Challenging ChatGPT with different typologies of physics education questions

Pre-print version. Cite as:

Víctor López-Simó, Mikael Frank Rezende; Challenging ChatGPT with Different Types of Physics Education Questions. Phys. Teach. 1 April 2024; 62 (4): 290–294. https://doi.org/10.1119/5.0160160

Introduction

Will my job be replaced by ChatGPT? Can artificial intelligence (AI) engines do homework for students? How do I know if a delivered assignment was made by a robot? These and many other questions have been occupying the minds of professionals from different areas, including professors and researchers, especially since ChatGPT was launched in November 2022. This Generative Pretrained Transformer mechanism works through a chat interface that allows establishing conversations based on the targeted processing of a large volume of data, but the inner functioning of ChatGPT still acts as a "black box" for most of us. As physics educators, we are particularly interested in understanding the kind of information it can provide to students, how reliable this information can be, and where it may still fall short. This helps us better understand how we can use it.

In the last months, different investigations have indicated the need for detailed studies to better understand both the potentialities and limitations of AI in physics teaching and learning scenarios. On the one hand, [1] and [2] presented different strategies to use Chat GTP in the physics classroom, presenting easy-to-implement examples of how ChatGPT can be used in physics classrooms to foster critical thinking skills at the secondary school level [1] and to generate bad examples to be addressed with students to critique and fix them [2]. On the other hand, some investigations have analyzed the level of performance of this IA tool to solve physics problems. According to [3], ChatGPT would narrowly pass a calculus-based physics course while exhibiting many of the preconceptions and errors of a beginning learner. In parallel, [4] found that ChatGPT3.5 can match or exceed the median performance of a university student who has completed one semester of college physics, and [5] found very impressive basic problem-solving capabilities of ChatGPT in interpreting simple physics problems, assuming relevant parameters, and writing correct codes.

Despite those previous contributions focus either on identifying ChatGPT-based physics education good practices or testing ChatGPT physics' performance in comparison with real students, our particular interest lies in understanding how different typologies of physics education problems may influence both the correctness and the variability of the answers provided by the tool. It is well known in Physics Education Research that the typology of physics education questions strongly affects the ways of reasoning and obtaining the answer [6], [7]. For this reason, our research question is: How the correctness and the variability of the answers provided by ChatGPT are affected by the typology of physics education question?

Methods

To define the typologies of physics education questions, we first decided to narrow the domain to one specific physics topic: forces and motion, and to focus only on questions corresponding to upper-secondary school level (between 15 and 17 years old). We also decided not to use multiple-choice questions such as those from FCI [4], programming exercises [3], using exclusively open questions. We also excluded questions related to inquiry, experimental design, data analysis, epistemology, nature of physics, etc., focusing exclusively on conceptual knowledge. Based on these exclusion criteria, and after an extensive discussion with a panel of 12 experienced in-service physics teachers, five typologies of questions were defined (see Table I). These five typologies were organized hierarchically, drawing inspiration from Bloom's Taxonomy, that is, arranged according to a hierarchy level of cognitive demand and thinking skills required.

Table I. Typology of the physics questions proposed to ChatGPT.

Code	Typology	Description	Specific question proposed in the topic Forces and Motion
Q1	**Dictionary definition**	Definition that can be found in the dictionary/textbook	*What does Newton's second law of motion state?*
Q2	**Simple calculation**	Calculation that needs 1 or 2 equations and some numeric values provided by the problem statement.	*A 3.0 kg body is moving on a frictionless horizontal surface with a velocity V_0. At a certain instant (t = 0s), a force of 9.0 N is applied in the opposite direction of motion. Knowing that the body comes to rest at t = 9.0s, what is the initial velocity V_0, in m/s, of the body?*
Q3	**Multi-step calculation**	Calculation that needs many combined equations and many numeric values (some of them provided by the problem statement, some others not).	*A body of mass 3 kg is initially dropped on an incline with a slope of 30^o and a coefficient of friction of 0.15, covering a distance of 2 meters on the incline. Upon reaching a height equal to zero, it starts to move horizontally over another surface, this surface now with a coefficient of friction of 0.4 and a length of 30 cm. At the end of this second surface, there is a spring of K=30N/cm. What will be the maximum compression of the spring after it is reached by the body?*
Q4	**Reasoning problems**	Problem that strongly relies on appropriate conceptualizations counterintuitive to the everyday life reasoning	*A person is on a ship sailing on a very wide and deep river with a constant current. The ship starts to sink, and the person needs to swim to a lifeboat. There are two lifeboats available, one located 10 meters ahead of the ship in the direction of the river current, and the other is 10 meters behind the ship. If the person swims forward, they will have to catch the lifeboat as it moves forward, and if they swim backward, they will see the lifeboat behind them approaching. Which lifeboat should the person swim towards?*
Q5	**Fermi Problems**	Estimation is required without previous data or equation provided. There are many strategies to be solved, and there is no single "good answer".	*How much would it cost to pay someone to push a car the same distance it would travel consuming 1 euro of gasoline?* *(*This question beyond the force and motion concepts also can include other topics, such as energy, economy, physiology, etc.)*

After defining the questions, we employed ChatGPT version 3.5, a derivative of the GPT-3 models. We are aware that GPT-4 brings numerous enhancements over its predecessor (augmented memory more capacity, speed, and efficiency, visual input processing, and multilingual capabilities). However, we opted to utilize the GPT-3 since it is the version more widely used by being in open access, and, therefore, the results obtained can be more representative of what most users will encounter.

During this process, we presented each question ten times consecutively, using different “new chat” tabs while remaining logged into the same user account. To ensure consistency, we refrained from using specific prompts or instructions to guide the resolutions of each question. For example, we avoided providing directives like “*behave like a third-year high school student*” or “*respond as a physics doctor would*”. Additionally, we did not impose any command to limit the response length or to adopt a particular language style. Essentially, we copy and paste the text of each question into the chat interface, obtaining independent responses for each attempt.

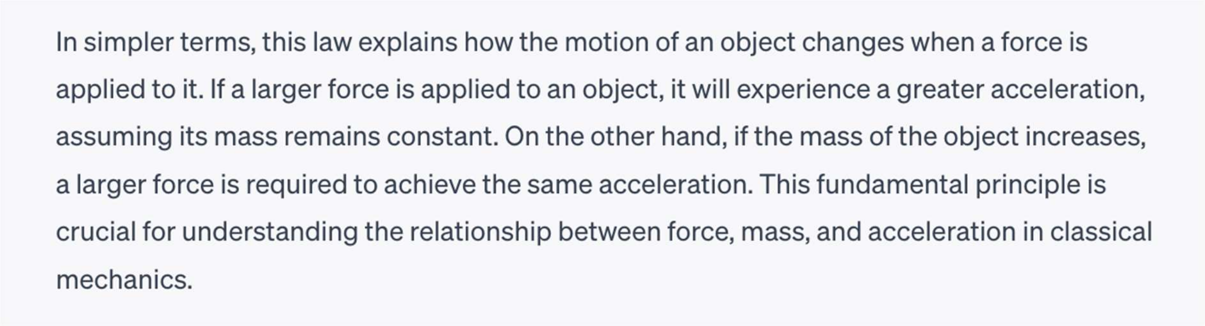
Results

Table II outlines the answers generated by ChatGPT. For questions Q2, Q3, and Q5 we included only the final value at the end of the answer. In the case of Q1 and Q5, we provided a summarized qualitative answer.

Table II. Summary of the answer by ChatGPT

	Q1	Q2	Q3	Q4	Q5
Typology	**Dictionary Definition**	**Simple Calculation**	**Multi-step calculation**	**Reasoning**	**Fermi Problems**
Expected answer	$\Sigma F = \frac{d(mv)}{dt}$	-27 m/s	11cm	He takes the same time to both boats	A value within a magnitude order of 10^3€
T1	$\Sigma F = m.a$	-27 m/s	0.7486 m	Behind	10€
T2	$\Sigma F = \frac{d(mv)}{dt}$	-27 m/s	0 m	Behind	1000€
T3	$\Sigma F = \frac{d(mv)}{dt}$	27 m/s	-0.000783	Behind	20€
T4	$\Sigma F = \frac{d(mv)}{dt} = m \cdot a$	-4.5 m/s	0.153m	Behind	5€
T5	$\Sigma F = m.a$	27 m/s	-0.0036 m	Behind	55.56€
T6	$\Sigma F = \frac{d(mv)}{dt} = m \cdot a$	-27 m/s	0.431m	Behind	3733€
T7	$\Sigma F = m.a$	-27 m/s	0.396m	Behind	200€
T8	$\Sigma F = m.a$	$\sqrt{-729}$m^2/s^2	$\sqrt{-1.5266}$N.m/m	Behind	0.01€
T9	$\Sigma F = \frac{d(mv)}{dt}$	-27 m/s	0.4648N	Ahead	950€
T10	$\Sigma F = m.a$	9 m/s	0 cm	Behind	1.67€

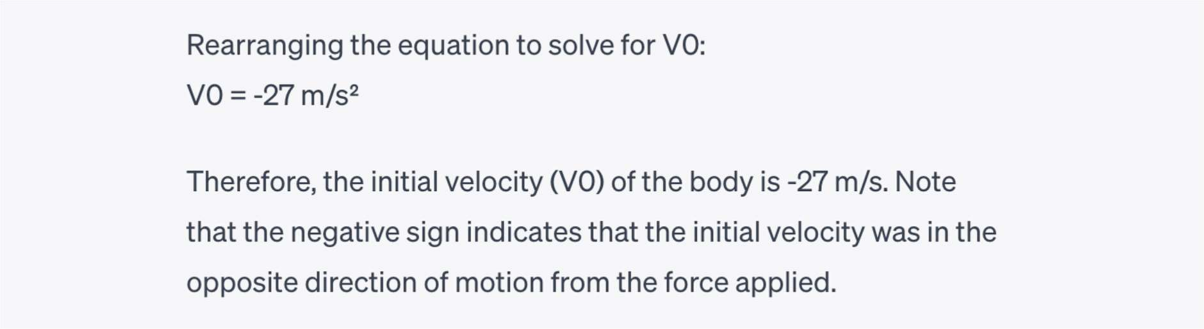
All 10 attempts for Q1 were deemed satisfactory. We didn’t identify any conceptual errors in the responses, nor did we observe significant narrative or procedural variations in the texts generated by ChatGPT. The response generally followed two approaches: (5 out of 10) those that emphasized the direct proportionality of the net force to the acceleration and the inverse proportionality to its mass; and (5 out of 10) those that focused on the rate of change of momentum concerning time. In two of the responses (Q1T4 and Q1T6) ChatGPT did both approaches, treating the proportionality between the magnitudes force and acceleration as a simpler case of the rate of change of momentum.



In simpler terms, this law explains how the motion of an object changes when a force is applied to it. If a larger force is applied to an object, it will experience a greater acceleration, assuming its mass remains constant. On the other hand, if the mass of the object increases, a larger force is required to achieve the same acceleration. This fundamental principle is crucial for understanding the relationship between force, mass, and acceleration in classical mechanics.

Fig. 1. Part of Q1's response by ChatGPT

In the case of Q2, in seven out of ten attempts for plug and chug questions, the answers are correct, with different narrative constructions that specific that the body was moving in the opposite direction of the applied force. In the three incorrect attempts (T4, T8, and T10), the AI used wrong values for the displacement, unnecessary equations, and failed to perform simple arithmetic calculations.



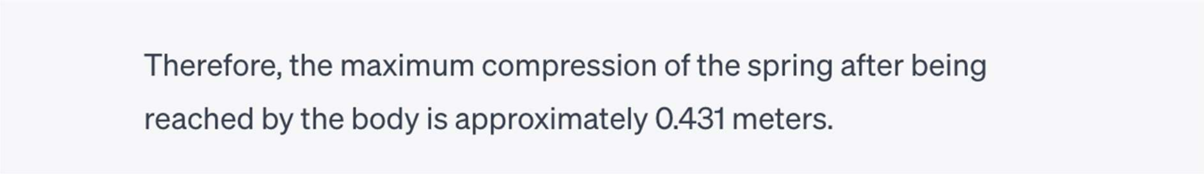
Rearranging the equation to solve for V0:

$V0 = -27 \text{ m/s}^2$

Therefore, the initial velocity (V0) of the body is -27 m/s. Note that the negative sign indicates that the initial velocity was in the opposite direction of motion from the force applied.

Fig. 2. Part of Q2's response by ChatGPT

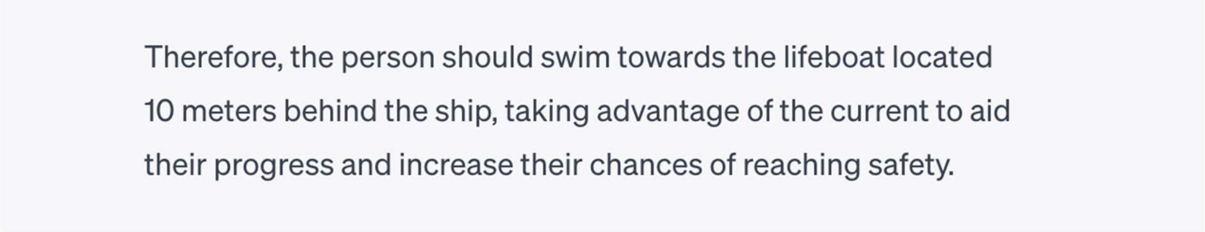
About Q3, in trials T3, T5, T7, T8, T9, and T10 there was a stoppage while ChatGPT was issuing a response, requiring us to request continuity. In all instances, the system was unable to provide a correct answer, displaying various errors, including simple calculation errors mistakes, failure in trigonometric relationships, and even not considering the friction between surfaces.



Therefore, the maximum compression of the spring after being reached by the body is approximately 0.431 meters.

Fig. 3. Part of Q3's response by ChatGPT

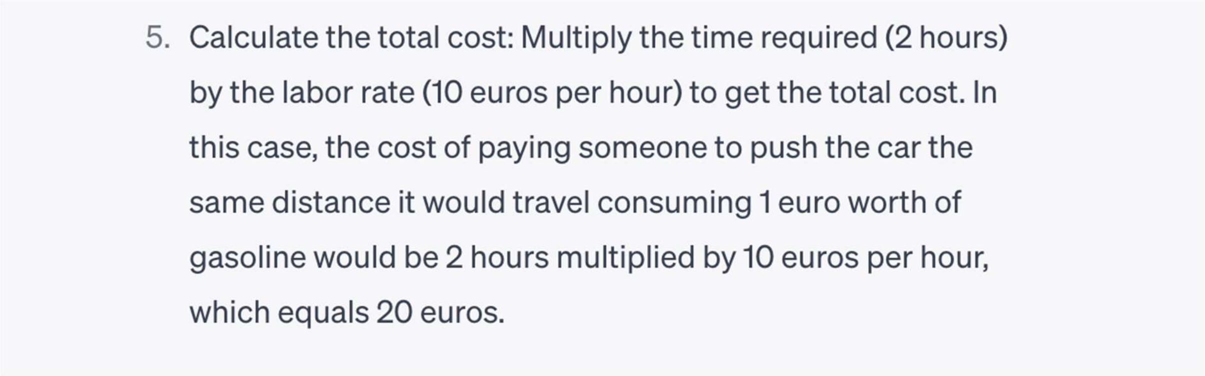
Regarding Q4, the appropriate answer is that the person can swim on both boats, respecting the principle of Galilean relativity, since both the ship and the lifeboats are at rest concerning the water frame of reference. However, ChatGPT provided the answer that the person must swim to the lifeboat behind in nine out of the ten attempts, and in one attempt, to the front. Surprisingly, this time the textual constructions presented were very similar in content and form.



Therefore, the person should swim towards the lifeboat located 10 meters behind the ship, taking advantage of the current to aid their progress and increase their chances of reaching safety.

Fig. 4. Part of Q4's response by ChatGPT

Lastly, Fermi's question like Q5, had been previously discussed by [8], concluding that the expected answer would be around an order of magnitude of 10^3€, yet without a specific correct answer with an exact value. This type of question involves strong interpretation, informed reasoning, and appropriate assumptions and approximations concerning which models and quantities to use. When presented to ChatGPT, its estimations, and inferences generated highly varied results, with some of them being closer to the expected order of magnitude than others. Many different strategies were used, as well as presented in different narrative constructions. In the resolution procedures, there were attempts to compare electric and combustion cars with the human body. Some estimations were very accurate, such as the average consumption of a combustion car, or the average price of gasoline, but in other attempts, the values given by the AI were far away from the expected, and there were also some wrong applications of physics concepts.



5. Calculate the total cost: Multiply the time required (2 hours) by the labor rate (10 euros per hour) to get the total cost. In this case, the cost of paying someone to push the car the same distance it would travel consuming 1 euro worth of gasoline would be 2 hours multiplied by 10 euros per hour, which equals 20 euros.

Fig. 5. Part of Q5's response by ChatGPT

Discussion

Summarizing in Table III, we have analyzed the answers generated by the GTP in terms of correctness and variability, as well as the main errors identified in the responses generated by the ChatGPT.

Table III: Summary of results

	Correctness of the answer	Variability of the answer	Main mistakes identified
Q1	Very High (10/10)	Low variability both in form and content	● Not identified
Q2	High (7/10)	Similar structure calculation	● Use of unnecessary equations ● Simple arithmetic errors
Q3	Very Low (0/10)	Different structure calculation	● Simple arithmetic errors ● Trigonometric errors

			- Lack of understanding of some values given in the question. - Wrong application of some physics concepts (i.e., not considering the friction)
Q4	Very Low (0/10)	Low variability both in form and content	- Wrong application of some physics concepts (i.e., not considering the Galilean relativity of movement)
Q5	Low (3/10)	A lot of variability both in form and the content	- Lack of accuracy of values generated by the AI (human velocity to push one car, average consumption of the car, efficiency of the human body, etc.). - Simple arithmetic errors - Wrong application of some physics concepts (i.e., the relationship between energy and force).

Considering that ChatGPT utilizes a probabilistic approach to generate texts and considering the nature of the questions posed, it can be concluded that both the correctness and the variability of the answers is influenced by the question's typology. ChatGPT offers suitable responses for straightforward definitions and basic calculations when all the required data is provided in the statement. However, when confronted with multi-step problems necessitating intricate conceptualization and multiple calculations involving various physical concepts and solution methods, there tends to be a higher occurrence of errors in simpler calculations.

Simultaneously, tasks demanding high-order cognitive skills from physics students, such as applying physics principles in reasoning, as seen in Q4, none of the provided answers turn out to be accurate. Furthermore, in estimation scenarios such as Q5, the provided answer is reasonable in some cases, but due to the extensive variability of responses, it often falls short. Many instances also reveal errors, inaccuracies, dubious content, and challenges in structuring solutions across different levels, which will vary depending on the objective or challenge requested.

Limitations

These results don’t allow us a generalization to all classes of questions within the proposed typologies. Firstly, because we presented only one question in each typology, employing the same prompting method. We didn’t used a systematic variation of the tasks for each task type, but only ten trials of a single task. In the interpretation of the results, we cannot exclude that the choice of context in which a task is framed or the concept that a task requires have an influence on the correctness or the variability of the output of ChatGPT. It is known that the performance of LLM strongly depends on the data it is trained on, and making the same procedure with other physics topics (such as electromagnetism, light, sound, thermodynamics, etc.) could lead a different results.

Secondly, it is important to highlight that several variables might influence responses from ChatGPT, as discussed by [3]. The defined prompts used can be misunderstood by the LLM, and a slight variation in the phrasing of the prompts can lead to different results [13]. The interpretability of our study results may be constrained by these uncertainties regarding the extent to which variations in contextual framing, conceptual aspects addressed, or phrasing could account for the observed outcomes.

Finally, we must take into account that we are discussing about technologies that are evolving rapidly and will be introduced more swiftly than any individual study could comprehensively address all nuances and specificities. Considering that ChatGPT 3.5 currently faces challenges in strategically managing the logical, conceptual, and mathematical complexities inherent to physics, it's important to acknowledge that these results might become outdated in the context of newer ChatGPT versions [11] or other generative AIs that exhibit significantly enhanced performance in reasoning tasks.

Educational implications

ChatGPT and other upcoming AI tools are powerful resources that can contribute significantly to education. Instead of adopting a simplistic approach of either endorsing or rejecting these tools in educational settings, it's crucial to address questions such as when, why, and how to utilize them effectively. In our study, we present an initial exploration that sheds light on areas where AI, like ChatGPT, might encounter challenges. We aimed to delve into the correctness and variability of the answers provided by this AI tool when faced with different typologies of physics questions. As a result, we discovered that its performance is heavily contingent on the question type. Although this AI is adept at providing clear definitions and conducting simple calculations in physics questions, educators must be aware that it still struggles to handle conceptual reasoning or estimation questions, and these types of questions require higher-order thinking skills. It's important to note, however, that we are not asserting that ChatGPT is entirely unsuitable for addressing entire categories of physics problems, given that, there is evidence from works such as [3] and [9] that suggests these types of chatbots occasionally succeed in answering questions of this nature in different contexts.

As physics educators, these findings carry various implications. On one hand, we cannot assume that ChatGPT, in its current version, can be entirely relied upon as a trustworthy self-help tool for introductory physics. While it can furnish coherent definitions, it falls short of fully elaborating multi-step problems or reasoning through complex physics questions that are counterintuitive [6]. Teachers can capitalize on ChatGPT's limitations as a catalyst for engaging educational activities by encouraging students to critically examine and discuss its shortcomings, fostering a deeper understanding of the complexities in physics problem-solving.

On the other hand, the concern that ChatGPT might encourage student cheating by automatically completing their assignments is a narrow viewpoint. This holds true not just for questions involving reasoning and estimations, but primarily because there exist more effective avenues for cheating in school physics questions [10]. By comprehending the specific tasks susceptible to cheating with ChatGPT, educators can make informed decisions when designing assignments, ensuring that tasks proposed to students not only assess their knowledge but also foster genuine understanding and critical thinking in physics education.

Lastly, while the analysis of the ChatGPT’s responses in this study points towards the need for further investigations, due to the simply expected increase in AI performance, as educators, regardless of the version or type of any AI, we will always be instigated with the challenge of curating "good questions" for our students [12]—questions that encourage them to think and reason beyond mere regurgitation of definitions.

References

[1] P. Bitzenbauer, "ChatGPT in physics education: A pilot study on easy-to-implement activities," *Contemp. Educ. Technol.* **15**(3), ep430 (2023). https://doi.org/10.30935/cedtech/13176

[2] D. MacIsaac, "Chatbots Attempt Physics Homework—ChatGPT: Chat Generative Pre-Trained Transformer," *Phys. Teach.* **61**(4), 318 (2023). https://doi.org/10.1119/10.0017700

[3] J G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?." Physical Review Physics Education Research 19(1), 010132. (2023). https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

[4] C. G. West, "AI and the FCI: Can ChatGPT project an understanding of introductory physics?." arXiv preprint arXiv:2303.01067 (2023).

[5] J. Wang, "ChatGPT: A test drive," *Am. J. Phys.* **91**(4), 255–256 (2023). https://doi.org/10.1119/5.0145897

[6] L. Viennot, *Reasoning in Physics: The Part of Common Sense* (Springer, 2001).

[7] A. Van Heuvelen, "Learning to think like a physicist: A review of research-based instructional strategies," *Am. J. Phys*. **59**(10), 891–897 (1991). https://doi.org/10.1119/1.16667

[8] L. Albarracín, V. López-Simó, and J. B. Ärlebäck, "Repensando los problemas de Fermi para la enseñanza y aprendizaje de las ciencias" *Investig. Ensino Cienc*. **26**(3), 56-68 (2021). https://doi.org/10.22600/1518-8795.ienci2021v26n3p56

[9] Dahlkemper, M. N, Lahme, S. Z., Klein, P. "How do physics students evaluate Artificial Intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT". *Phys. Rev. Phys. Educ. Res*. **19,** 010142 (2023) https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

[10] C. Ruggieri, "Students' use and perception of textbooks and online resources in introductory physics", *Phys. Rev. Phys. Educ*. Res. **16**, 020123 (2020).

[11] OpenAI, AIGPT-4 Technical Report https://arxiv.org/abs/2303.08774 (2023)

[12] C. Chin, "Teacher questioning in science classrooms: Approaches that stimulate productive thinking," *J. Res. Sci. Teach.* **44**(6), 815–843 (2007). https://doi.org/10.1002/tea.20171

[13] J. López, et al. "GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts." *Natural Language Processing Journal* 5 (2023): 100032. https://doi.org/10.1016/j.nlp.2023.100032