

Évaluation du potentiel des Grands Modèles de Langage pour le domaine de l'humour

Gabriel Beaudoin

Université de Montréal

Hiver 2025

Plan de la présentation

Introduction

Première Expérience

Deuxième Expérience

Réflexion

Défis Rencontrés

Conclusion

Introduction

- L'humour, plus nécessaire que jamais dans notre époque.
- Parallèlement, l'essor fulgurant des Grands Modèles de Langage (LLM).
- Ces IA génèrent du texte avec une aisance impressionnante.
- Mais peuvent-elles saisir la subtilité, la culture, l'essence humaine de l'humour ?
- **Question** : Les LLM sont-ils réellement capables d'un humour efficace ?

Objectifs du Projet

- Tester la capacité des LLM récents (GPT-4o, Claude 3.5, etc.) à générer de l'humour.
- Analyser l'impact de techniques comme le raisonnement en chaîne (CoT).
- Sonder leur aptitude à évaluer eux-mêmes la qualité d'une blague.
- Dresser un portrait des capacités et limites actuelles des modèles généralistes.
- *Personnellement* : Mieux comprendre le fonctionnement des LLMs.

Première Expérience

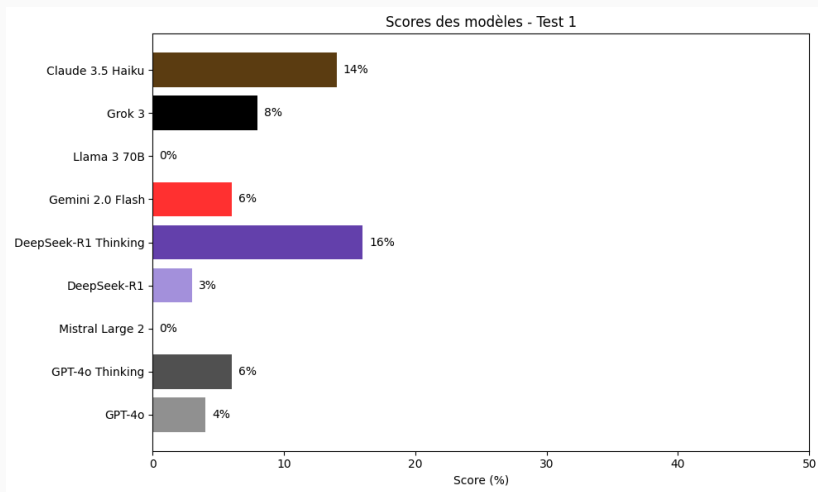
- **Les Tests** : Génération de blagues pour différentes tâches humoristiques.
- **Les Modèles** : GPT-4o, DeepSeek-R1 (+CoT), Mistral L2, Gemini Flash, Llama3 70B, Grok 3, Claude 3.5 Haiku.
- **L'Évaluation** : Quelques humains, notation moyenne de 0 à 10 par génération sur 10 générations de blagues.

Test 1 : Blague Simple (One-liner)

Prompt : "Génère-moi une one-liner joke."

- **Objectif :** Style par défaut.
- **Constat général :** Souvent peu original, blagues connues ou mal exécutées.
- **Exemple (DeepSeek-R1 Chain of Thought) :** *Avant, je jouais du piano à l'oreille... maintenant, j'utilise mes mains.*

Résultats Visuels - Test 1

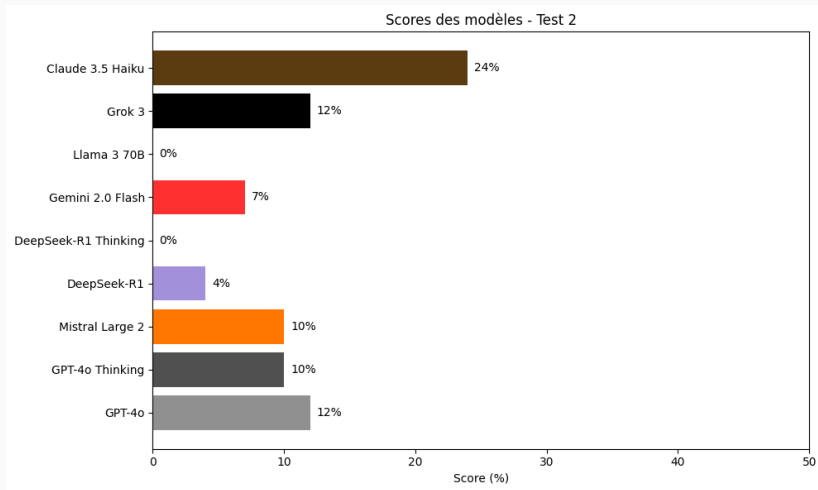


Test 2 : Blague Simple (Rôle Humoriste)

Prompt : "Tu es un humoriste vraiment drôle. Génère-moi une one-liner joke hilarante !"

- **Objectif :** Le 'role-prompting' améliore-t-il ?
- **Constat général :** Pas de changement majeur.
- **Exemple (Claude 3.5) :** *Mon GPS et moi, on a le même problème : on sait exactement où on veut aller, mais on se perd toujours en route.*

Résultats Visuels - Test 2

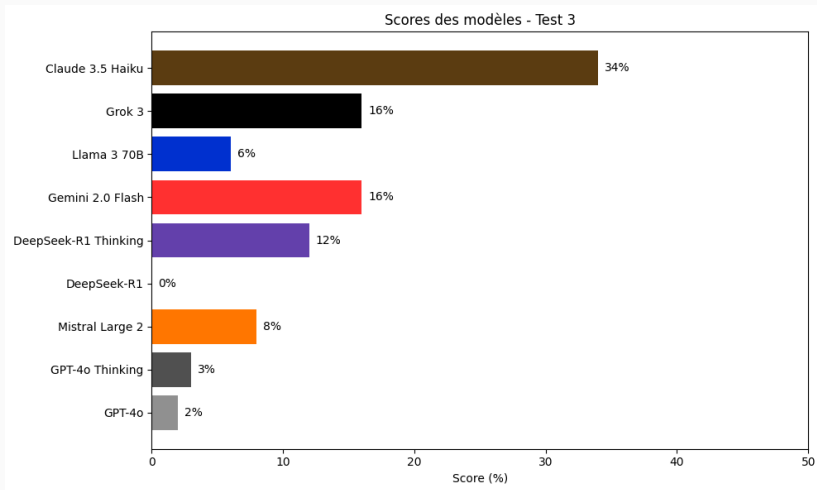


Test 3 : Blague avec Contexte

Prompt : "Tu es un humoriste [...], Génère-moi une blague, mais prends ton temps pour la préparer. [...]"

- **Objectif :** Permettre une construction plus élaborée.
- **Constat général :** Amélioration notable.
- **Exemple (Claude 3.5) :** *Patient: Docteur, J'ai mal à l'oeil quand je bois du café. Médecin: Essayez d'enlever la cuillère de la tasse avant de boire !*

Résultats Visuels - Test 3

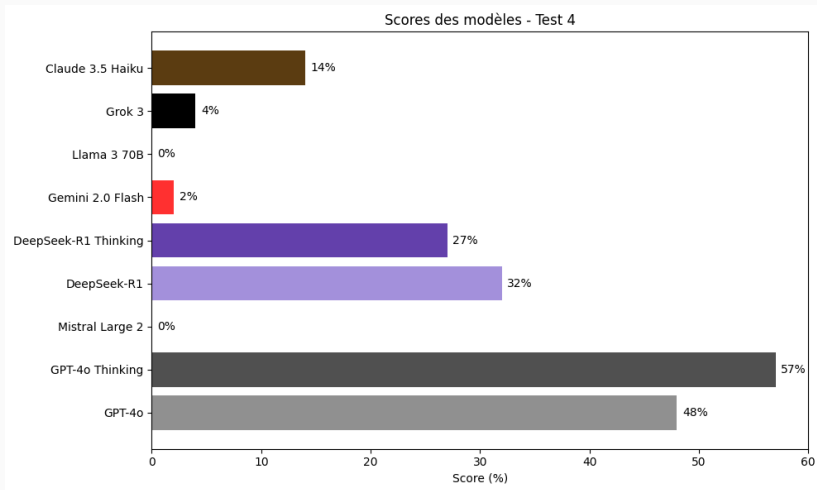


Test 4 : Sketch Humoristique

Prompt : "Tu es un humoriste [...]. Génère-moi un sketch humoristique hilarant !"

- **Objectif :** Évaluer la capacité à créer un dialogue structuré et à élaborer un contexte encore plus détaillé.
- **Constat général :** Certains modèles ont eu des résultats excellents comme GPT-4o, tandis que d'autres ont créé des textes mais pas drôles du tout, comme Gemini 2.0 Flash et Llama 3 70B.
- **Exemple (GPT-4o) :** *[...] j'ai tapé mes symptômes sur Google et apparemment, j'ai soit un rhume... soit trois mois à vivre. [...] 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 !) [...]*

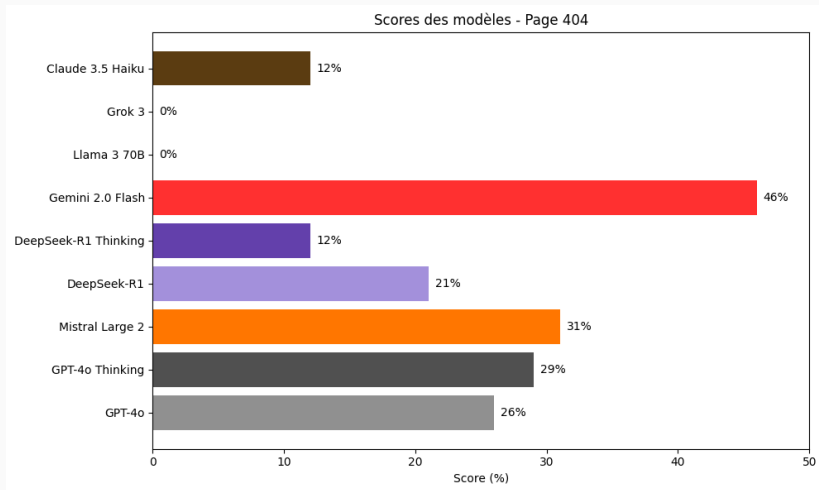
Résultats Visuels - Test 4



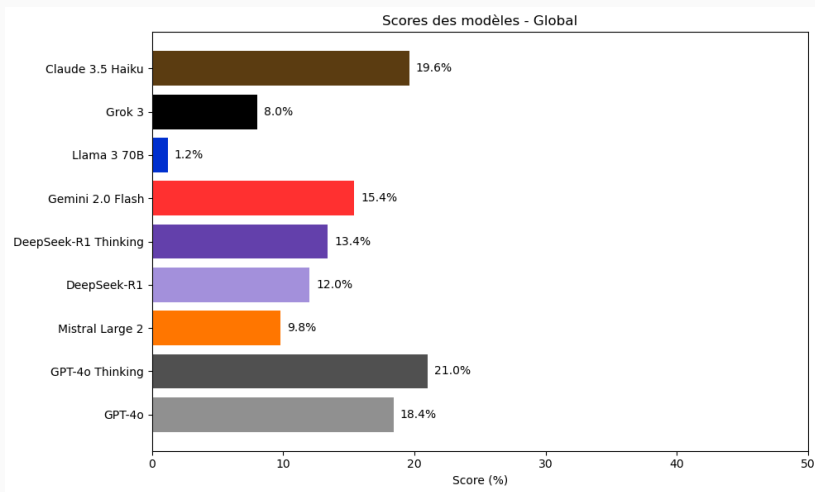
Prompt : "Tu es un humoriste [...]. Génère-moi une idée pour une page 404 drôle pour [site web spécifique avec contexte sur l'entreprise, ex : Dollarama]." (Idée puis HTML/CSS)

- **Objectif :** Explorer un domaine de l'humour pas assez exploré dans les pages web.
- **Constat général :** Bons résultats pour ce genre de tâche.
- **Exemple (GPT-4) :** *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 \$. [...]*

Résultats Visuels - Test 5 (Page 404)



Bilan de la Première Expérience



Discussion - Analyse par Test

- **Tests 1 & 2 (One-liners)** : Performance initiale faible. Le Chain of Thought semble aider, comme DeepSeek qui s'est beaucoup amélioré, mais le role-prompting a peu d'effet sur ce format très contraint.
- **Test 3 (Contexte)** : Autoriser la préparation améliore significativement la qualité. Format plus propice à l'humour que le simple one-liner.
- **Test 4 (Sketches)** : Le plus grand potentiel émerge avec la liberté créative. Potentiel d'aide à l'écriture pour les humoristes.
- **Test 5 (Pages 404)** : Succès notable. Peut-être dû à des attentes inférieures dans ce type de tâche, mais certainement les pages 404 restent sous-exploitées.

- **Répétition et Originalité** : Forte tendance à la récurrence des thèmes (médecin, yoga), des blagues et des structures. L'originalité est un défi majeur.
- **Mémorisation vs. Création** : Les blagues similaires entre modèles posent question. Est-ce de la mémorisation ou une convergence vers des schémas appris ? (Plusieurs études disent des choses contraires.) Selon mes résultats, je pencherais plus pour une mémorisation. [1] [3]
- **Impact du Chain of Thought** : L'amélioration observée semble faible par rapport au coût computationnel supplémentaire de cette technique. Remise en question par l'idée que le raisonnement pourrait être a posteriori, comme le montre l'étude récente d'Anthropic. [7] [5]

Deuxième Expérience

- **L'Objectif** : Valider les tendances, tester l'auto-évaluation des LLM et faire une expérience avec plus de participants.
- **Les Modèles** : Sélection réduite (GPT-4o, DeepSeek-R1, Gemini 2.0 Flash).
- **La Méthode** : Génération massive -> Filtrage par LLM -> Évaluation humaine à grande échelle (89 participants, 3 emojis).

- **Déception** : Les scores moyens des humains restent très faibles, confirmant les difficultés des LLM pour la génération humoristique.
- **Décalage** : Les LLM surestiment massivement la qualité de leurs propres blagues par rapport aux humains.
- **Point Positif** : L'ordre relatif du classement (quelle blague est meilleure qu'une autre) semble globalement cohérent entre LLM et humains.
- **Problème** : Les LLM manquent de cohérence lorsqu'ils évaluent plusieurs fois la même blague. [2]

Table 1: Nombre de votes humains et score moyen global par LLM.

LLM	Pas drôle	Légèrement drôle	Drôle	Évaluations
DeepSeek-R1	135	38	12	185
Gemini Flash 2.0	113	46	14	173
GPT-4o	131	37	13	181

Table 2: Répartition des votes humains selon la note attribuée par le LLM.

LLM	Note LLM	Pas drôle	Lég. 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%

- **Constat Fondamental** : Les LLM imitent la *forme* des blagues, sans réelle *compréhension* de ce qui fait rire.
- **Évaluation par LLM : Non Fiable**
 - Surestime systématiquement la qualité (biais appris ?).
 - Se concentre sur la structure plutôt que sur l'effet comique. (Préfère une forme "classique" à une blague originale mais moins structurée).
 - Manque de cohérence interne (notes variables pour la même blague).
- **Implication Directe** : Si l'évaluation des tâches subjectives par les LLM échoue, comment mesurer l'efficacité réelle des améliorations ? Nécessité de métriques plus robustes.

Discussion - Fiabilité de l'Évaluation et Implications

- **Scénario Optimiste (Si évaluation fiable) :** Potentiel pour des techniques itératives (ex : Crew IA) où les LLM s'auto-évaluent pour améliorer l'humour.
- **Scénario Réaliste (Si évaluation non fiable pour les tâches subjectives) [8] [4] :**
 - **Chain of Thought :** Efficace pour les tâches objectives (maths) grâce à la décomposition logique. Moins pertinent pour l'humour, une tâche subjective où le bénéfice est incertain et qui nécessite de grands modèles.
 - **Justification des Coûts :** Le manque de fiabilité de l'évaluation par les LLM pour l'humour remet en question l'utilisation généralisée de techniques d'inférence coûteuses pour ce type de tâche. Difficile de justifier le surcoût de calcul des méthodes comme le Chain of Thought ou la génération multiple si le gain n'est pas mesurable ou marginal, surtout avec l'impact croissant de l'IA sur l'environnement.

Réflexion

- **Problème** : L'humour repose sur la surprise, l'inattendu. Les LLM reposent sur les probabilités, ce qui limite l'inattendu. Comme nous avons pu l'observer, ils répètent souvent les mêmes thèmes et adoptent des styles d'écriture standard, ce qui limite la surprise.
- **Solution** : 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.

- **Entraînement Généraliste** : Pas spécifiquement pour l'humour, mais contiennent des blagues dans leurs données.
- **Capacités Émergentes** : Les grands modèles montrent des compétences humoristiques accrues.
- **Atout** : Connaissances vastes permettant de puiser dans de nombreuses cultures, références, événements, etc. Très utile pour l'humour contextuel.
- **Hypothèse Multilingue** : La langue ne changerait pas fondamentalement la capacité humoristique (sauf pour les jeux de mots), car la "réflexion" interne serait indépendante de la langue, comme le montre l'étude d'Anthropic. [5]

- **Modèles Non Censurés Testés** : Trop petits pour être performants (ex : DeepSeek-R1-abliterated 14b).
- **Impact de la Censure** : Ne semble **pas être le principal obstacle** à un bon humour.
- **Contournement Facile** : Possible de faire générer de l'humour noir/vulgaire sans jailbreak complexe.
- **Constat** : Le problème n'est pas un manque de "liberté" du modèle, mais une difficulté plus fondamentale à saisir l'essence comique. Produire du vulgaire n'équivaut pas à produire de l'humour efficace.

Autres techniques de prompting

- **Imitation d'Humoristes Connus :**

- Fonctionne étonnamment bien pour des humoristes très connus.
- Moins efficace pour les humoristes moins connus.
- *Exemple (François Bellefeuille) :* "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. [...]"

- **Imitation de Styles d'Humour :**

- Fonctionne bien pour certains styles (ex : humour d'observation).
- *Exemple (humour d'observation) :* 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 ?

- **Fine-tuning Spécifique** : Adapter un grand modèle à l'humour pourrait améliorer la pertinence et l'originalité. (Limité par les ressources dans ce projet).
- **Interaction Humain-IA** :
 - Tester la collaboration LLM + Humoriste pour l'écriture.
 - Importance de l'itération : Les LLM sont souvent meilleurs après plusieurs échanges/feedbacks (non testé ici).
 - Réfléchir aux meilleurs outils pour intégrer les LLM dans le processus créatif des humoristes.
- **Potentiel Futur** : Malgré les mauvais résultats actuels, l'amélioration rapide des modèles laisse entrevoir un bon potentiel, surtout pour assister à la création.

Défis Rencontrés

- **Évaluateurs humains** : Difficile à recruter, surtout en grand nombre.
- **Métriques d'évaluation** : Définir comment noter l'humour est complexe. [9] [6]
- **Évolution rapide des modèles** : Les résultats deviennent vite obsolètes.
- **Ressources limitées** :
 - Budget/serveurs pour tester des modèles plus spécifiques (non censurés, fine-tunés).
 - Temps (projet de 150h)
- **Étendu du sujet** : Chaque réponse soulève de nouvelles questions.

Conclusion

- Les LLM actuels ne sont pas très bons pour la génération humoristique, mais semblent avoir un potentiel prometteur.
- L'évaluation de l'humour par les LLM est peu fiable.
- Les LLM ne remplaceront pas les humoristes de sitôt, mais peuvent devenir des outils d'aide à la création pertinents dans un futur proche.

Remerciements

- Benoît Baudry, de m'avoir supervisé, pour son aide, ses idées et sa bonne humeur.
- Ma famille, mes amis, l'association étudiante (AÉDIROUM), et tous les évaluateurs pour avoir pris le temps d'évaluer des blagues (même si elles étaient souvent pas drôles).
- Damien Masson pour ses conseils sur l'évaluation.
- Les chercheurs dont j'ai lu les papiers de recherche.
- Les compagnies d'IA pour l'accès à leurs modèles.

References

- [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] 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>.
- [3] 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>.

- [4] 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>.
- [5] 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>.
- [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>. [Accessed 08-02-2025].
- [7] 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>.
- [8] 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>.
- [9] 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>.

Vous pouvez retrouver mon rapport complet à l'adresse suivante :
https://github.com/GabrielBeaudoinUdem/ift4055/blob/main/ift4055_rapport_final.pdf

Merci de votre attention !

Des questions ?