Centre for Teaching and Learning

UNIVERSITY OF OXFORD

Responsible use of AI in academic context: AI Task Integrity Cards

Contents

What it is

What is the card

This document provides a template for describing AI tasks in academic context in a responsible manner that takes into account the probabilistic nature of all generative AI tools.

Any responsible description or suggestion to try to “use AI to do something” should describe:

- possible paths to success
- variability across contexts
- potential sources of failure
- strengths and weaknesses of different tools and/or models
- examples of prompts and sample chats

What is an AI task and how do we describe it?

For the purposes of the task integrity card, an AI task is anything that we may ask AI to do either via a prompt or another interface. This may be intentionally directly or under the surface as part of a workflow.

Unlike an AI model, AI task does not have a specific level of granularity. We may choose to define it for a particular communicative or investigative purpose or with respect to a dimension of difference from other tasks or a combination of the two.

For example, in certain contexts and for communicating with certain users we may decide to create a task card for:

- Using AI for translation
- Using AI for poetry translation
- Using AI for poetry translation for educational purposes
- Using AI for poetry translation for educational purposes in German teaching

At every level, the template provides opportunities to add meaningful information in a systematic and comparable way to users of AI tools.

We can also define a task through the perspective of due diligence, for example, people estimating the academic integrity of examinations (again at multiple levels of granularity):

- Potential for uses of AI in answering multiple choice questions
- Potential for uses of AI in answering multiple choice questions in the sciences

- Potential for uses of AI in answering multiple choice questions on a particular examination

In this case, the purpose is not necessarily to guide end-users but to give a detailed and robust information to test designers or policy makers.

Why is this needed

Artificial intelligence is increasingly being used for a variety of academic tasks in many contexts. Many recommendations and suggestions for possible uses are available. However, unlike with other tools, the behaviour of AI tools is highly unpredictable across uses, tools, and contexts. In particular:

- AI tools produce different output every time in response to identical task
- It is impossible to tell ahead of time if AI will succeed at a particular task in a particular context
- There are many seemingly similar tasks where AI will succeed at one and fail at another
- AI tools will mix in accurate and hallucinated information without any indication of a difference
- AI tools are constantly improving and changing their capabilities
- AI outputs can be greatly improved by different ways of formulating a prompt
- AI tools do not reliably report
- Users of AI tools do not have reliable intuition about interpreting the output generated by AI

Very often, these characteristics of AI are introduced in the form of separate guidance but it is not always certain that the users will have accessed the guidance or are able to apply it to the particular task.

The authors of the guidance have not always made it clear whether this is something that has been suggested or it has been tried.

Similar approaches to responsible disclosure

This approach is modelled on the idea of the Model Card introduced by Google in 2020 as a template for responsible disclosure about the limits and capabilities of new artificial intelligence models.

This practice has been widely adopted in the field and most new models are now accompanied by a disclosure modelled on the Model Card.

However, no such unified practice exists in guidance for the use of AI in education. There are some examples of guidance that includes prompts and sample chats but not in a systematic way.

Another example of a more systematic approach is the Assessment Ideas for an AI enabled world from JISC but it does not contain tips on where AI will fail and how it will succeed.

How to use the template

The template is intended to guide the process of formulating as well as sharing guidance on how to use AI for a particular task.

1. Formulate a task from the perspective of AI or an existing task
2. Try the task several times in different contexts and with different prompts
3. Note the failures and think about how they are instances of general limitations of AI
4. Describe the potential failures with reference to a general description of AI functionality
5. Share sample prompts and where possible links to actual chats

The template can be modified for different purposes:

1. Evaluating individual AI tools or models
2. Any educational task where success depends on context and failure is possible

Template: Responsible AI Use Task Integrity Card

Overview

This overview evaluates assigns values across seven categories using 3 descriptors. The idea is to give a user a quick overview of what to pay attention to. It will also make it possible to compare different AI tasks.

Condition	Low	Medium	High
Suitability of AI for task	Not an ideal task for AI in most circumstance	Suitable in some contexts with provisos	Suitable in most contexts
Variability across contexts	Will succeed in most situations for most input	Success is dependent on the right context in predictable ways	Will fail in many situations and for different input without an easy to predict
Need for iteration	Very often, the first output of the	Additional prompting is often required to	Acceptable results are almost always the result

	prompt produces adequate results	produce adequate results	of long chain of interactions
Need for manual modification	Output can be often used as is with small tweaks and changes	Output often has to be modified significantly in at least one area	Output can rarely be used without significant changes across many aspects
Hallucination problem	Hallucination is usually not frequent or does not represent a problem	Hallucination can be a problem depending on context	Hallucination is very likely to significantly impact the output
Prompt sensitivity	High chance of success even with generic prompts	Richer prompts will contribute to better output	Highly sensitive to exact prompt wording or style
Context window dependence	Task is not likely to run against context window limits in most circumstance	There are some contexts in which the task will run up against the context window limitations	High attention is required to the length of the input to avoid running against context window limits
Variability across tools/models	Most generative AI tools will be suitable for this task regardless of models	Some tools may be better than others in certain contexts	Only using specific tools is likely to lead to success

Task aim and description

See above on defining task and task granularity.

- What is the task trying to accomplish
- What learning and productivity outcomes can AI contribute to
- How can the task be performed
- What context is it most appropriate for
- How is it different from or similar to other tasks
- Is it a more specific version of a more general category of tasks that has some specific requirements

How AI can succeed at the task

Describe what are the criteria for success are:

- What the user can expect from the AI

- What the user has to contribute
- How the user can take advantage of imperfect output
- Best ways of formulating prompts

Where AI can fail at the task

Describe specific potential for failure as it related to the task and if known/available how prompts can alleviate it:

- Where hallucination can be a problem
- Where context window can be a problem
- Where lack of logic can be a problem
- Where the lack of metalanguage can be an issue

Sample prompts and chats

List exact text of prompts known to produce successful results. Ideally with links to shared chats.

Notes on suitable tools

List tools that can be best used for this task, noting as appropriate:

1. Choice of generic chat vs a special purpose tool
2. Strengths and weaknesses of specific tools/models
3. Comparison of free vs paid tools
4. Data privacy concerns, if any
5. Date of last test

Example: Using AI for text summarisation

Overview

Task	Level	Notes
Suitability of AI	High	One of the most common uses.
Variability across contexts	Low	This task is likely to succeed with most texts in most languages.
Need for iteration	Low	Acceptable summaries are often produced after first prompt but additional queries often improve the results
Need for manual modification	Low	Summaries can often be used with minor edits
Hallucination problem	Low	Low in most contexts but increases with length of text
Prompt sensitivity	Low	Often produces good results with generic prompts
Context window dependence	High	Check context window of tool before use
Variability across tools	Mediu m	Only Claude.ai for longer contexts

Task aim and description

Use AI to summarise longer text in various languages, styles and formats. You can use this to:

- Help you decide whether to read a longer text
- Help you improve your own writing by generating comparisons
- Check your own understanding of a text
- Check your own texts for completeness
- Create summaries of longer writing you produced

How AI can succeed

Summarisation is one of the strengths of generative AI. Often this does not require any special prompting but you can improve your results if you use a better prompting technique:

- Specify a persona: Summarise this as an expert in the field
- Specify audience: Summarise this as if for a non-expert audience

- Specify a style: Summarise this as an abstract to an academic paper
- Specify a format: Summarise this as an outline
- Asking for alternatives: simply refresh

How AI can fail

Although AI is generally very good at summarisation but it can frequently encounter issues.

Sources of failure

1. AI can not access the whole text if the text is too long (as it generates its response, the text may no longer be available), this may result in hallucination.
2. Texts in languages other than English may exceed the context.
3. AI can “hallucinate” things that could be in the text but are actually not in the text. This could be very subtle and mixed in with things that are in the text:
    a. Terms that are also used for something but are not used in the text
    b. People or facts that could be in the text but are not
    c. Conclusions that are often drawn in this subject but are not actually made by the authors
4. AI may also not mimic the style or language specified accurately or appropriately
5. AI cannot count words, do not rely on it to produce exact word counts. You can specify length but will need to check the exact number of words.

Mitigating failure and responsible use

1. Ask AI to generate multiple summaries and compare them
2. Choose a tool with sufficient context window (Claude or ChatGPT Plus)
3. Strip out any text that is not essential for the summary (references, acknowledgments, abstract, etc.) to shorten what is presented to AI
4. Start a new chat for additional interactions with the text
5. Check any factual information (names, dates, numbers, etc.) against the text

Sample prompts and chats

Note: The best tool for this task is Claude.ai which does not allow sharing chats.

1. “You are an expert in economics who is also very good at popularising. Summarise this paper in 200 words.

2. “You are an expert editor of an economics journal. Write a summary of this article that I can send to the editor.”
3. “Create an outline of the key arguments in this text. Note sources of evidence, research methods and conclusion the text makes.”

You can also include an example of a similar summary and ask AI to follow the same style.

Choosing the right tool

Always check the “context window” of the tool you are using. This is specified in tokens – 100 tokens usually = about 75 words in English. Not all tools will disclose this.

Best tools

1. Claude.ai: This is the best tool for summarising longer texts (up to about 70,000 words in English). You can upload PDFs. Works on the mobile phone.
2. ChatGPT Plus (paid): Can summarise texts up to about 6,000 words. Best quality models. Also offers plugins for summarising of longer texts.
3. ChatGPT (free): Can summarise texts up to 2-3,000 words. Note: Longer summaries may need shorter text.

Other possibilities

There are many other tools dedicated to summarizing texts. They will often not provide as good an outcome as Claude or ChatGPT plus.

- Elicit
- Scholarcy
- Sciscape

Sample AI Task Analysis Comparison

This is a sample table comparing various AI tasks as to their suitability and common issues.

Task	Suitability of AI for task	Variability across situations	Need for iteration	Need for manual modification	Hallucination problem	Prompt sensitivity	Context window dependence	Variability across tools/models
Generating a summary	High	Low	Low	Low	Low	Low	High	Medium
Creating tables from unstructured text	High	Low	Low	Low	Low	Medium	High	Low
Identifying people, terms of words in text	High	Low	Low	Low	Low	Medium	High	Low
Providing information about famous books	Medium	High	Medium	Low	High	Medium	Low	Low
Asking for biographies people	Low	High	Medium	Low	Low	Medium	Low	Low
Correcting or translating existing code	High	Medium	Medium	Medium	Medium	High	Medium	Medium
Writing computer code based on specification	High	Medium	High	Medium	Medium	High	Medium	Medium
Generating a general illustration image in different styles	High	Medium	Medium	Low	High	High	N/A	Medium
Generating image containing complicated text	Low	High	High	Medium	Medium	High	N/A	High
Interpreting an image	Medium	High	Medium	High	High	Low	N/A	High
Generating questions about a text	High	Low	Low	Medium	Medium	Low	High	Low