

# EDUCATIONAL ASSESSMENT & GENERATIVE AI

Less talk, more evidence...

---

A/Professor Joshua McGrane  
Assessment and Evaluation Research Centre  
The University of Melbourne  
Kellogg College, University of Oxford



aea '24  
EUROPE



UNIVERSITY of  
NICOSIA



MINISTRY OF EDUCATION  
SPORT AND YOUTH



Pearson



IEA

Researching education, improving learning

GA Partnership  
*Moving tests online*

AQA  
Questions matter

AQA | Global  
Assessment  
Services



CAMBRIDGE  
UNIVERSITY PRESS & ASSESSMENT





THE 'MOON LANDING'...

November 30, 2022

# Introducing ChatGPT

[Try ChatGPT ↗](#)[Download ChatGPT desktop >](#)[Learn about ChatGPT >](#)

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

ChatGPT is a sibling model to [InstructGPT](#), which is trained to follow an instruction in a prompt and provide a detailed response.

We are excited to introduce ChatGPT to get users' feedback and learn about its strengths and weaknesses. During the research preview, usage of ChatGPT is free. Try it now at [chatgpt.com](https://chatgpt.com).

The Imminence of . . .

## Grading Essays by Computer

Breakthrough? Or buncombe and ballyhoo?

You should know, after reading this careful description of efforts at the University of Connecticut to rescue the conscientious English teacher from his backbreaking burden. It is authored by the researcher whose very first computer strategy for essay grading yielded marks indistinguishable from those of experts. Mr. Page, himself a refugee from English teaching, answers questions that will occur to the skeptical educator.

By ELLIS B. PAGE

“GRADING essays by computer?” For many of us, the idea at first seems utter nonsense, to be rejected out of hand. In my experience, the fastest rejection came from an English professor at a certain well-known New England institution. He heard of the idea, digested it thoroughly, and pronounced an indignant *“Impossible!”* all within ten seconds. With further conversation, I learned that this professor knew almost nothing of the various disciplines, even linguistics, which must participate in grading by computer. For those who know more of these disciplines, and especially for those who understand computers, a mechanical essay grader at first seems a delightful toy. Upon closer inspection, and with more and more background knowledge, the notion takes on a certain fascinating inevitability: We *will* soon be grading essays by computer, and this development *will* have astonishing impact on the educational world.

The aim of this article is to persuade educators of these views: First, there is a serious need for computer grading of essays. Second, such grading is feasible, and a very promising beginning has already been made. Third, some striking improve-

ments in the quality of instruction may be foreseen as a result.

### Educational Need

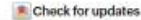
Anyone who has spent much time in the faculty room of a secondary school will realize that there is a sort of hierarchy of privilege about grading papers. Probably at the top is the teacher of mathematics or business, who can often finish his correcting during the work periods in class, and is rarely seen lugging home an oppressive load of student papers. Perhaps in the middle is the teacher of science or history, who does not have primary responsibility for the language usage of the students. This teacher has an occasional set of term papers, but not usually the constant pressure of large out-of-class grading. And at the bottom is the English teacher, who is apt to be either driven by exorbitant extra work or pursued by endless guilt.

The trouble is, almost everyone knows what students *should* do in English: preferably a daily writing stint; at least a weekly theme. Each exercise should be carefully planned, and each should be returned to the student with extensive and wise

*“Just for a moment, then, imagine what the result would be if all student essays could be turned over to a computer, which would perform a stylistic and subject-matter analysis according to the general rules desired, and deliver extensive comment and suggestion for the student to the teacher by the first bell the next day. Surely, this seems a kind of magic today, yet a beginning system of this sort may be instituted in a very few years...”* (Page, 1966, p.239)

**MORE TALK...**

# Large language models challenge the future of higher education



ChatGPT is a chatbot based on a large language model (LLM) that generates text in dialogue format. It was publicly released by OpenAI in December 2022 and has sent shockwaves through the higher education sector for its ability to create polished, confident-sounding text, which could be used to write essays and assignments. While for now it can produce answers that are only competent enough to achieve a passing mark, it is capable of correctly answering multiple-choice questions across several subject areas, including passing sample questions from high-profile licensing examinations. The rate of progress of such applications has been such that it is not difficult to imagine that a much-improved successor of ChatGPT will be released soon.

One question that arises is whether and how higher education should react. Should universities ban its use? Or should academics instead accept that language models will become integral to their professional toolkit, and incorporate them in our teaching and assessment practices?

On a practical level, allowing the use of LLM-based tools would impact the structure of assessment. And on the level of professional conduct, many share the sentiment that using text that is produced by a LLM is on a par with committing plagiarism. As universities already have harsh penalties in place to sanction plagiarism by other means, it seems natural to extend them to LLMs. A problem with this approach, however, is that it will be challenging to enforce. Unlike copy-and-pasting or paraphrasing, LLMs produce new text that is not traceable to a single source, and although software to check the likelihood of LLM-aided cheating has been released (ref. 2), their reliability appears to be low for now. Moreover, any attempt to upgrade detection software is likely to fail in the face of fast-evolving LLMs.

Another reaction by some universities has been to (at least temporarily) revert to old-fashioned pen-and-paper, invigilated examinations as their primary mode of assessment. While this solution will dramatically reduce LLM-related cheating in the short-term,

it is unlikely to be a sustainable or widely applicable one. The approach can only be used in traditional institutions where students are physically present, and it is a regressive move with respect to the digital transformations in higher education<sup>3</sup> delivery and assessment that were instigated by the global COVID-19 pandemic. Transforming written assessment into oral exams may be better suited to digital environments, yet this brings concerns of reliability, validity and scalability.

A third type of reaction to LLMs, and perhaps the only sustainable one, is to adapt and embrace them, as envisaged in a recent editorial<sup>4</sup> in this journal and consistent with the International Baccalaureate's recent announcement regarding their qualifications<sup>5</sup>. There are many possibilities to experiment and be creative with ChatGPT when teaching and assessing students. However, the adoption of ChatGPT (or similar privately owned applications) as part of standard practice raises serious risks of negative operational, financial, pedagogical and ethical consequences for universities. In particular, OpenAI is under no obligation to cater to the needs of educational institutions when it comes to maintenance and access to its model, thus creating basic operational issues if this forms part of the assessment.

The long-term pedagogical implications of accepting LLMs as learning tools also need consideration. Practising academic writing is a common way to teach and assess logical argumentation and critical thinking<sup>6</sup> (which ironically are necessary skills to evaluate a LLM's output). Foreign-language students or students who are educationally disadvantaged are likely to be the most affected, with educators placing less emphasis on learning how to craft well-written and argued texts. This could end up strengthening social divides and diminishing social mobility once students graduate and are thrown into working environments where LLMs may not be available or useful.

Another challenge concerns the trust that educators can put in the model, how it was trained and on what data. Text produced by LLMs is a reflection of patterns<sup>7</sup> in the training data. Its use in education could further

entrench representational harms in ways that are insidiously difficult to document and redress<sup>8</sup>. OpenAI made some progress in improving the accuracy of ChatGPT on factual prompts and also in moderating toxic content. However, the limits of this engineering are impossible to test, and they have come at the cost of exploiting the labour of data workers who, it has emerged<sup>9</sup>, were contracted to view and label toxic content. Educators adopting ChatGPT in their teaching would implicitly validate these harmful and extractive practices.

Finally, there should be concern about the resources that are required for running LLMs, particularly in light of hundreds of universities' net-zero and low-carbon commitments. A recent article estimates ChatGPT's daily carbon footprint to be around 23 kg CO<sub>2</sub>e, about the same as a single return trip from London to Paris on the Eurostar, but this does not include the cost of training the model. While this may appear relatively small, it will rapidly increase as the technology becomes ubiquitous. Educational institutions should, therefore, be mindful of asking students to use a model whose operation is actively contributing to the climate crisis, unless the value that can be derived from its use demonstrably exceeds the environmental cost.

Given these challenges, what can academics do? One step could be the creation of publicly funded LLMs in collaboration with open, stakeholder-led initiatives like the BigScience project. Such models could be specifically developed for educational settings, ensuring that they are auditable and transparent with regards their human and environmental costs. This will require a forward-looking vision, substantial investments and the active involvement and lobbying of educational institutions and their funders. Excitement about ChatGPT and other LLM tools foreshadows the huge political issue of who owns and sets the standards for education in the age of AI.

Silvia Miliano<sup>1</sup>, Joshua A. McGrane<sup>2</sup> & Sabina Leonelli<sup>3</sup>

<sup>1</sup>Exeter Centre for the Study of the Life Sciences (Egonis), University of Exeter, Exeter, UK. <sup>2</sup>Melbourne Graduate School



## OPEN ACCESS

EDITED BY  
Gavin T. L. Brown,  
The University of Auckland, New Zealand

REVIEWED BY  
Syamsul Nor Azian Mohamad,  
MARA University of Technology, Malaysia  
Jason M. Lodge,  
The University of Queensland, Australia  
Kim Schildkamp,  
University of Twente, Netherlands

\*CORRESPONDENCE  
Therese N. Hopfenbeck  
✉ Therese.hopfenbeck@unimelb.edu.au

RECEIVED 01 August 2023  
ACCEPTED 02 November 2023  
PUBLISHED 23 November 2023

CITATION  
Hopfenbeck TN, Zhang Z, Sun SZ,  
Robertson P and McGrane JA (2023)  
Challenges and opportunities for classroom-  
based formative assessment and AI: a  
perspective article.  
*Front. Educ.* 8:1270700.  
doi: 10.3389/educ.2023.1270700

COPYRIGHT  
© 2023 Hopfenbeck, Zhang, Sun, Robertson  
and McGrane. This is an open-access article  
distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with accepted  
academic practice. No use, distribution or  
reproduction is permitted which does not  
comply with these terms.

# Challenges and opportunities for classroom-based formative assessment and AI: a perspective article

Therese N. Hopfenbeck<sup>1,2,3\*</sup>, Zhonghua Zhang<sup>1</sup>, Sundance Zhihong Sun<sup>1</sup>, Pam Robertson<sup>1</sup> and Joshua A. McGrane<sup>1,2</sup>

<sup>1</sup>Assessment and Evaluation Research Centre, Graduate School of Education, The University of Melbourne, Parkville, VIC, Australia, <sup>2</sup>Kellogg College, University of Oxford, Oxford, England, <sup>3</sup>The University of Science and Technology, Trondheim, Norway

The integration of artificial intelligence (AI) into educational contexts may give rise to both positive and negative ramifications for teachers' uses of formative assessment within their classrooms. Drawing on our diverse experiences as academics, researchers, psychometricians, teachers, and teacher educators specializing in formative assessment, we examine the pedagogical practices in which teachers provide feedback, facilitate peer- and self-assessments, and support students' learning, and discuss how existing challenges to each of these may be affected by applications of AI. Firstly, we overview the challenges in the practice of formative assessment independently of the influence of AI. Moreover, based on the authors' varied experience in formative assessment, we discuss the opportunities that AI brings to address the challenges in formative assessment as well as the new challenges introduced by the application of AI in formative assessment. Finally, we argue for the ongoing importance of self-regulated learning and a renewed emphasis on critical thinking for more effective implementation of formative assessment in this new AI-driven digital age.

## KEYWORDS

artificial intelligence, formative assessment, self-regulation, critical thinking, classroom based assessment

## Introduction

In an era marked by rapid technological advancements, artificial intelligence (AI) is now increasingly used in diverse sectors of our society, fundamentally transforming the way we live.





# SOME CHALLENGES AND OPPORTUNITIES...

- **Challenges**

- Plagiarism

- The 'essay' is on life support

- Human cost of the models

- Environmental cost of the models

- Cultural and linguistic biases

- Who 'owns' education

- **Opportunities**

- Democratising AI

- Address longstanding issues in assessment

- Greater focus on 'higher order' skills & SRL

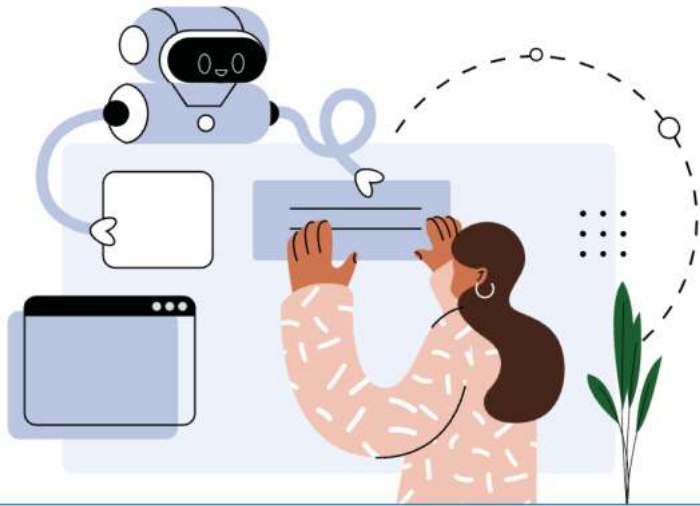
- Personalised learning

- Leveraging our collective datasets to train and fine-tune smaller, smarter and more ethical models

- Automated grading and feedback**



# Guidance for generative AI in education and research



Education  
2030



Australian Framework  
for Generative Artificial  
Intelligence in Schools



Policy paper  
**Generative artificial intelligence (AI) in  
education**  
Updated 26 October 2023



**MEANWHILE...**



edcafe

almanack

MAGIC  
SCHOOL



lessonplans.ai

Brisk  
Teaching

QUIZGECKO

Learnt.ai

# Nurture AI in Education with Microsoft

We are taking enormous leaps with our 'Nurture Assistant' to empower teachers and students with Artificial Intelligence in a safe & responsible way that gives them superpowers within Microsoft Teams.

Our Framework Is  
Research-Backed By



"Feedback is critical to improving learning as it both influences students' motivation to learn and their ability to do so" ([Hattie 1999](#)).

nurture

Research  
Framework

+



OpenAI

=

Unique,  
pedagogically  
sound EdTech

# Advice on using artificial intelligence tools for student assessment and feedback

## Advice on using artificial intelligence tools for student assessment and feedback

### Summary

[Background to these guidelines](#)

[Staff responsibility, roles and pedagogical use](#)

[Managing risks](#)

## Summary

The University is updating its guidelines relating to staff use of new Generative AI (GenAI) tools to assist with assessing students' work or providing feedback on students' work. These new guidelines seek to achieve a balance between managing the risks associated with the use of AI tools for feedback and assessment, while supporting innovation in AI use where this provides benefits to students and staff.

From Semester 1, 2024, staff can use new AI systems to support their evaluation of students' work and to provide feedback to students. However, staff remain responsible for any academic judgements made on students' submitted work and any feedback provided to students.

The outputs from any AI system used for assessment or feedback must be reviewed by staff and the prompts and inputs the AI system is using must be well understood and managed.

**HOW ABOUT THE EVIDENCE...**

# CAN LARGE LANGUAGE MODELS GRADE?

<https://doi.org/10.1007/s40593-024-00431-z>

ARTICLE



## Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset

Owen Henkel<sup>1</sup> · Libby Hills<sup>2</sup> · Bill Roberts<sup>3</sup> · Joshua McGrane<sup>4</sup>

Accepted: 19 September 2024

© International Artificial Intelligence in Education Society 2024

### Abstract

Formative assessment plays a critical role in improving learning outcomes by providing feedback on student mastery. Open-ended questions, which require students to produce multi-word, nontrivial responses, are a popular tool for formative assessment as they provide more specific insights into what students do and do not know. However, grading open-ended questions can be time-consuming and susceptible to errors, leading teachers to resort to simpler question formats or conduct fewer formative assessments. While there has been a longstanding interest in automating the grading of short answer questions, previous approaches have been technically complex, limiting their use in formative assessment contexts. The newest generation of Large Language Models (LLMs) potentially makes grading short answer questions more feasible, as the models are lexible and easier to use. This paper addresses the lack of empirical research on the potential of the newest generation LLMs for grading of short answer questions in two ways. First, it introduces a novel dataset of short answer reading comprehension questions, drawn from a battery of reading assessments conducted with over 150 students in Ghana. This dataset allows for the evaluation of LLMs in a new context, as they are predominantly designed and trained on data from high-income countries. Second, the paper empirically evaluates how well various configurations of generative LLMs can grade student short answer responses compared to expert human raters. The findings show that GPT-4, with minimal prompt engineering, performed extremely well on evaluating the novel dataset (Cohen's  $\kappa = 0.87$ ), matching performance with expert human raters. The...



- Rising & Oxford African Reading Comprehension Short Answer (ROARs) Dataset
- 162 students in Ghana
- 9-18 years old, average age 13
- 67% female
- 1068 answers to a set of reading comprehension questions adapted from pre-PIRLS
  - Triple scored by humans
- 72% written, 28% spoken (and transcribed)
- Simple model prompting approach
  - Zero vs. few shot prompting



## EXAMPLE TASK

<b>Passage</b>
<p>“The river is flooding,” said the giraffe. “A wall of water is racing down the valley and will soon be here.”</p> <p>“What can we do?” asked the gazelle. “It’s too late to run away.” “Climb up here,” called the monkey from the treetops. “The river won’t reach the high branches.” The animals raced to the trees. But some of them could not climb up the slippery tree trunks. Their hooves and tails were not made for climbing.</p>
<b>Question</b>
Why were the animals trying to climb to the treetops?
<b>Student Answer 1</b>
<p>The reason why they were trying to climb to the treetops is because the river <del>can't</del><sup>won't</sup> reach the high branches</p>

# CAN LARGE LANGUAGE MODELS GRADE?

<https://doi.org/10.1007/s40593-024-00431-z>

ARTICLE



## Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset

Owen Henkel<sup>1</sup> · Libby Hills<sup>2</sup> · Bill Roberts<sup>3</sup> · Joshua McGrane<sup>4</sup>

Accepted: 19 September 2024

© International Artificial Intelligence in Education Society 2024

### Abstract

Formative assessment plays a critical role in improving learning outcomes by providing feedback on student mastery. Open-ended questions, which require students to produce multi-word, nontrivial responses, are a popular tool for formative assessment as they provide more specific insights into what students do and do not know. However, grading open-ended questions can be time-consuming and susceptible to errors, leading teachers to resort to simpler question formats or conduct fewer formative assessments. While there has been a longstanding interest in automating the grading of short answer questions, previous approaches have been technically complex, limiting their use in formative assessment contexts. The newest generation of Large Language Models (LLMs) potentially makes grading short answer questions more feasible, as the models are lexible and easier to use. This paper addresses the lack of empirical research on the potential of the newest generation LLMs for grading of short answer questions in two ways. First, it introduces a novel dataset of short answer reading comprehension questions, drawn from a battery of reading assessments conducted with over 150 students in Ghana. This dataset allows for the evaluation of LLMs in a new context, as they are predominantly designed and trained on data from high-income countries. Second, the paper empirically evaluates how well various configurations of generative LLMs can grade student short answer responses compared to expert human raters. The findings show that GPT-4, with minimal prompt engineering, performed extremely well on evaluating the novel dataset (Cohen's  $\kappa = .87$ ), matching performance with expert human raters. The...



- GPT-4 strikingly outperformed GPT-3.5-Turbo in inter-rater reliability with human scoring
- GPT-4 with few shot prompting achieved a Linear Weighted Kappa of .87 (two-class scoring) and Quadratic Weighted Kappa of .91 (three-class scoring)
  - Zero shot prompting achieved .83 and .86
- GPT-4 equally accurate and reliable compared to human scoring
- Preliminary analysis showed no indication of bias in terms of students' demographic factors

# CAN LARGE LANGUAGE MODELS GRADE?

<https://doi.org/10.1007/s40593-024-00431-z>

ARTICLE



## Can LLMs Grade Open Response Reading Comprehension Questions? An Empirical Study Using the ROARs Dataset

Owen Henkel<sup>1</sup> · Libby Hills<sup>2</sup> · Bill Roberts<sup>3</sup> · Joshua McGrane<sup>4</sup>

Accepted: 19 September 2024

© International Artificial Intelligence in Education Society 2024

### Abstract

Formative assessment plays a critical role in improving learning outcomes by providing feedback on student mastery. Open-ended questions, which require students to produce multi-word, nontrivial responses, are a popular tool for formative assessment as they provide more specific insights into what students do and do not know. However, grading open-ended questions can be time-consuming and susceptible to errors, leading teachers to resort to simpler question formats or conduct fewer formative assessments. While there has been a longstanding interest in automating the grading of short answer questions, previous approaches have been technically complex, limiting their use in formative assessment contexts. The newest generation of Large Language Models (LLMs) potentially makes grading short answer questions more feasible, as the models are lexible and easier to use. This paper addresses the lack of empirical research on the potential of the newest generation LLMs for grading of short answer questions in two ways. First, it introduces a novel dataset of short answer reading comprehension questions, drawn from a battery of reading assessments conducted with over 150 students in Ghana. This dataset allows for the evaluation of LLMs in a new context, as they are predominantly designed and trained on data from high-income countries. Second, the paper empirically evaluates how well various configurations of generative LLMs can grade student short answer responses compared to expert human raters. The findings show that GPT-4, with minimal prompt engineering, performed extremely well on evaluating the novel dataset (Cohen's  $\kappa = 0.87$ ), matching responses with expert human raters. The...



- Subsequently conducted further analysis on scoring bias (to be published)
- Generalised Linear Mixed Effect modelling
  - Crossed and nested random effects by student, question and reading passage
- No significant effects of gender, age or home language
- GPT was significantly more accurate for written responses than transcribed spoken responses

# CAN LLMS GRADE MORE COMPLEX TASKS?



The screenshot shows the Kaggle interface for the competition 'The Hewlett Foundation: Short Answer Scoring'. The page includes a search bar, navigation tabs (Overview, Data, Code, Models, Discussion, Leaderboard, Rules, Team), and a main content area with text describing the competition and prizes. A QR code is visible at the bottom right of the page.

**The Hewlett Foundation: Short Answer Scoring**

**Overview** Data Code Models Discussion Leaderboard Rules Team

The William and Flora Hewlett Foundation (Hewlett Foundation) is sponsoring the Automated Student Assessment Prize (ASAP) in hopes of discovering new tools to support schools and teachers. The competition aspires to solve the problem of the high cost and the slow turnaround of hand scoring thousands of written responses in standardized tests. As a result many schools exclude written responses in favor of multiple-choice questions, which are less able to assess students' critical reasoning and writing skills. ASAP has been designed to help determine whether computerized systems are capable of grading written content accurately for schools and teachers to adopt those solutions. ASAP aspires to inform key decision makers, who are already considering adopting these systems, by delivering a fair, impartial and open series of trials to test current capabilities and to drive greater awareness when outcomes warrant further consideration.

Critical reasoning is one of a suite of skills that experts believe students must be taught to succeed in the new century. The Hewlett Foundation makes grants to educators and nonprofit organizations in support of these skills, which it calls "deeper learning." They include the mastery of core academic content, critical reasoning and problem solving, working collaboratively, communicating effectively, and learning how to learn independently. With ASAP, Hewlett is appealing to data scientists to help solve an important problem in the field of educational testing.

Hewlett is sponsoring the following prizes as part of Phase Two:

\$50,000: 1 <sup>st</sup> place
\$25,000: 2 <sup>nd</sup> place
\$15,000: 3 <sup>rd</sup> place



- The Hewlett Foundation Short Answer Scoring competition dataset (ASAP-SAS)
- 10 questions taken from Grade 10 standardised assessments in the USA
  - 3 Science, 2 Biology, and 5 English
- Approximately 1600 responses per task
  - Analysis included a stratified random sample of 600 responses per task
- Two human scores and no demographic information for students
- Holistic scoring
  - 0 - 3 for Biology and two Science tasks
  - 0 - 2 for English tasks and one Science task

# MORE COMPLEX PROMPTING



The screenshot shows the Kaggle website interface. On the left is a navigation sidebar with the Kaggle logo and links for 'Create', 'Home', 'Competitions', 'Datasets', 'Models', 'Code', 'Discussions', 'Learn', 'More', 'Your Work', and 'View Active Events'. The main content area features a search bar and a title 'The Hewlett Foundation: Short Answer Scoring'. Below the title are tabs for 'Overview', 'Data', 'Code', 'Models', 'Discussion', 'Leaderboard', 'Rules', and 'Team'. The 'Overview' tab is active, displaying a detailed description of the competition: 'The William and Flora Hewlett Foundation (Hewlett Foundation) is sponsoring the Automated Student Assessment Prize (ASAP) in hopes of discovering new tools to support schools and teachers. The competition aspires to solve the problem of the high cost and the slow turnaround of hand scoring thousands of written responses in standardized tests. As a result many schools exclude written responses in favor of multiple-choice questions, which are less able to assess students' critical reasoning and writing skills. ASAP has been designed to help determine whether computerized systems are capable of grading written content accurately for schools and teachers to adopt those solutions. ASAP aspires to inform key decision makers, who are already considering adopting these systems, by delivering a fair, impartial and open series of trials to test current capabilities and to drive greater awareness when outcomes warrant further consideration.' Below this is a paragraph about 'Critical reasoning' and a table of prizes. A QR code is located at the bottom right of the page.

Prize	Amount	Rank
Hewlett is sponsoring the following prizes as part of Phase Two:	\$50,000:	1 <sup>st</sup> place
	\$25,000:	2 <sup>nd</sup> place
	\$15,000:	3 <sup>rd</sup> place

- RISE Framework for prompt construction
  - Role - define the AI's role and the perspective it should take toward the prompt
  - Input - provide with relevant information and data for the AI to consider while producing its response
  - Steps - outline the series of actions that the AI should undertake while addressing the prompt
  - Expectation - describe the desired output
- A mix of zero, one and few-shot prompting
- Systematically varied the models' temperature setting

## A MIX OF LLMS



- Closed-weights, OpenAI models
  - GPT-3.5-Turbo
  - GPT-4-Turbo
  - GPT-4o-mini



- Open-weights, Meta models
  - Llama-3.1-8b
  - Llama-3.1-70b
  - Llama-3.1-405b

# TASK 10 - SCIENCE

## Prompt—Doghouse Item

Brandi and Jerry did the following controlled experiment to find out how the color of an object affects its temperature.

**Question:** What is the effect of different lid colors on the air temperature inside a glass jar exposed to a lamp?

**Hypothesis:** The darker the lid color, the greater the increase in air temperature in the glass jar, because darker colors absorb more energy.

### Materials:

glass jar  
lamp  
four colored lids: black, dark gray, light gray, and white  
thermometer  
meterstick  
stopwatch

### Controlled Experiment Setup

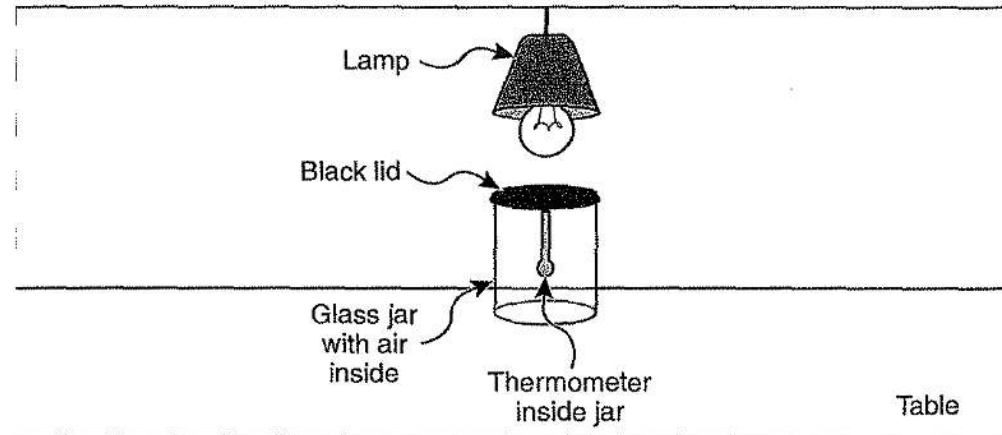


Diagram not to scale

### Procedure:

1. Put the black lid with the attached thermometer on the glass jar.
2. Make sure the starting temperature inside the jar is 24° C.
3. Place lamp 5 centimeters away from the lid and turn on the lamp.
4. After 10 minutes measure the air temperature inside the glass jar and record as Trial 1.
5. Turn off lamp and wait until the air in the jar returns to the starting temperature.
6. Repeat steps 2 through 5 for Trials 2 and 3.
7. Repeat steps 1 through 6 for the dark gray, light gray, and white lids.
8. Calculate and record the average air temperature for each lid color.

### Data:

Lid Color vs. Air Temperature Inside Glass Jar

Lid Color	Air Temperature Inside Glass Jar After 10 Minutes (° C)			
	Trial 1	Trial 2	Trial 3	Average
Black	54	52	54	53
Dark gray	48	48	48	48
Light gray	44	45	46	45
White	42	43	41	42

Note: Starting temperature was 24° C for every trial.

Brandi and Jerry were designing a doghouse. Use the results from the experiment to describe the best paint color for the doghouse.

In your description, be sure to:

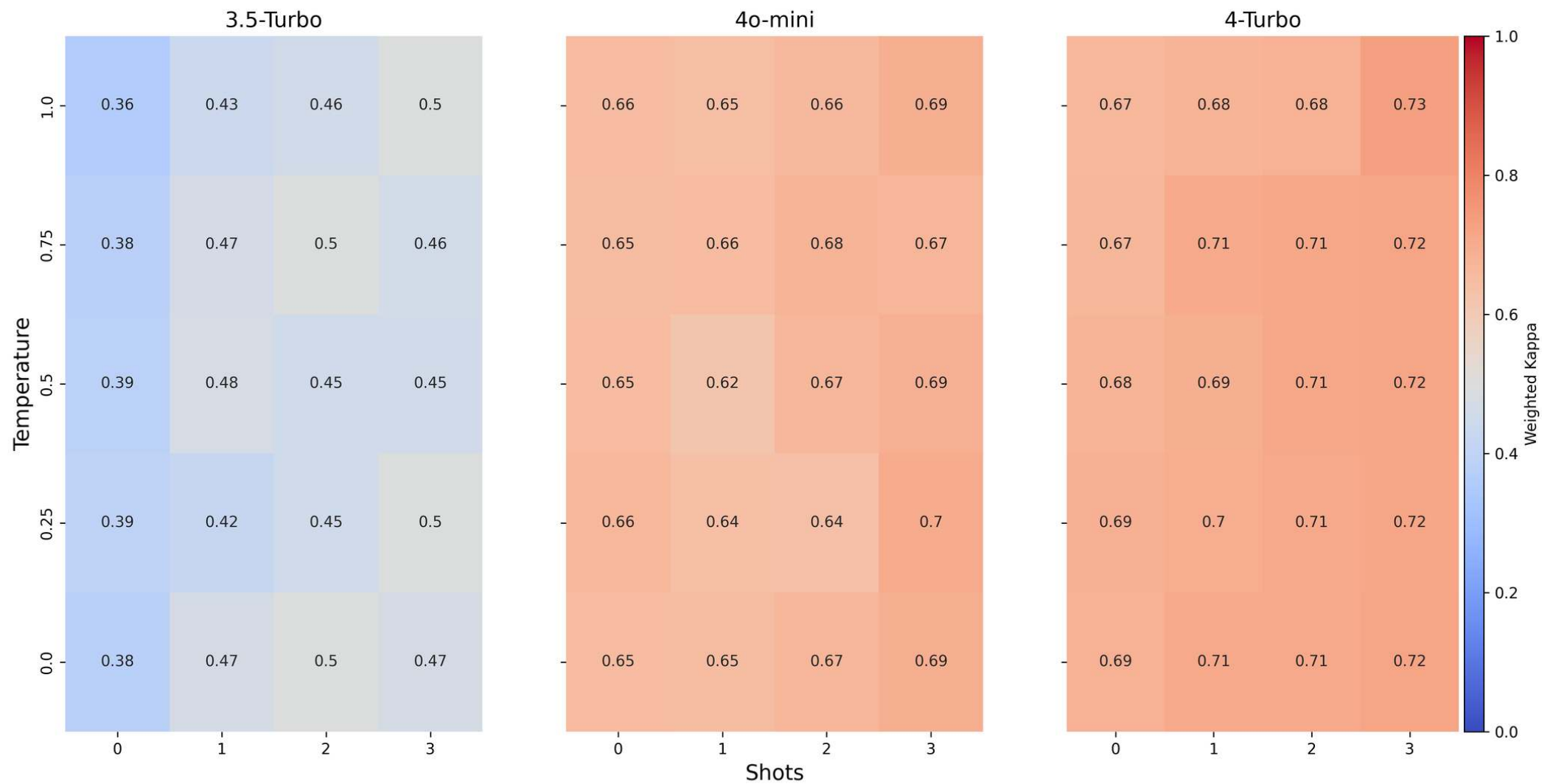
- Choose a paint color.
- Describe how that color might affect the inside of the doghouse.
- Use results from the experiment to support your description.

Choose a color:

- Black                       Dark gray                       Light gray                       White

# TASK 10 - GPT MODELS - QWK

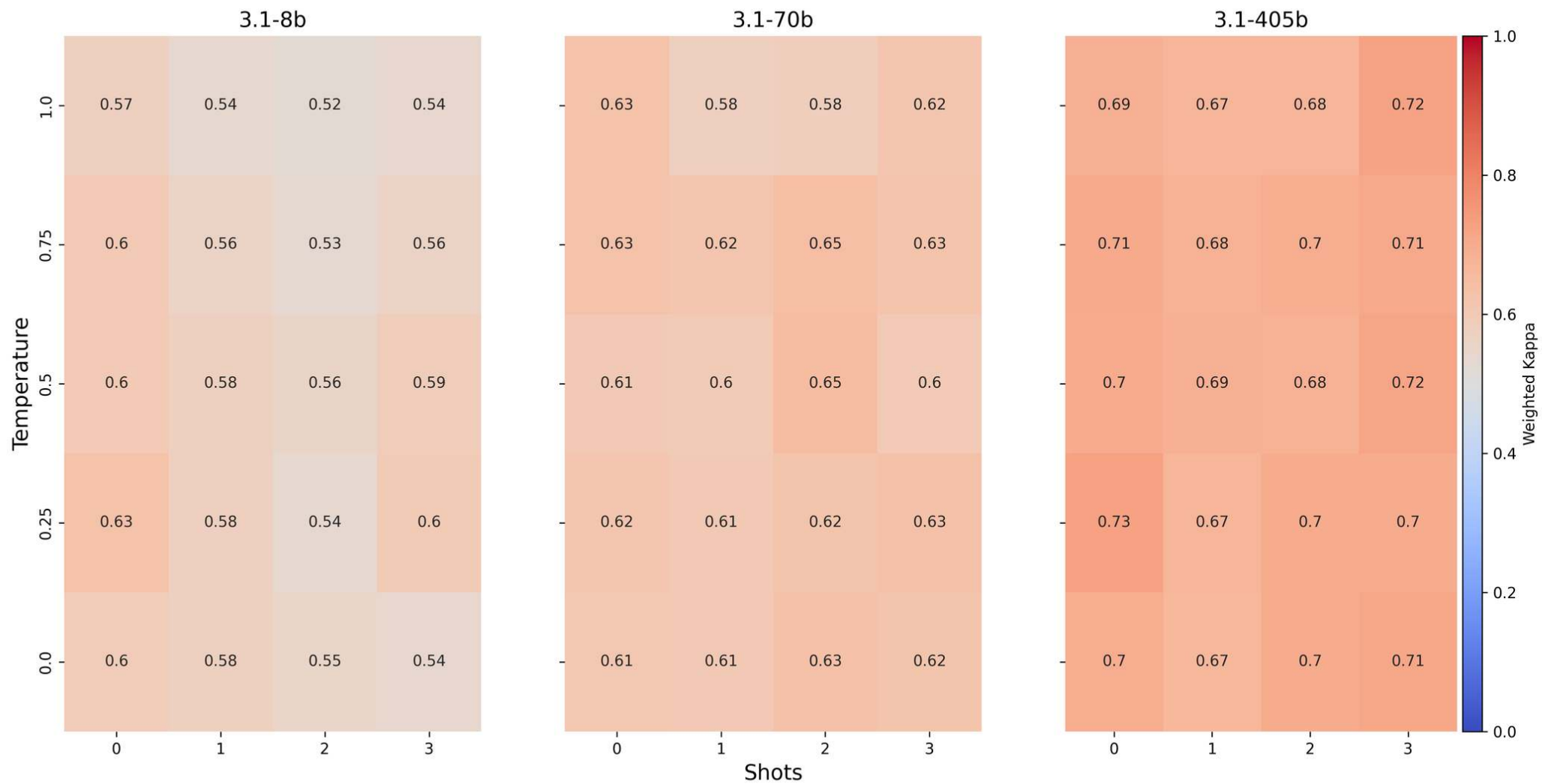
Task 10 - GPT models





# TASK 10 - LLAMA MODELS - QWK

Task 10 - LLaMa models



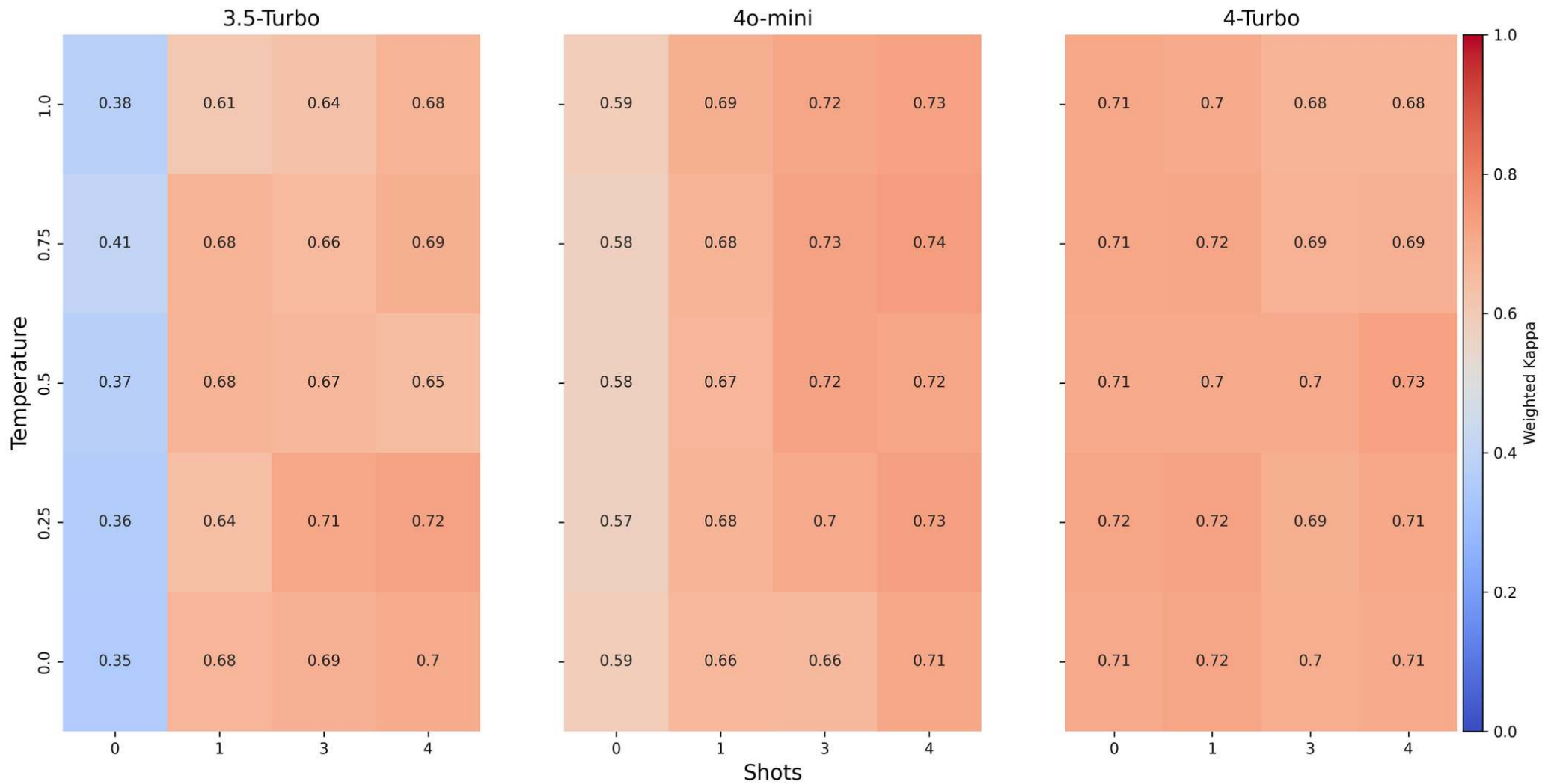
## TASK 5 - BIOLOGY

*Prompt—Protein Synthesis Item*

Starting with mRNA leaving the nucleus, list and describe four major steps involved in protein synthesis.

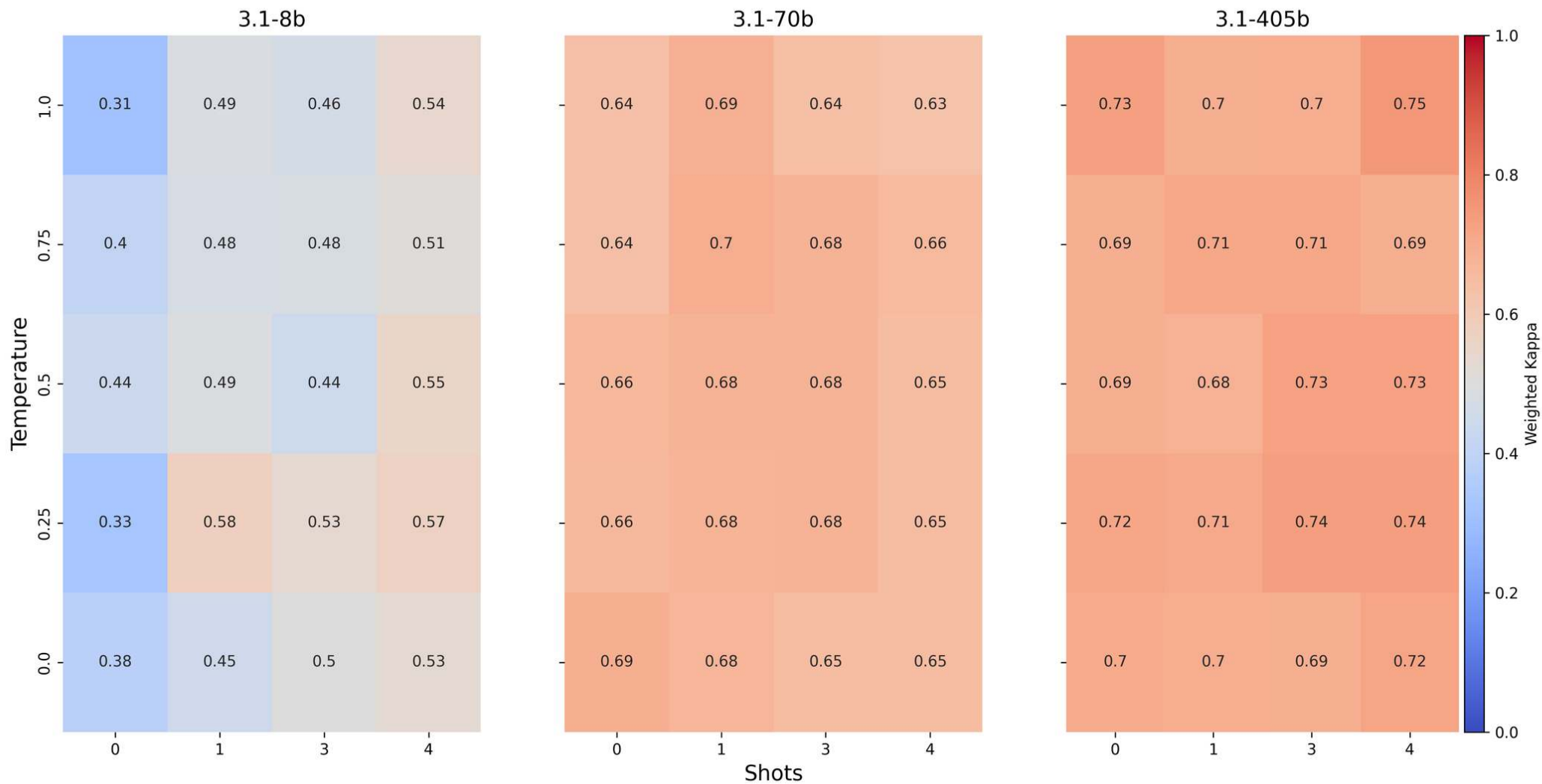
# TASK 5 - GPT MODELS - QWK

Task 5 - GPT models



# TASK 5 - LLAMA MODELS - QWK

Task 5 - LLaMa models



# TASK 9 - ENGLISH

## Reading Passage—Organization of Article Item

### Orbiting Junk

Grab your telescope! Look up in the sky! It's a comet! It's a meteor!

It's a . . . tool bag?

Such an observation isn't as strange as it seems. Orbital pathways around our planet that were once clear are now cluttered with the remains of numerous space exploration and satellite missions. This "space junk" is currently of great concern to government space agencies around the globe.

### What Is Space Junk?

In 1957, the Soviet Union launched the first artificial satellite. The United States followed suit, and thus began the human race's great space invasion.

Over the past 52 years, a variety of spacecraft, including space capsules, telescopes, and satellites, have been sent beyond Earth's atmosphere. They explore the vast reaches of our solar system, monitor atmospheric conditions, and make global wireless communication possible. The rockets that are used to power these spacecraft typically fall back to Earth and disintegrate in the intense heat that results from friction with Earth's atmosphere. The objects themselves, however, are positioned hundreds of miles above Earth, far from elements that would cause them to degrade or burn up. In this airless environment, some of them continue to circle the planet indefinitely. While this is ideal for a fully functioning object that was launched for that purpose—for example, a communications satellite—what happens when a satellite "dies" or malfunctions and can't be repaired? The disabled object becomes a piece of high-tech junk, circling the globe in uncontrolled orbit.

### Crash Course

With no one at the controls, dead satellites run the risk of colliding with each other. That's exactly what happened in February 2009. Two communications satellites, one American and one Russian, both traveling at more than 20,000 miles per hour, crashed into each other 491 miles above the Earth. The impact created hundreds of pieces of debris, each assuming its own orbital path. Now, instead of two disabled satellites, there are hundreds of microsattellites flying through space.

It's not only spectacular crashes that create debris. Any objects released into space become free-orbiting satellites, which means that astronauts must take great care when they leave their spacecraft to make repairs or do experiments. Still, accidents do happen: in 2008, a tool bag escaped from the grip of an astronaut doing repairs on the International Space Station (ISS).

### Little Bits, But a Big Deal

So who cares about a lost tool bag or tiny bits of space trash?

Actually, many people do. Those bits of space debris present a very serious problem. Tiny fragments traveling at a speed of five miles per second can inflict serious damage on the most carefully designed spacecraft. If you find that hard to believe, compare grains of sand blown by a gentle breeze to those shot from a sandblaster to strip paint from a concrete wall. At extreme speeds, little bits can pack a punch powerful enough to create disastrous holes in an object moving through space.

Scientists are hard-pressed for an easy solution to the problem of space junk. Both the National Aeronautics and Space Agency (NASA) and the European Space Agency maintain catalogues of known objects. The lost tool bag, for example, is listed as Satellite 33442. But while military radar can identify objects the size of a baseball, anything smaller goes undetected. This makes it difficult for spacecraft to steer clear of microdebris fields. Accepting the inevitability of contact, engineers have added multiple walls to spacecraft and stronger materials to spacesuits to diminish the effects of impact.

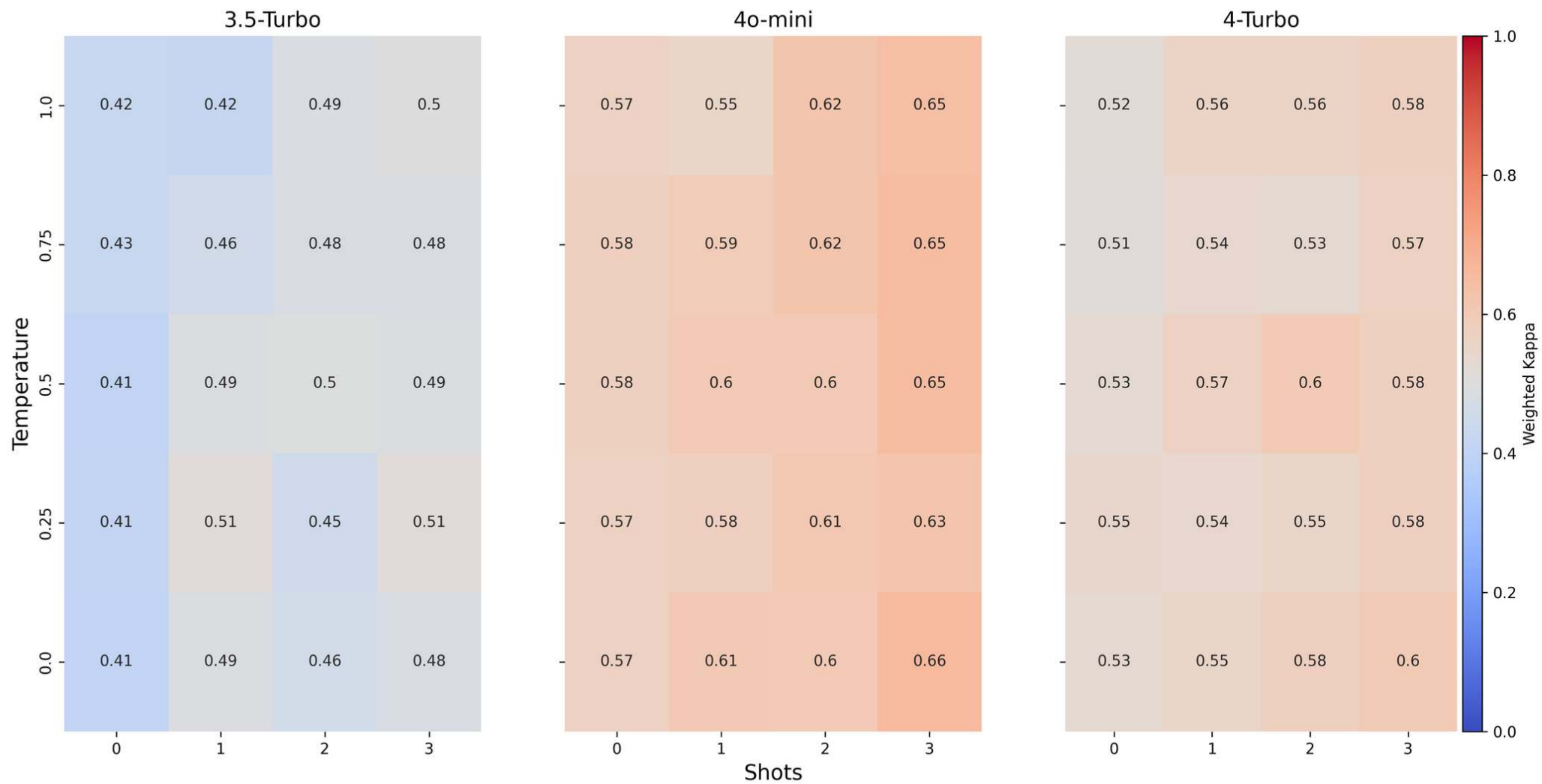
Yet the problem is certain to persist. In fact, the amount of space trash is actually increasing because commercial space travel is on the rise and more nations have undertaken space exploration. Space agencies hope that the corporations and nations involved can work together to come up with a viable solution to space pollution.

## Prompt—Organization of Article Item

How does the author organize the article? Support your response with details from the article.

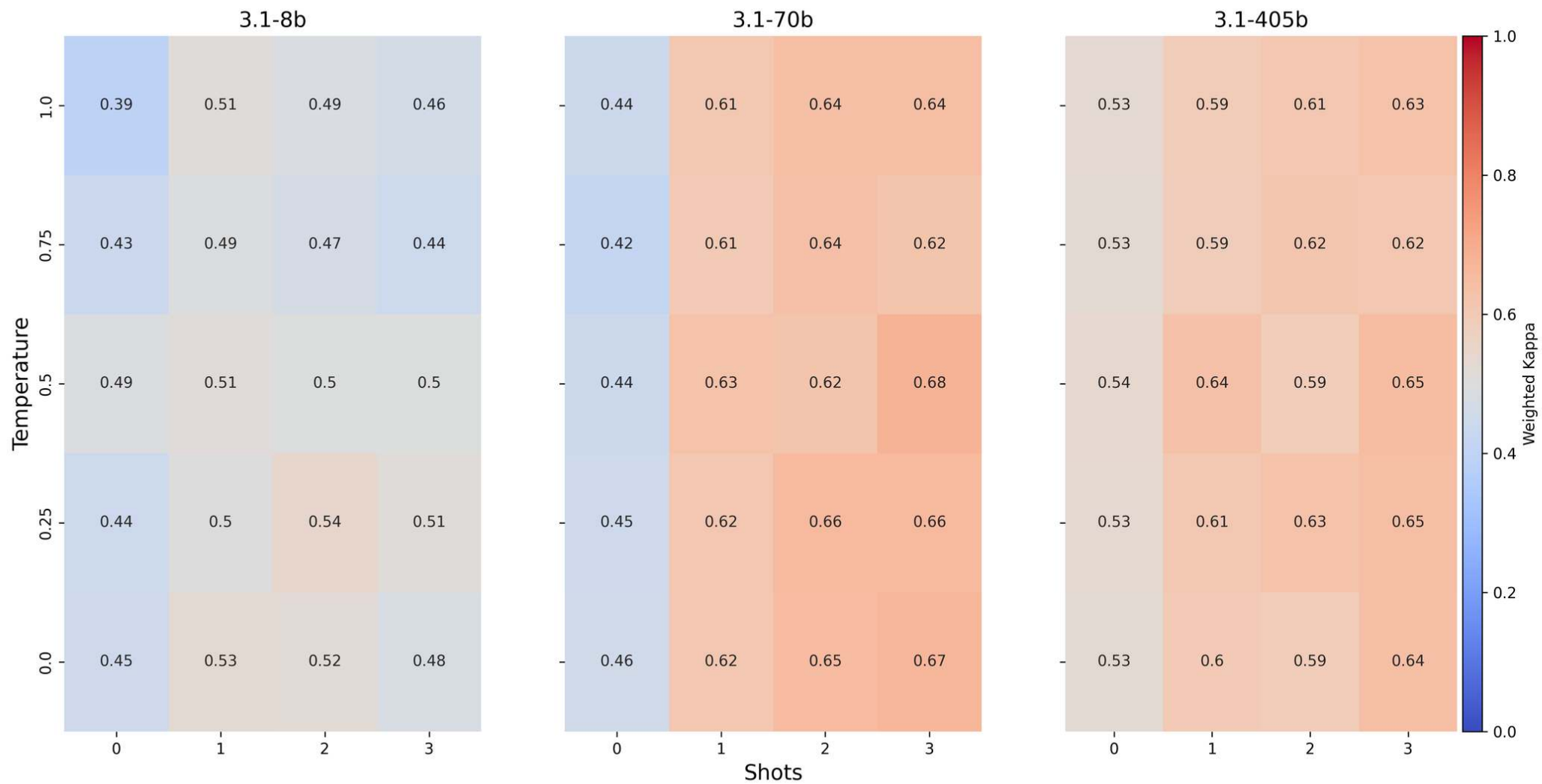
# TASK 9 - GPT MODELS - QWK

Task 9 - GPT models



# TASK 9 - LLAMA MODELS - QWK

Task 9 - LLaMa models



# SUMMARY OF FINDINGS



The screenshot shows the Kaggle interface for the competition "The Hewlett Foundation: Short Answer Scoring". The left sidebar contains navigation links: Create, Home, Competitions (selected), Datasets, Models, Code, Discussions, Learn, More, Your Work, and Viewed. The main content area has a search bar and tabs for Overview, Data, Code, Models, Discussion, Leaderboard, Rules, and Team. The Overview tab is active, displaying a description of the competition and a table of prizes.

**The Hewlett Foundation: Short Answer Scoring**

**Overview** Data Code Models Discussion Leaderboard Rules Team

The William and Flora Hewlett Foundation (Hewlett Foundation) is sponsoring the Automated Student Assessment Prize (ASAP) in hopes of discovering new tools to support schools and teachers. The competition aspires to solve the problem of the high cost and the slow turnaround of hand scoring thousands of written responses in standardized tests. As a result many schools exclude written responses in favor of multiple-choice questions, which are less able to assess students' critical reasoning and writing skills. ASAP has been designed to help determine whether computerized systems are capable of grading written content accurately for schools and teachers to adopt those solutions. ASAP aspires to inform key decision makers, who are already considering adopting these systems, by delivering a fair, impartial and open series of trials to test current capabilities and to drive greater awareness when outcomes warrant further consideration.

Critical reasoning is one of a suite of skills that experts believe students must be taught to succeed in the new century. The Hewlett Foundation makes grants to educators and nonprofit organizations in support of these skills, which it calls "deeper learning." They include the mastery of core academic content, critical reasoning and problem solving, working collaboratively, communicating effectively, and learning how to learn independently. With ASAP, Hewlett is appealing to data scientists to help solve an important problem in the field of educational testing.

Hewlett is sponsoring the following prizes as part of Phase Two:

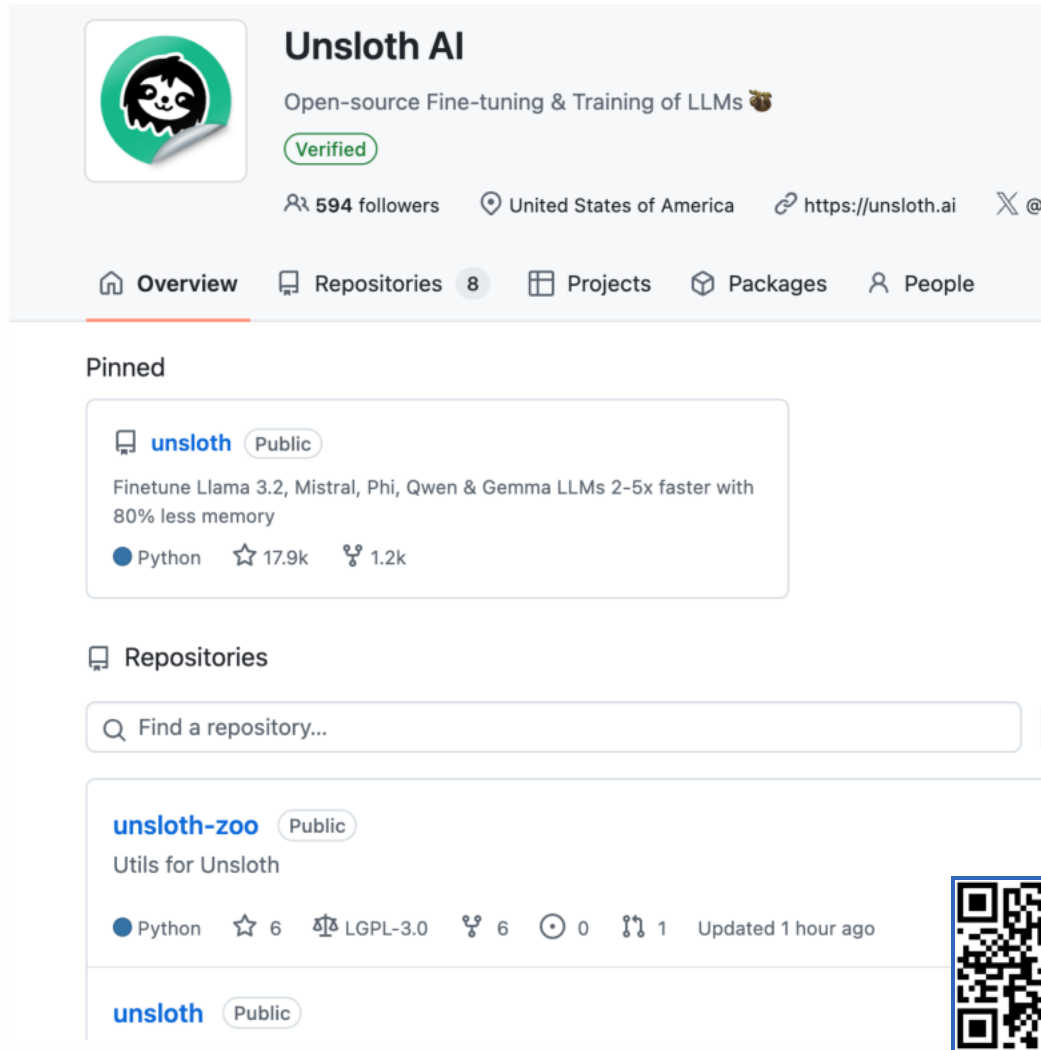
\$50,000:	1 <sup>st</sup> place
\$25,000:	2 <sup>nd</sup> place
\$15,000:	3 <sup>rd</sup> place



- There were no overall significant differences in accuracy and reliability between GPT-4-Turbo, GPT-4o-mini & Llama-3.1-405b
  - GPT-3.5-Turbo, Llama-3.1-8b & Llama-3.1-70b were significantly worse
- QWKs varied by subject area for the best performing model combinations:
  - .68 to .72 for the Science tasks
  - .73 to .76 for the Biology tasks
  - .59 to .68 for the English tasks
- Few-shot prompting with all possible score points exemplified led to significantly higher accuracy



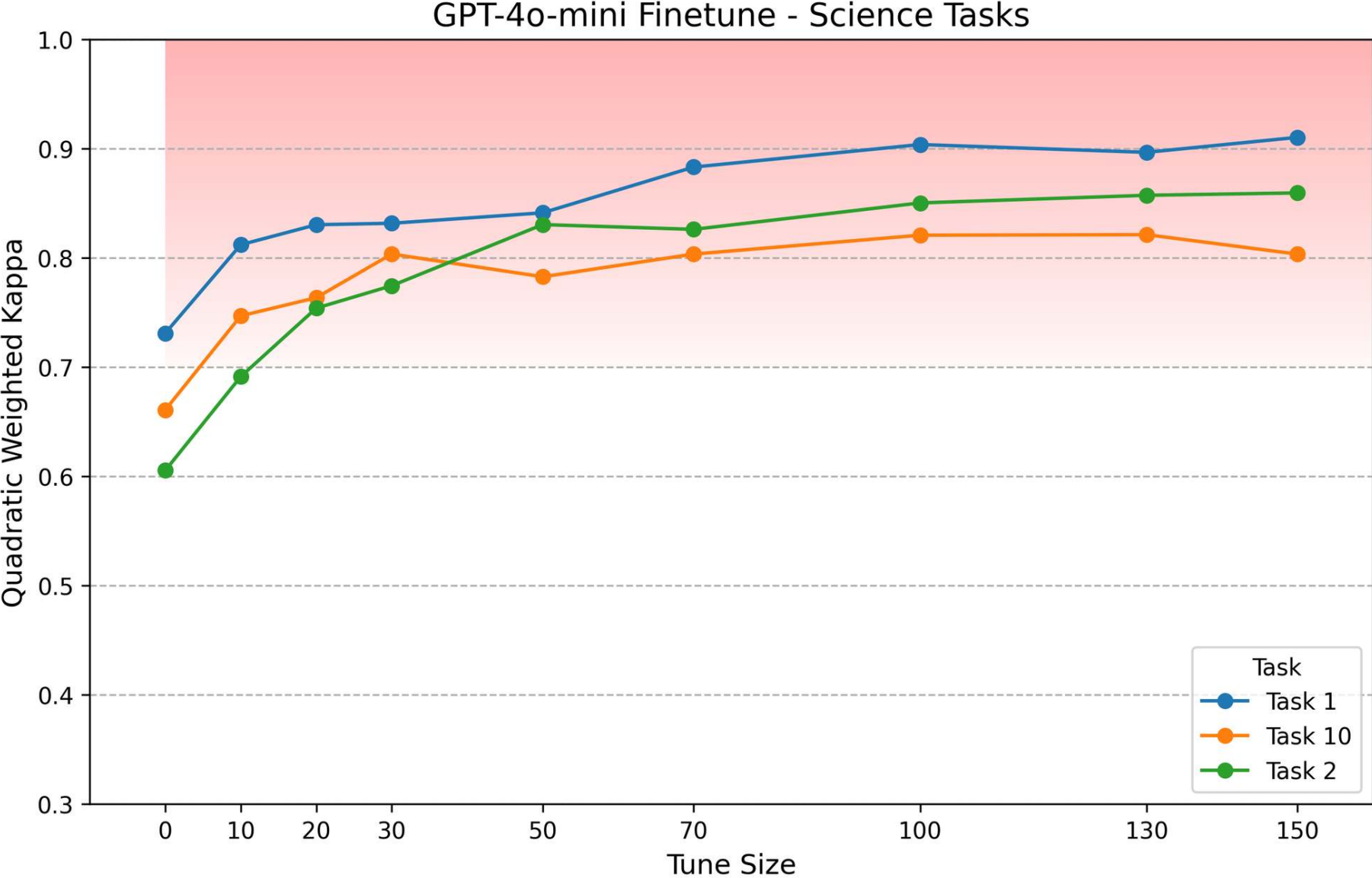
# CAN WE 'TUNE' OUR WAY THERE...



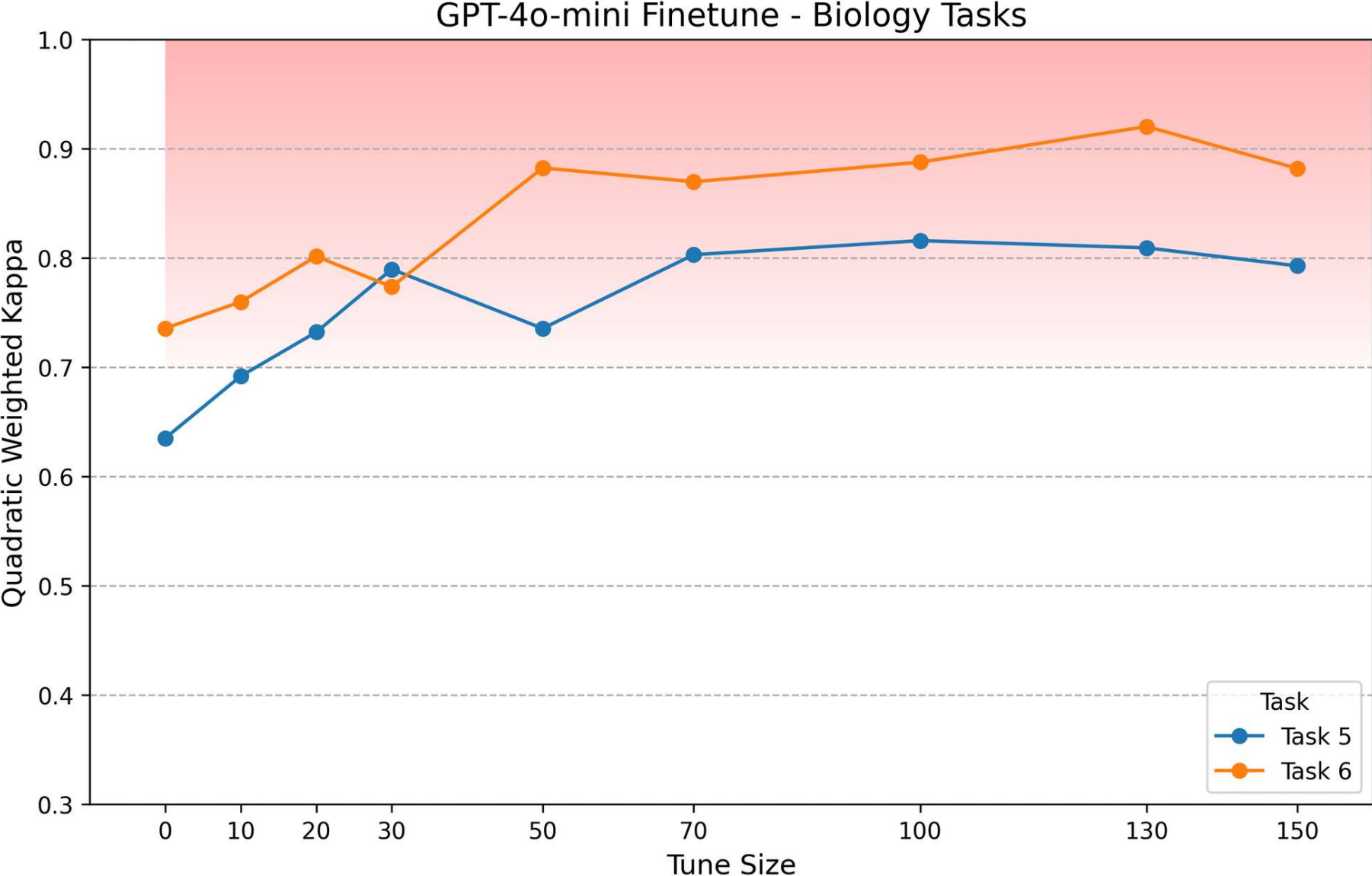
The screenshot shows the GitHub profile for Unloth AI. The profile includes a profile picture of a cartoon sloth, the name 'Unloth AI', and a bio: 'Open-source Fine-tuning & Training of LLMs 🦥'. It is a verified account with 594 followers, located in the United States of America, and has a website link to https://unsloth.ai. The navigation bar shows 'Overview' as the active tab, along with 'Repositories' (8), 'Projects', 'Packages', and 'People'. Under the 'Pinned' section, there is a repository card for 'unsloth' (Public) with the description 'Finetune Llama 3.2, Mistral, Phi, Qwen & Gemma LLMs 2-5x faster with 80% less memory'. It has 17.9k stars and 1.2k forks. Below this, the 'Repositories' section is visible with a search bar. The first repository listed is 'unsloth-zoo' (Public), described as 'Utils for Unloth', with 6 stars, 6 forks, 0 issues, and 1 pull request, updated 1 hour ago. A QR code is located at the bottom right of the repository list.

- Fine-tuned both the GPT-4o-mini and Llama-3.1-8b models with varying exemplar dataset sizes
  - 10 examples per score point through to 150 examples per score point
- Llama fine-tuning was done using the rank-stabilised LORA method with the unsloth python repository

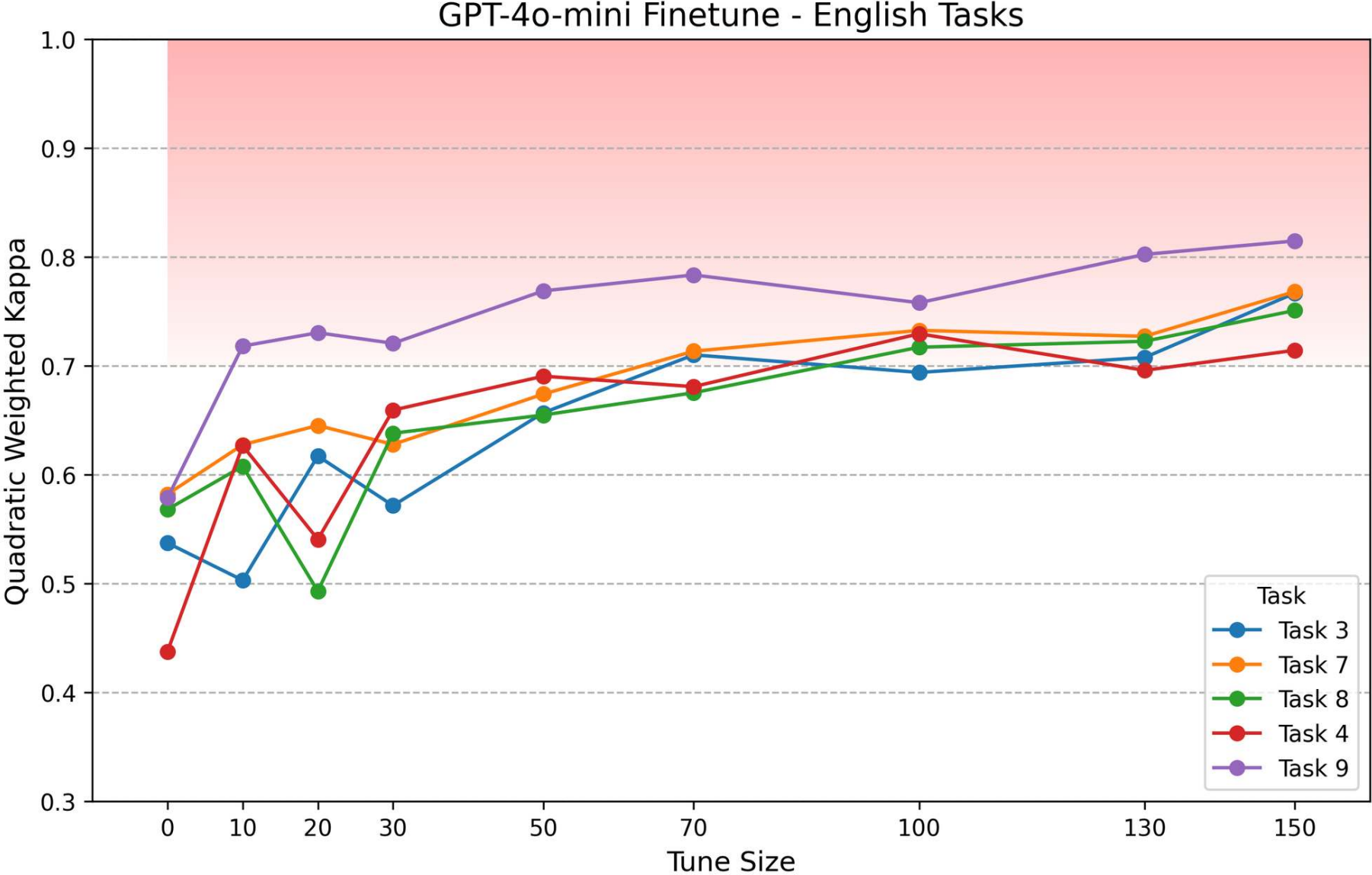
# GPT QWK BY TUNE SIZE - SCIENCE TASKS



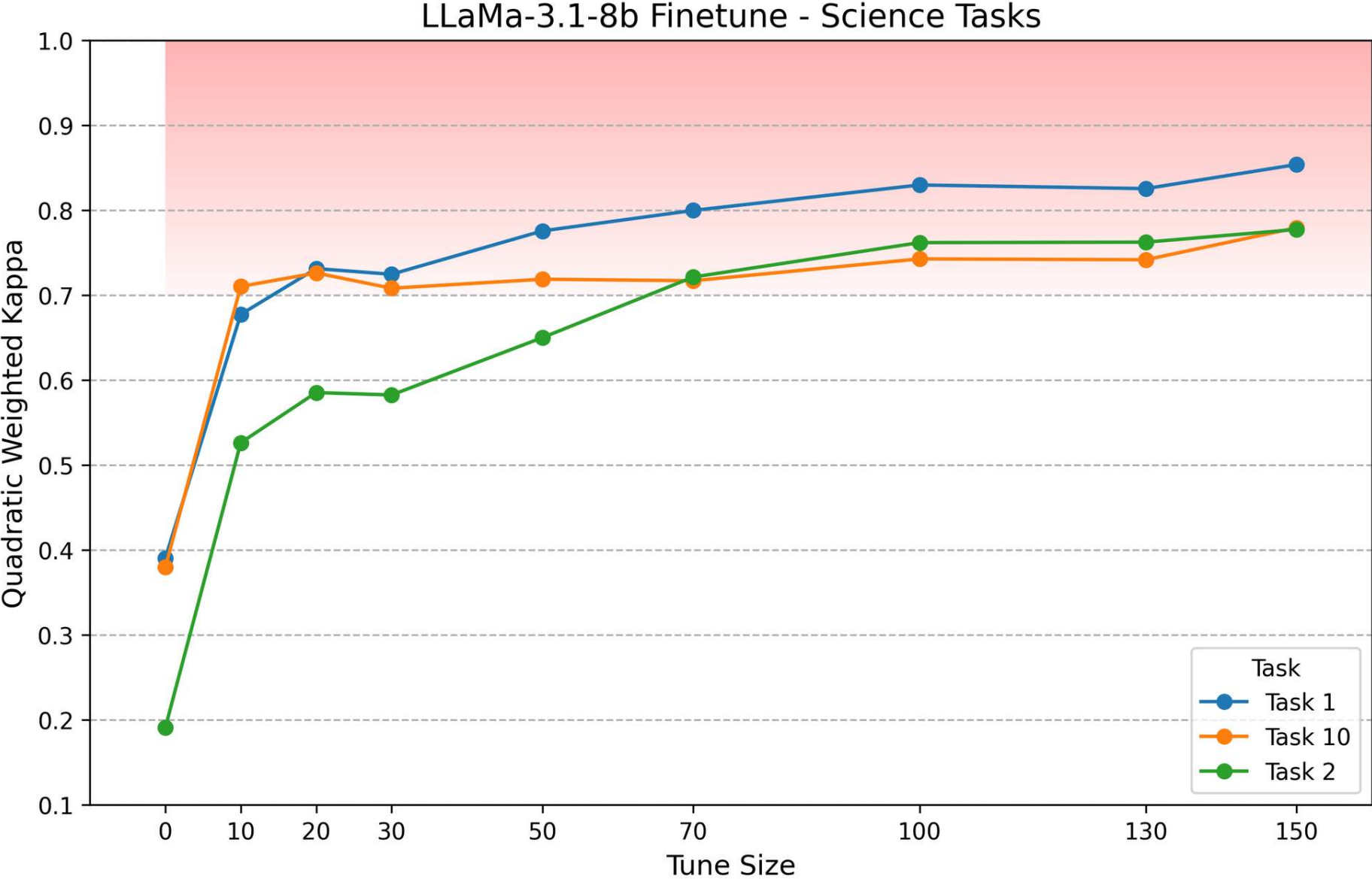
# GPT QWK BY TUNE SIZE - BIOLOGY TASKS



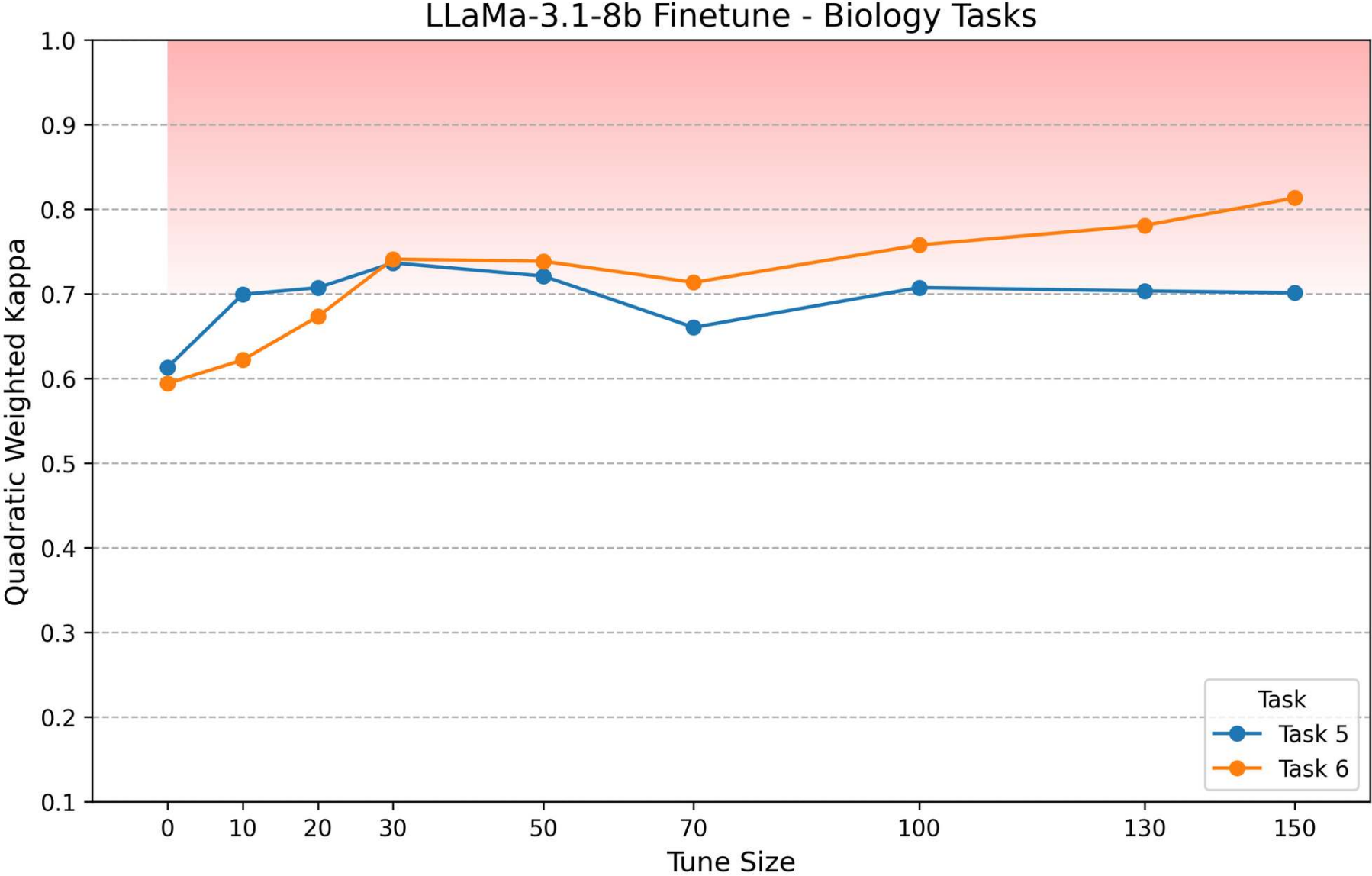
# GPT QWK BY TUNE SIZE - ENGLISH TASKS



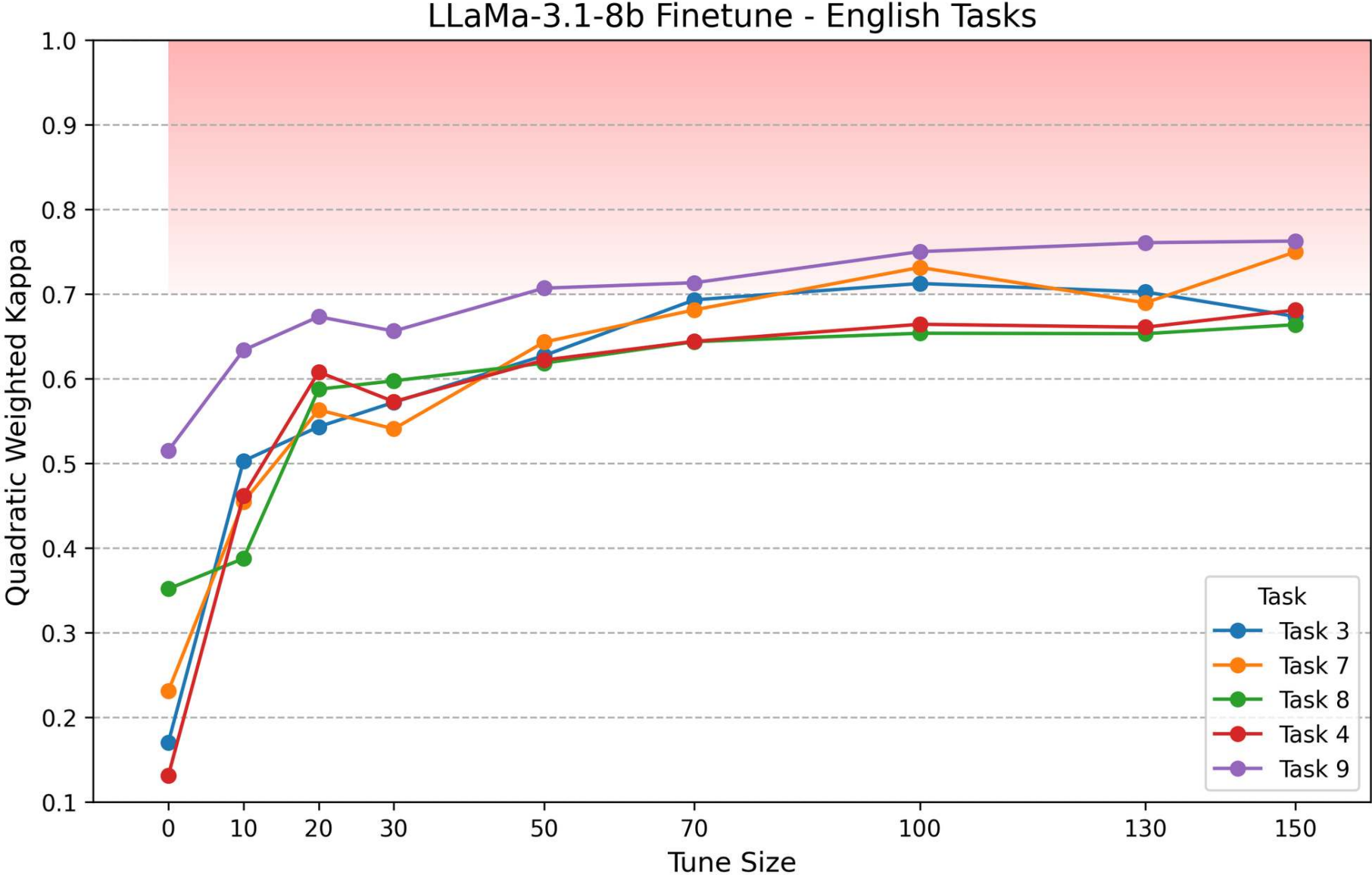
# LLAMA QWK BY TUNE SIZE - SCIENCE TASKS



# LLAMA QWK BY TUNE SIZE - BIOLOGY TASKS



# LLAMA QWK BY TUNE SIZE - ENGLISH TASKS



**THE ROAD FORWARD...**



# CONCLUSIONS

- LLMs have consistently been shown to automatically grade at a non-expert human marker level of accuracy and reliability
- We can potentially fine-tune the human out of the loop...
- Further exploration with more and varied educational assessment datasets by discipline, language and culture
- More extensive investigation of bias and other threats to validity
- A shift to focusing on production and quality of feedback rather than grading

## CONCLUSIONS

- A focus on open-weights models, collaborative data sharing, and a democratic, non-profit driven commitment to the development, use and validation of AI in educational assessment - "*AI by all, for all...*"