il y a des chercheurs qui se consacrent à chaque sujet et qui ne font pas le tour. Alors, un projet de 150 heures, ce n'est pas surprenant que je n'aie fait qu'effleurer la surface. Cela a néanmoins été une belle expérience pour toucher au monde de la recherche et apprendre à lire des articles.

Conclusion

Cette étude visait à évaluer la capacité des grands modèles de langage actuels à générer de l'humour. Les résultats obtenus montrent que les LLM peinent à produire des blagues courtes de manière consistante et tendent à reproduire des schémas et thèmes récurrents. Ils montrent cependant un potentiel plus encourageant lorsque la tâche leur offre une plus grande liberté créative, comme dans la génération de sketches. Parmi les modèles testés, GPT-4, Gemini Flash 2.0, Deepseek-R1 et Claude ont été les plus performants, bien qu'aucun modèle ne se soit révélé très bon pour la génération d'humour. Un constat majeur de cette analyse est la difficulté des LLM à évaluer la qualité humoristique de manière fiable. Ils tendent à surévaluer les blagues en se basant sur des aspects de forme du texte plutôt que sur si c'est réellement comique, ce qui remet en question leur utilisation comme évaluateurs objectifs pour des tâches subjectives comme l'humour.

Plusieurs pistes futures émergent de cette étude. Le fine-tuning de grands modèles spécifiquement pour l'humour, l'exploration de techniques de prompting plus avancées comme le role-prompting, et surtout le développement de méthodes d'évaluation de l'humour plus fiables sont des axes de recherche futurs intéressants.

Quel est l'impact potentiel sur le secteur de l'humour ? À ce stade, les LLM ne représentent pas une menace de remplacement pour les humoristes. Ils offrent cependant un potentiel certain comme outils pour les humoristes, pouvant servir à générer des idées initiales, explorer des pistes créatives ou surmonter une page blanche. Dans les prochaines années, il est certain que l'on verra de plus en plus d'outils pour aider à l'écriture de sketches humoristiques. Même si les LLM actuels ne sont pas encore des maîtres de l'humour, le domaine évolue très rapidement et il sera donc intéressant de suivre l'avancée de ce sujet.

Remerciement

Je voudrais remercier en premier lieu Benoît Baudry qui a bien voulu me superviser pour ce projet. Il m'a aidé à rester motivé et joyeux, m'a donné des idées, m'a bien fait rire et a rendu cette expérience plaisante. Je voudrais aussi remercier ma famille et mes amis, les membres de l'association des étudiants en informatique de l'UdeM ainsi que tous les autres qui ont bien voulu évaluer des blagues, d'avoir pris le temps de le faire même si les blagues n'étaient pas souvent drôles. Damien Masson qui m'a conseillé sur plusieurs sujets durant mon étude. Tous les chercheurs avant moi qui ont fait des travaux que j'ai lus pour m'informer sur le sujet. Et finalement, les grandes compagnies d'IA qui nous permettent d'utiliser gratuitement leurs outils.

Références

- [1] Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. “You Told Me That Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13701–13715. DOI: 10.18653/v1/2023.emnlp-main.845. URL: <https://aclanthology.org/2023.emnlp-main.845>.
- [2] 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>.
- [3] Fabricio Goes et al. *Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges*. 2022. arXiv: 2212.11214 [cs.AI]. URL: <https://arxiv.org/abs/2212.11214>.
- [4] Drew Gorenz and Norbert Schwarz. “How funny is ChatGPT? A comparison of human- and A.I.-produced jokes”. In: *PLOS ONE* 19.7 (July 2024), pp. 1–13. DOI: 10.1371/journal.pone.0305364. URL: <https://doi.org/10.1371/journal.pone.0305364>.
- [5] Eric Hartford. *Uncensored Models — erichartford.com*. <https://erichartford.com/uncensored-models>. [Accessed 08-02-2025].
- [6] Zachary Horvitz et al. *Getting Serious about Humor: Crafting Humor Datasets with Unfunny Large Language Models*. 2024. arXiv: 2403.00794 [cs.CL]. URL: <https://arxiv.org/abs/2403.00794>.
- [7] *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/>.
- [8] *Humor Detection and Appreciation: Human versus LLM performance*. <https://osf.io/yzupx>.
- [9] *Humor Understanding Multi-task Optimization & Ranking — huggingface.co*. <https://huggingface.co/blog/TuringsSolutions/humortest>.
- [10] Sophie Jentzsch and Kristian Kersting. *ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models*. 2023. arXiv: 2306.04563 [cs.AI]. URL: <https://arxiv.org/abs/2306.04563>.
- [11] Sean Kim and Lydia B. Chilton. *AI Humor Generation: Cognitive, Social and Creative Skills for Effective Humor*. 2025. arXiv: 2502.07981 [cs.HC]. URL: <https://arxiv.org/abs/2502.07981>.
- [12] Nathan Lambert et al. *Tulu 3: Pushing Frontiers in Open Language Model Post-Training*. 2025. arXiv: 2411.15124 [cs.CL]. URL: <https://arxiv.org/abs/2411.15124>.
- [13] Jack Lindsey et al. “On the Biology of a Large Language Model”. In: *Transformer Circuits Thread* (2025). URL: <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>.
- [14] 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>. [Accessed 08-02-2025].
- [15] Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. *From Punchlines to Predictions: A Metric to Assess LLM Performance in Identifying Humor in Stand-Up Comedy*. 2025. arXiv: 2504.09049 [cs.CL]. URL: <https://arxiv.org/abs/2504.09049>.
- [16] Sander Schulhoff et al. *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. 2025. arXiv: 2406.06608 [cs.CL]. URL: <https://arxiv.org/abs/2406.06608>.

- [17] Alexey Tikhonov and Pavel Shtykovskiy. *Humor Mechanics: Advancing Humor Generation with Multistep Reasoning*. 2024. arXiv: 2405.07280 [cs.CL]. URL: <https://arxiv.org/abs/2405.07280>.
- [18] Joe Toplyn. *Witscript 2: A System for Generating Improvised Jokes Without Wordplay*. 2023. arXiv: 2302.03036 [cs.CL]. URL: <https://arxiv.org/abs/2302.03036>.
- [19] Joe Toplyn. *Witscript 3: A Hybrid AI System for Improvising Jokes in a Conversation*. 2023. arXiv: 2301.02695 [cs.CL]. URL: <https://arxiv.org/abs/2301.02695>.
- [20] Joe Toplyn. *Witscript: A System for Generating Improvised Jokes in a Conversation*. 2023. arXiv: 2302.02008 [cs.CL]. URL: <https://arxiv.org/abs/2302.02008>.
- [21] Dmitry Vikhorev et al. *CleanComedy: Creating Friendly Humor through Generative Techniques*. 2024. arXiv: 2412.09203 [cs.CL]. URL: <https://arxiv.org/abs/2412.09203>.
- [22] 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>. [Accessed 08-02-2025]. 2024.
- [23] Han Wang et al. *Innovative Thinking, Infinite Humor: Humor Research of Large Language Models through Structured Thought Leaps*. 2025. arXiv: 2410.10370 [cs.AI]. URL: <https://arxiv.org/abs/2410.10370>.
- [24] Jason Wei et al. “Chain of Thought Prompting Elicits Reasoning in Large Language Models”. In: *CoRR* abs/2201.11903 (2022). arXiv: 2201.11903. URL: <https://arxiv.org/abs/2201.11903>.
- [25] Jason Wei et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL]. URL: <https://arxiv.org/abs/2206.07682>.
- [26] Zhikun Wu, Thomas Weber, and Florian Müller. “One Does Not Simply Meme Alone: Evaluating Co-Creativity Between LLMs and Humans in the Generation of Humor”. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*. IUI ’25. ACM, Mar. 2025, 1082–1092. DOI: 10.1145/3708359.3712094. URL: <http://dx.doi.org/10.1145/3708359.3712094>.
- [27] Yue Zhang et al. *LLMEval: A Preliminary Study on How to Evaluate Large Language Models*. 2023. arXiv: 2312.07398 [cs.AI]. URL: <https://arxiv.org/abs/2312.07398>.