Generation of Punning Riddles in Portuguese with Prompt Chaining
Paper type: Late Breaking Results

Marcio Lima Inácio and **Hugo Gonçalo Oliveira**
Centre for Informatics and Systems of the University of Coimbra (CISUC)
Intelligent Systems Associate Laboratory (LASI)
Department of Informatics Engineering
Polo II, Pinhal de Marrocos, 3030-290, Coimbra, Portugal
{mlinacio, hroliv}@dei.uc.pt

Abstract

We present an approach to Humor Generation that adapts the rule-based system SECO into a more modern paradigm using Large Language Models. We introduce a prompt chain method replicating SECO's functionality to create punning riddles in Portuguese. We believe this is the first research on using LLMs for Humor Generation in the Portuguese language. We further employ the CMOS for evaluation, a metric not previously applied to Humor Generation systems. Results indicate no significant difference in funniness between the rule-based system and the LLM approach, with a CMOS of -0.0758, with a slight lean towards SECO. An agreement analysis reveals that raters diverged significantly, reflecting humor's subjective nature. However, the flexibility of LLMs offers a valuable framework for Humor Generation, unrestricted by template limitations and lexica sizes.

Introduction

Natural Language Processing systems have been gaining much traction during the last decade, especially with the development of neural networks and transformers, ultimately leading to the creation of Large Language Models (LLMs). This kind of model has been proven to perform extremely well in Natural Language Generation (NLG) scenarios (Mialon et al. 2023), creating fluent natural-sounding texts. However, LLMs still have their limitations, creating significant paths for future research.

As mentioned by Li et al. (2023a), creativity is one of the new frontiers to pursue with this type of system, as it has to deal not only with text form but also with open-ended outputs and subjective perceptions. In this context, we consider Humor Generation as a creative task and aim at exploring it with LLMs, especially with recent observations that such models can produce humor as well as, or even better than, an average person (Gorenz and Schwarz 2024).

Specifically, our work deals with Humor Generation in the Portuguese language, which has less research than languages such as English. To the extent of our knowledge, there is no work on using LLMs for Humor Generation in Portuguese. The most recent work that aims at creating humorous riddles in this language is SECO, a rule-based system created by Gonçalo Oliveira and Rodrigues (2018) to build punning riddles from lexical-semantic relations.

In this paper, we use SECO as an inspiration for creating a prompt chain that generates riddles in the same format. This way, we aim at understanding if using an LLM can result in better jokes, due to its capacity to produce fluent text and encode lexical relations in its deep structures. Thus, we evaluate the outputs using CMOS, a method largely used in the speech generation community to conduct subjective evaluations of automatically generated content (Loizou 2011).

Evaluation results, with 10 native speakers, show that our LLM prompt chain creates riddles comparable to those of SECO, with no significant differences related to funniness. We also show that the scoring decisions largely differ across evaluators, probably due to the subjective nature of the task. We believe that this approach, based on prompt engineering, is important to open future research paths for Humor Generation in Portuguese, with more flexible and adaptable systems. This is possible by using LLMs rather than a fixed set of rules and templates to create texts.[1]

The remainder of the paper is organized as follows. First, we discuss some related work on automatic Humor Generation, focusing on works that inspired this paper. Afterward, we present the resources and methodologies used in this work, followed by a discussion of the evaluation results. The conclusions of our research are then discussed, together with a review of some limitations of our work and considerations about ethical aspects of Humor Generation.

Related Work

Humor generation dates back to the 1990s, with rule-based systems such as LIBJOG (Raskin and Attardo 1994) and JAPE (Binsted and Ritchie 1994). This paradigm, usually with fixed templates, has been used until very recently, with approaches based on lexical replacement (Valitutti et al. 2016) or gap filling (Winters, Nys, and De Schreye 2018).

Some works also explored Machine Learning and NLG, with Recurrent Neural Networks (Yu, Tan, and Wan 2018), Generative Adversarial Networks (Diao et al. 2020), and more recently, LLMs and prompt engineering (Mittal, Tian, and Peng 2022; Gorenz and Schwarz 2024).

Directly related to this work, there is SECO (Gonçalo Oliveira and Rodrigues 2018), which

[1] All code, data, and results are available on: `https://github.com/NLP-CISUC/SECO-LLM`.

is, to the extent of our knowledge, one of the few systems for Humor Generation in Portuguese, alongside Memegera (Gonçalo Oliveira, Costa, and Pinto 2016). More specifically, SECO creates riddles by exploring compound words and amalgams, i.e. words that can be split into two parts that have a meaning on their own. Then, their method uses semantic relations (for example, antonymy, hyponymy, etc.) to fill templates and produce jokes. Examples of SECO-like riddles are: "What's the opposite of artificial intelligence? Natural stupidity!" and "What do you get when you cross a chicken with a vegetable? A peacock!". The authors also propose a scoring function for ranking and filtering the generated jokes, to show the user only the best ones. This ranking prioritizes riddles with more frequent words that should be recognizable by most people.

We also get inspiration from previous works by Chen, Shi, and Si (2023) and Toplyn (2022), who explored a step-by-step approach for Humor Generation with LLMs, by incrementally prompting the model to follow the theory of humor created by Toplyn (2014). These works are important as they highlight how humor can be constructed through reasoning and exploration of associations, relating to the rule-based methods of SECO.

Methodology

Our methodology derives from that of Gonçalo Oliveira and Rodrigues (2018), using some of the same resources and following a similar rationale. More details are presented below.

Corpora and Resources

As starting point, SECO uses a list of amalgams, i.e. words that may be split into two sub-words that have their own meaning. Examples of amalgams in English are "nobody" ("no" + "body") and "hardship" ("hard" + "ship"). Another resource used in SECO is a list of various compound expressions (Ramisch et al. 2016), such as "french fries", "solar system", or "dead-end".

In this paper, we utilized the exact same input concepts used by Gonçalo Oliveira and Rodrigues (2018) in their work to enable a fair comparison between the two systems.

Prompt Chaining

To follow the same rationale of SECO, which is a rule-based method comprised of multiple steps, we decided to follow a prompt chaining approach, similarly to what Chen, Shi, and Si (2023) did with the theory of Toplyn (2014). To start, the model receives a system prompt, which controls its general behavior. The prompt given was[2]:

> *You're an assistant with a great sense of humor who loves to create puns and funny word games. You will help the user follow a step-by-step reasoning to generate a pun, wordplay, or joke related to that topic. The joke must be original, creative, and make the reader laugh.*

First, we present the input concept (w) and its sub-words (w_1 and w_2) according to the corresponding lexicon, asking for related concepts to each sub-word, the chosen relations (r_1 and r_2) are the same used by SECO to produce its riddle. The prompt template can be seen below:

> *Look at this word: w. It can be split into two parts: w_1; and w_2. What are the r_1 of w_1 and the r_2 of w_2?*

This step of obtaining a list of related words corresponds to SECO's usage of lexical-semantic resources, such as OpenWordNet-PT (de Paiva, Rademaker, and de Melo 2012) and ConceptNet (Speer, Chin, and Havasi 2017). Then, we prompt the model to create a list of jokes that encompasses all these concepts following one of the riddle patterns (p) used by SECO.

> *Create a list of jokes in a question-answer format that combines w with the r_1 of w_1 and the r_2 of w_2. The joke must follow, in general terms, the following pattern: p.*

The different templates available for p are: (i) What is the opposite of w? X; (ii) What does w mean? X; (iii) What results from crossing X and Y? w. Again, for each riddle, we used the same template used by SECO in the corresponding rule-based generated joke.

Finally, to account for the scoring that SECO has to determine the best jokes to be presented for the user, we prompted the model to select the best joke from the ones it generated. The prompt is as follows:

> *From this list of jokes, choose only one that is the funniest and makes the most people laugh. Present the text of the joke including the pattern's question, without any further information.*

The LLM used in this work was Sabiá-2 Medium (Almeida et al. 2024) via MariTalk API[3] calls. We choose this model especially because it was trained exclusively with data in Brazilian Portuguese and is reported to produce results similar to or better than GPT-3.5 while being cheaper. Since this is a proprietary model, the authors do not disclose their training methodology or architecture. The total cost of this work, including early experimentation and final generation, was R$8.61 (US$1.64, €1.54). Examples of outputs for each prompt can be seen in Table 1.

Evaluation Protocol

To evaluate Sabiá-2 on this task of riddle regeneration, we wanted to avoid using absolute scales (e.g. a Likert scale of quality), since it is not possible to have a sufficiently controlled environment that this kind of evaluation requires, e.g. using the same people with access to the same resources. Therefore, we employ the Comparative Mean Opinion Score (CMOS), a subjective evaluation methodology largely used to compare Text-to-Speech (TTS) systems (Loizou 2011).

In this sense, during evaluation, ten native speakers of Portuguese, from both Brazil and Portugal, were presented, each, a set of 100 pairs of riddles, one produced by SECO and another created by the LLM, generated from the same

[2]The original prompts, which can be seen in Table 1, were translated to English for accessibility of the paper.

[3]https://www.maritaca.ai/

Prompt	Response
Observe essa palavra: planta baixa. Ela pode ser dividida em duas partes: planta; e baixa. Quais são os hipônimos de planta e os sinônimos de baixa?	Os hipônimos de “planta” são palavras que estão contidas dentro do significado de “planta”, como “árvore”, “grama”, “flor”, “arbusto”, “vegetal”, etc. Já os sinônimos de “baixa” são palavras que têm significados semelhantes, como “pequena”, “reduzida”, “curta”, “baixo” (quando usado como adjetivo), “diminuta”, etc. Agora, vamos criar um trocadilho usando essas informações: Pergunta: O que a planta baixa disse para a árvore alta? Resposta: “Você está me deixando com inveja de crescer!” Este trocadilho brinca com a ideia de que a “planta baixa” está se sentindo inferior em comparação com a “árvore alta”, e a “inveja de crescer” é uma expressão idiomática que significa querer alcançar o mesmo nível ou status que outra pessoa.
Crie uma lista de piadas em forma de pergunta e resposta que combinem planta baixa com os hipônimos de planta e os sinônimos de baixa. A piada deve seguir, em linhas gerais, o seguinte modelo: “Que resulta do cruzamento entre X e Y ? planta baixa”.	1. Que resulta do cruzamento entre um jardim e um anão? Planta baixa. 2. Que resulta do cruzamento entre um gramado e um hobbit? Planta baixa. 3. Que resulta do cruzamento entre um vaso de flor e um índio pequeno? Planta baixa. 4. Que resulta do cruzamento entre um arbusto e um duende? Planta baixa. 5. Que resulta do cruzamento entre um pé de feijão e um anão? Planta baixa. 6. Que resulta do cruzamento entre um canteiro de horta e um pigmeu? Planta baixa. 7. Que resulta do cruzamento entre um pote de cacto e um homem pequeno? Planta baixa. 8. Que resulta do cruzamento entre um bonsai e um goblin? Planta baixa. 9. Que resulta do cruzamento entre um jardim de inverno e um anão de jardim? Planta baixa. 10. Que resulta do cruzamento entre um vaso de planta e um elfo pequeno? Planta baixa. Essas piadas jogam com a ideia de cruzamento entre diferentes tipos de plantas ou ambientes relacionados a plantas e seres pequenos ou anões, resultando na “planta baixa” como um trocadilho.
Dessa lista de piadas, escolha somente uma que seja a mais engraçada e que faça o maior número de pessoas rir. Apresente o texto da piada incluindo a pergunta do modelo, sem mais informações.	Que resulta do cruzamento entre um jardim e um anão? Planta baixa.

Table 1: Example of prompts and corresponding LLM responses. An English translated version is made available as appendix.

input concept, word relations, and question pattern. Each pair has to be evaluated on a scale from -2 to 2, indicating which riddle is better in the matter of funniness. The riddles were presented randomly, and the evaluator could stop whenever they desire.

Results

Out of the ten people who evaluated the generated pairs, seven of them covered all 100 jokes. For each evaluation pair, we combined the different scores by calculating the average across all evaluators. In this sense, the final CMOS (average of averaged scores) is -0.0758 ± 0.7672.

With this result, we observe that no system is better in this scenario ($|\text{CMOS}| < 0.2$), especially with such a high standard deviation (0.7672). The median value of the aggregated scores is 0, with scores ranging from -1.67 to 1.89. A general distribution of the evaluation scores is in Figure 1.

To check if the results are reliable, especially if the median differs significantly from zero (i.e. the systems differ according to the sample distribution), we carried out some statistical testing. Namely, we used the one-way Wilcoxon signed-rank test (Corder and Foreman 2011, p. 40), since the distribution we obtained is not considered normal (Shapiro-Wilk p-value $= 0.01848 < 0.05$) (Royston 1995). The test did not reject the null hypothesis (Wilcoxon p-value $= 0.4112 > 0.05$, CI $= [-0.25, 0.12]$); thus, we cannot state that the systems are significantly different from one another in terms of funniness.

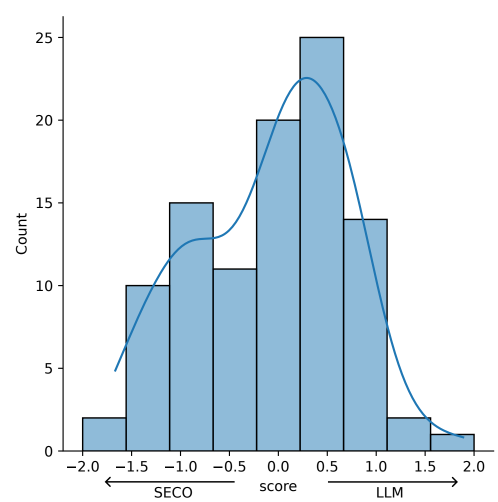

Figure 1: Histogram of evaluation scores

Given that the standard deviation of scores is high, we also decided to check if the raters agreed with each other regarding which system is better. To this extent, we used Krippendorf's Alpha (Krippendorff 2019), resulting in $\alpha = 0.341$, which is poor; this means that the scores differ significantly depending on the person who is evaluating the pair. If we consider the subjective nature of humor, i.e. it depends not only on the text itself but also on the person reading such text, this result is expected. Even though we first believed that using CMOS could circumvent this challenge, since it is a methodology designed for subjective evaluation, it does not seem up to par with the difficulties that humor poses.

To illustrate some outputs, we present the highest (LLM is preferred by the raters) and lowest scored (SECO is preferred) pairs in Table 2. In these examples, we can see that the LLM can draw more complex relations, such as linking "leiteiro" (*milkman*) and "matemático" (*mathematician*) with "leite integral" (*whole milk*) instead of relying on direct links in a lexicon. On the other hand, SECO is more predictable and robust against word repetition, such as in "ca, causa, causa", it also guarantees that the punchline will be related to the sub-words and not exclusively to the whole input concept ("efeito" being the antonym of "causa").

SECO	LLM	CMOS
Que resulta do cruzamento entre um branco e o que é global? Leite integral.	Que resulta do cruzamento entre um matemático e um leiteiro? Leite integral.	1.89
Qual é o contrário de causarei? Efeito-rainha.	Qual é o contrário de "ca, causa, causa"? "Efeito, efeito, efeito".	-1.67

Table 2: Examples of riddles with highest absolute CMOS.

Conclusion

In this paper, we adapted SECO, a rule-based system that generates humorous riddles in Portuguese, into an LLM-based paradigm. To this extent, we developed a prompt chain that mimics the functioning of SECO by receiving a compound word or an amalgam, looking up other words through lexical-semantic relations, generating candidate jokes, and finally selecting the best ones to present to the user. To the best of our knowledge, this is the first research on using LLMs for Humor Generation in Portuguese.

For the evaluation, we used Comparative Mean Opinion Score (CMOS), a subjective evaluation metric largely used in the TTS community. To the extent of our knowledge, this evaluation metric has not yet been explored to evaluate humor generation systems.

Results show no significant difference, with a CMOS of -0.0758, between the two systems regarding funniness. An agreement analysis with Krippendorff's Alpha showed that raters hardly agreed with each other ($\alpha = 0.341$), which is expected given the highly subjective nature of humor.

On the other hand, we argue that creating such an approach with LLMs is still valuable, as it is largely more flexible than rule-based ones. When using an LLM, the generation process is not limited to specific templates nor conditioned by the quality and size of underlying lexica. Having such an adaptable approach for Humor Generation, that performs as well as the previous existing one, can open many paths for future research on this matter.

As examples of future investigations, we mention exploring evolutionary computation to create better jokes based on fitness metrics (Winters and Delobelle 2021), which could also lead to the exploration of high-quality automatic metrics for assessing humor quality. Our prompt chain can also be expanded with other steps not originally present in SECO, such as concept expansion through brainstorming (Li et al. 2023b) or enhancing generated jokes with feedback (Madaan et al. 2023).

Author Contributions

Marcio Lima Inácio carried out the experimentation, methodology definition, evaluation, result analysis, and paper writing. Hugo Gonçalo Oliveira provided the original code and results for SECO, advised throughout the work, and helped write and revise the paper.

Limitations

As the main limitation of this work, we mention that the evaluation process, due to resource and time constraints, was restricted to evaluating only one aspect of the text (funniness) within a fixed CMOS scale, which does not capture other aspects such as novelty or naturalness. We also acknowledge that CMOS exclusively captures a relation between the LLM and SECO, without evaluating the actual quality of the jokes produced (if they are funny by themselves); anecdotally, some volunteers mentioned that in most of the pairs, none of the jokes were funny at all.

We also understand that our results are conditioned to the LLM and prompts used. Using other models and phrasings can produce significantly different outputs, which could outperform our methods and SECO.

Ethical Considerations

A major ethical concern we mention is that the Sabiá-2 models are closed, i.e. we do not have access to their sizes, architecture, or training data. However, we decided to continue using this model as it was specifically trained for the Portuguese language with instruction-following capabilities, enabling us to use prompt chaining and avoid creating unrelated texts in other languages, which happened in some exploratory experiments with open models such as Llama2 (Touvron et al. 2023).

Additionally, we mention that humor has been historically used to perpetuate prejudice in society (Drakett et al. 2018) and maintain relations of power (Bitterly 2022). Therefore, we advise caution when creating, reporting, and disseminating systems that generate this kind of content. On this matter, we acknowledge that our system does not have a filtering or moderation step to avoid problematic outcomes.

Acknowledgments

We thank all volunteering participants who took part in our evaluation process. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology, I.P. (grant number UI/BD/153496/2022), within the scope of the project CISUC (UID/CEC/00326/2020).

References

Almeida, T. S.; Abonizio, H.; Nogueira, R.; and Pires, R. 2024. Sabiá-2: A New Generation of Portuguese Large Language Models. arXiv:2403.09887.

Binsted, K., and Ritchie, G. 1994. An implemented model of punning riddles. In Hayes-Roth, B., and Korf, R. E., eds., *Proceedings of the 12th National Conference on Artificial Intelligence*, volume 1, 633–638. Seattle: AAAI Press / The MIT Press.

Bitterly, T. B. 2022. Humor and power. *Current Opinion in Psychology* 43:125–128.

Chen, Y.; Shi, B.; and Si, M. 2023. Prompt to GPT-3: Step-by-Step Thinking Instructions for Humor Generation. In *Proceedings of the 14th International Conference on Computational Creativity*. Waterloo: ACC.

Corder, G. W., and Foreman, D. I. 2011. *Nonparametric Statistics: A Step-by-Step Approach*. Wiley.

de Paiva, V.; Rademaker, A.; and de Melo, G. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, 353–360. Mumbai: The COLING 2012 Organizing Committee.

Diao, Y.; Yang, L.; Fan, X.; Chu, Y.; Wu, D.; Zhang, S.; and Lin, H. 2020. AFPun-GAN: Ambiguity-Fluency Generative Adversarial Network for Pun Generation. In Zhu, X.; Zhang, M.; Hong, Y.; and He, R., eds., *Natural Language Processing and Chinese Computing*, volume 12430. Cham: Springer International Publishing. 604–616.

Drakett, J.; Rickett, B.; Day, K.; and Milnes, K. 2018. Old jokes, new media – Online sexism and constructions of gender in Internet memes. *Feminism & Psychology* 28(1):109–127.

Gonçalo Oliveira, H., and Rodrigues, R. 2018. Exploring Lexical-Semantic Knowledge in the Generation of Novel Riddles in Portuguese. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*, 17–25. Tilburg, the Netherlands: ACL.

Gonçalo Oliveira, H.; Costa, D.; and Pinto, A. M. 2016. One does not simply produce funny memes! - Explorations on the Automatic Generation of Internet humor. In Pachet, F.; Cardoso, A.; Corruble, V.; and Ghedini, F., eds., *Proceedings of the Seventh International Conference on Computational Creativity*, 238–245. Paris: Sony CSL Paris, France.

Gorenz, D., and Schwarz, N. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. PsyArXiv:5yz8n.

Krippendorff, K. 2019. *Content Analysis: An Introduction to Its Methodology*. 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc.

Li, S.; Han, C.; Yu, P.; Edwards, C.; Li, M.; Wang, X.; Fung, Y.; Yu, C.; Tetreault, J.; Hovy, E.; and Ji, H. 2023a. Defining a New NLP Playground. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11932–11951. Singapore: ACL.

Li, X.-Y.; Xue, J.-T.; Xie, Z.; and Li, M. 2023b. Think Outside the Code: Brainstorming Boosts Large Language Models in Code Generation. arXiv:2305.10679.

Loizou, P. C. 2011. Speech Quality Assessment. In Kacprzyk, J.; Lin, W.; Tao, D.; Kacprzyk, J.; Li, Z.; Izquierdo, E.; and Wang, H., eds., *Multimedia Analysis, Processing and Communications*, volume 346. Berlin, Heidelberg: Springer Berlin Heidelberg. 623–654.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651.

Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; Grave, E.; LeCun, Y.; and Scialom, T. 2023. Augmented Language Models: A Survey. arXiv:2302.07842.

Mittal, A.; Tian, Y.; and Peng, N. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1053–1062. Seattle, United States: ACL.

Ramisch, C.; Cordeiro, S.; Zilio, L.; Idiart, M.; and Villavicencio, A. 2016. How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 156–161. Berlin, Germany: ACL.

Raskin, J. D., and Attardo, S. 1994. Non-literalness and non-bona-fîde in language: An approach to formal and computational treatments of humor. *Pragmatics & Cognition* 2(1):31–69.

Royston, P. 1995. Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Applied Statistics* 44(4):547.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence and the Twenty-ninth Innovative Applications of Artificial Intelligence Conference*.

Toplyn, J. 2014. *Comedy Writing for Late-Night Tv: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*. Rye, New York: Twenty Lane Media, LLC.

Toplyn, J. 2022. Witscript 2: A System for Generating Improvised Jokes Without Wordplay. In *International Con-*

ference on Computational Creativity. Bolzano-Bozen: Association for Computational Creativity (ACC).

Touvron, H., et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Valitutti, A.; Doucet, A.; Toivanen, J. M.; and Toivonen, H. 2016. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering* 22(5):727–749.

Winters, T., and Delobelle, P. 2021. Survival of the Wittiest: Evolving Satire with Language Models. In *Proceedings of the Twelfth International Conference on Computational Creativity*, 82–86. Mexico City, Mexico: ACC.

Winters, T.; Nys, V.; and De Schreye, D. 2018. Automatic Joke Generation: Learning Humor from Examples. In Streitz, N., and Konomi, S., eds., *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*, volume 10922. Cham: Springer International Publishing. 360–377.

Yu, Z.; Tan, J.; and Wan, X. 2018. A Neural Approach to Pun Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1660. Melbourne, Australia: ACL.

Appendix A: Prompt translations

Since not all people interested in humor generation speak or understand Portuguese, in Table 3, we provide the same information as in Table 1, but translated to English. We highlight that, in the first prompt, we make it explicit that the pun word “floor plan” (*planta baixa*) is literally translated as “low plant”, which is the origin of the ambiguity in these jokes.

Prompt	Response
Look at this word: floor plan *("low plant")*. It can be divided into two parts: plant; and low. What are the hyponyms of plant and the synonyms of low?	The hyponyms of "plant" are words contained within the meaning of "plant", such as "tree", "grass", "flower", "shrub", "vegetable", etc. On the other hand, synonyms of "low" are words that have similar meanings, such as "small", "reduced", "short", "low" (when used as an adjective), "diminutive", etc. Now, let's create a pun using this information: Question: What did the floor plan say to the tall tree? Answer: "You're making me jealous growing up!" This pun plays on the idea that the "floor plan" is feeling inferior compared to the "tall tree", and "making me jealous growing up" is an idiom that means wanting to reach the same level or status as someone else.
Create a list of jokes in a question-and-answer format that combines floor plan with the hyponyms of plant and the synonyms of floor plan. The joke should broadly follow the following model: "What results from crossing X and Y ? floor plan".	1. What results from crossing a garden and a dwarf? Floor plan. 2. What results from crossing a lawn and a hobbit? Floor plan. 3. What results from crossing a flower pot and a small Indian? Floor plan. 4. What results from crossing a bush and an elf? Floor plan. 5. What results from crossing a bean plant and a dwarf? Floor plan. 6. What results from crossing a vegetable garden bed and a pygmy? Floor plan. 7. What results from crossing a cactus pot and a small man? Floor plan. 8. What results from crossing a bonsai and a goblin? Floor plan. 9. What results from crossing a winter garden and a garden gnome? Floor plan. 10. What results from crossing a plant pot and a small elf? Floor plan. These jokes play on the idea of crossbreeding between different types of plants or plant-related environments and small or dwarf beings, resulting in "floor plan" as a pun.
From this list of jokes, choose only one that is the funniest and makes the most people laugh. Present the text of the joke including the model question, without any further information. What results from crossing a garden and a dwarf? Floor plan.	

Table 3: Example of translated prompts and corresponding translated LLM responses.