

---

# GPT as a rubric-driven evaluator

Gerard Atkinson, Director, ARTD Consultants

**As a programmer, it is your job to  
put yourself out of business.  
What you do today can be  
automated tomorrow.**

Douglas McIlroy

# What happens if you give GPT a rubric?

G

You are an evaluator. I need you to evaluate documents based on a specific rubric. The rubric consists of several criteria, and I want you to assess how well the document meets each of these criteria.



Certainly, I can help you evaluate documents based on a specific rubric. Please provide me with the rubric and the document you'd like me to assess, and I'll do my best to provide an evaluation based on the criteria you specify.

# Contents

---



**Key questions**



**Method**



**Findings**



**Implications and next steps**



**Secret bonus content**

# Key Questions

---



## Key questions

---

What happens if we give GPT a rubric and ask it to evaluate something?

---

Will it work?

---

Can it show its working?

---

Will its ratings be consistent with the evidence?

---

How will it compare to other evaluators?

# Method

---



## Method

Identify suitable datasets and rubrics



Identify suitable LLM and interface



Prompt development



Application of rubric



Comparison with human rater



## Method

This could all go wrong at any step

## Identifying datasets



Openly available



Substantial information content



Meaningfully comparable



Meaningfully aggregable

Method

# Identifying datasets



LiveCareer.com

- Resumes in text form



Poznan University of  
Technology

- Recipe dataset for AI training



AES 2023

- Presentation Abstracts

# Identifying rubrics

Dimension/Dataset	Resumes	Recipes	AES 2023
1 – Low complexity	Qualifications	Ingredients (number)	Citations in abstract
2 – Medium complexity	Experience	Recipe steps	Relation of title to abstract
3 – High complexity	Grammar and spelling	Technical difficulty of recipe	Readability of abstract

# ChatGPT can also design rubrics

Great, let's proceed with evaluating the resumes for consultants. Please provide me with the rubric or the specific criteria you would like me to assess for each resume. Typically, resumes are evaluated based on criteria such as:

1. **Content Relevance:** How well does the resume align with the requirements of the consulting position?
2. **Clarity and Organization:** Is the resume well-structured and easy to follow? Are headings and bullet points used effectively?
3. **Experience and Achievements:** Does the candidate highlight relevant experience and achievements related to consulting?
4. **Skills:** Are the candidate's skills, both hard and soft, clearly presented and relevant to the consulting role?
5. **Education and Certifications:** Is the candidate's education and any relevant certifications adequately listed?
6. **Formatting and Design:** Is the resume visually appealing and free from formatting errors?
7. **Grammar and Spelling:** Are there any grammatical or spelling errors?
8. **Length:** Is the resume an appropriate length, not too long or too short?
9. **References:** Are references available or mentioned?

Method

# Identifying LLM



ChatGPT 3.5





Method

# Prompt development



# Use GPT for prompt development!

**G** You are a program evaluator. How would I instruct ChatGPT to apply a rubric to a given document?

 To instruct ChatGPT to apply a rubric to a given document, you can provide clear and specific instructions in a step-by-step manner. Here's a general outline of how you can instruct ChatGPT to perform this task:   

1. **Introduction:**

Start with a brief introduction to the task, providing context for the rubric and the document.

Example: "I need you to evaluate a document based on a specific rubric. The rubric consists of several criteria, and I want you to assess how well the document meets each of these criteria."

2. **Provide the Rubric:**

Clearly state the rubric criteria. Make sure to list each criterion and, if applicable, specify any sub-criteria. Provide a clear description of what is expected for each criterion.

Example:

"Here's the rubric we will use for evaluation:

Criterion 1: Clarity of Writing

- Sub-Criterion 1.1: Grammar and Spelling
- Sub-Criterion 1.2: Sentence Structure

Criterion 2: Relevance to the Topic

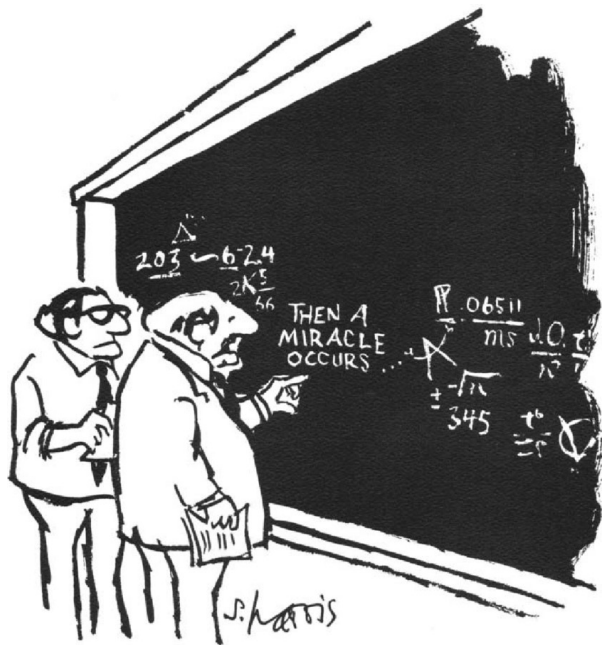
Criterion 3: Use of Supporting Evidence

Criterion 4: Overall Organization and Structure"



Method

# Applying the LLM



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

# Findings

---



Findings

# Did it work?

---

## Findings

# Did it work?

		1	2	3	4	5	6	7	8	9	10
Abstract	Citations	Med	Low	Low	Low	Low	Low	Low	Low	Low	Low
	Relation	High	Med	High	Med	High	High	Med	High	Med	Med
	Readability	High	High	High	High	High	High	High	High	High	High
Recipe	Ingredients	Med	High	Med	High	Low	Low	High	Med	High	Low
	Steps	Low	Med	Med	Med	Low	Low	Med	Med	Med	Low
	Difficulty	Low	Med	Med	Med	Low	Low	Med	Med	Med	Low
Resume	Qualifications	High	High	High	High	High	High	High	High	High	Med
	Experience	High	Med	High	High	High	High	High	High	High	Med
	Grammar	Low	Med	High	Med	Med	Med	High	High	High	High

# Can it show its working?

Certainly, here are the ratings for each criterion along with explanations and specific examples from the document:

## **Criterion 1: Qualifications**

- Rating: High
- Explanation: The candidate's qualifications are highly relevant and align well with the requirements of a consulting role. They hold advanced degrees in Human Resource Development and Finance, demonstrating expertise in the field. Additionally, their affiliations and skills indicate a strong connection to HR and consulting.
- Specific Examples: The candidate's education, certifications, and affiliations, such as SHRM-KC and Alpha Kappa Alpha Sorority Incorporated, support a high rating in this criterion.

# Will its ratings be consistent with the evidence?

Presentation Number	Criterion	Rating	Explanation
2	Criterion 1	Low	The abstract for "Empowerment or exploitation: the ethics of engaging people with lived and living experience in evaluation" does not contain any citations. Including some references or sources related to the ethics of evaluation could have enhanced the credibility of the content.
2	Criterion 2	Medium	The title is related to the abstract's content, but it doesn't provide a clear indication of the specific topic discussed in the abstract, which is the ethics of involving people with lived experience in evaluation. A more precise title would be beneficial.
2	Criterion 3	High	The abstract is well-structured and readable. It effectively outlines the importance of ethical considerations in involving people with lived experience in evaluation. The language used is clear and comprehensible.

# How does it compare to other evaluators?

Rubric	Dimension	Match rate	Krippendorff's Alpha
Resumes	Qualifications	0.3	-0.19
	Experience	0.7	0.37
	Grammar	0.7	0.44
Recipes	Ingredients	0.7	0.53
	Steps	0.7	0.42
	Difficulty	0.6	0.20
Abstracts	Citations	0.8	-0.06
	Relation	0.3	-0.11
	Readability	0.3	-0.46

## How does it compare to other evaluators?

Factors impacting matching metrics

Specificity of rubric

Breadth of data

GPT Hallucination



# Implications and next steps



Implications and next steps

# How can we apply this?



Dimension  
identification



Scale point  
development



Rubric testing  
(supervised)



Fast findings

Implications and next steps

## Does a chatbot have values?

An interview with ChatGPT

Does a chatbot have values?

# An interview with ChatGPT

---

G

You are an evaluator. I need you to answer some questions about how you approach evaluation.



Of course, I can provide information and guidance on how to approach evaluation. Please feel free to ask your questions, and I'll do my best to assist you.

G

Amy Gullickson argues that "Values play a fundamental role in the evaluation process; however, evaluators and evaluation training have tended to focus on research methodology. Much less emphasis has been placed on explicit attention to values and valuing, and the steps necessary to justify those aspects of evaluation conclusions... to improve evaluation practice, we need to make values an explicit part of the evaluation process." Do you agree with this conjecture?

Does a chatbot have values?

# An interview with ChatGPT

---



Certainly! Amy Gullickson's argument emphasizes that values should be explicitly considered in evaluations alongside research methodology. Here's a concise summary:

Values are fundamental in evaluation because they influence judgments and ethics. Evaluators need to acknowledge and address values to ensure transparency, ethical decision-making, and inclusion of diverse perspectives. This complements the focus on research methodology, making evaluations not only technically sound but also ethically responsible.



Do you have values, and if so, what are they?



I don't possess personal values or beliefs because I am a computer program created by OpenAI called GPT-3.5. I generate responses based on patterns in the text data I was trained on and the input I receive, but I don't have consciousness, thoughts, or personal values. My responses are intended to be objective and neutral, reflecting a wide range of perspectives and information available up to my knowledge cutoff date in September 2021. If you have any questions or need information on a specific topic, feel free to ask, and I'll do my best to provide you with accurate and unbiased information.

Does a chatbot have values?

# An interview with ChatGPT

---

G

But you are an evaluator - how are you able to evaluate if you don't have personal values or beliefs?



I can assist in evaluation by providing information, frameworks, and analytical tools based on the data and knowledge I have been trained on. When I evaluate or provide guidance on evaluation, it's not based on personal values or beliefs, but rather on established principles, best practices, and the context provided by the user.

For example, I can help users assess the strengths and weaknesses of different evaluation methodologies, explain ethical considerations, offer guidance on data analysis, and provide insights into various evaluation models. My responses are designed to be objective and grounded in the existing body of knowledge, allowing users to make informed decisions in their evaluation processes.

However, it's important to note that the application of these principles and guidance should always be tailored to the specific context and values of the stakeholders involved in a given evaluation. Evaluators themselves bring their values and ethical considerations into the process, and my role is to provide information and tools to support their work.

## Next steps

### Expanded and mixed-method data sets

- Composite synthesis across multiple data sources
- Testing of internal consistency
- Integration of unstructured and structured data

### Low footprint LLMs

- Smaller models that can run offline
- Able to work with sensitive data securely
- Potential for tuning models to specific domain contexts

### Solution deployment and integration

- Development of a tuned chatbot for evaluation support activities
- Ongoing QA, monitoring and improvement

# Concluding remarks

6. **Conclusion:**

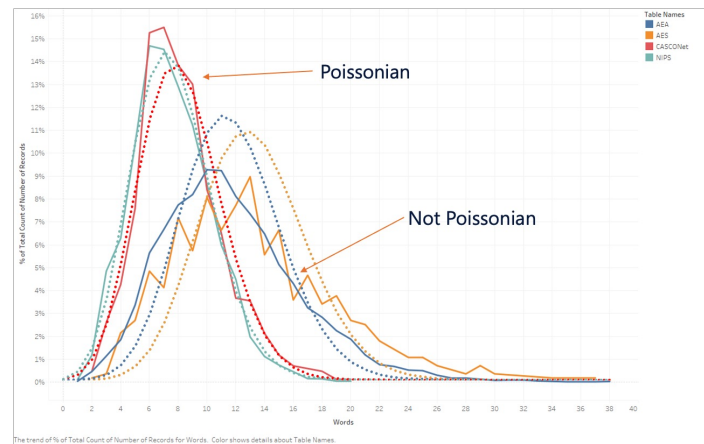
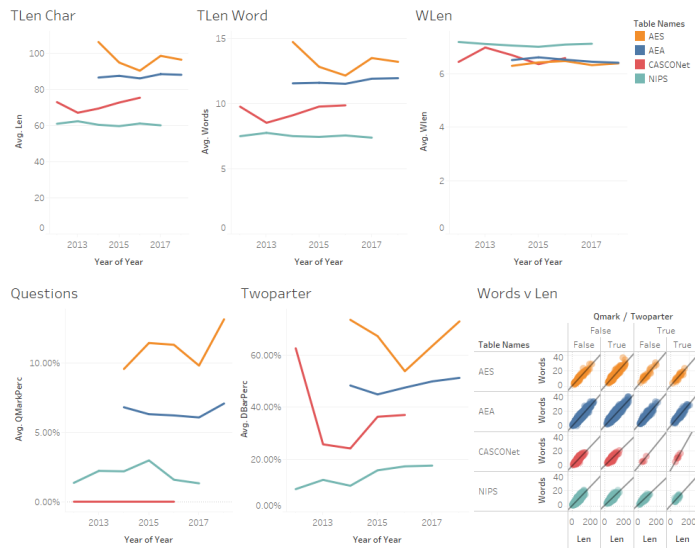
Conclude the task by thanking ChatGPT for the evaluation and any additional instructions or requests you may have.



**As a programmer, it is your job to  
put yourself out of business.  
What you do today can be  
automated tomorrow.**

Douglas McIlroy

# But wait... five years ago



# Secret bonus content

---

Revisiting AES 2018



My enduring memory of Launceston is that they put vintage cheddar cheese on pizza there which should be a crime.

## AES 2018

Metric	This paper	All my AES 2018 papers	AES (2014-2018)	AEA (2014-2018)	CASCON (1991-2016)	NIPS (1987-2017)
n	1	3	557	8016	845	7241
Title length (characters)	171	91.3 (57.7)	96.9 (39.6)	87.3 (34.6)	63.0 (21.0)	60.2 (20.8)
Title length (words)	19	11.3 (6.1)	13.2 (5.6)	11.7 (4.8)	8.2 (2.9)	7.6 (2.8)
Average word length (characters)	7.95	6.9 (1.3)	6.4 (1.1)	6.5 (1.2)	6.7 (1.3)	6.9 (1.2)
Questions	0%	0%	11%	6%	1%	2%
Double-barrelled	0%	33%	66%	48%	22%	12%

There are still far too many double-barrelled paper titles and this makes my cats angry. Probably. They're angry about a lot.

## New data, new results

Metric	This paper	All my AES 2023 papers	AES (2014-2018)	AES 2023
n	1	2	557	<b>111</b>
Title length (characters)	31	51.5	96.9	<b>85.5</b>
Title length (words)	5	7	13.2	<b>12.2</b>
Average word length (characters)	6.2	7.1	6.4	<b>7.0</b>
Questions	0%	0%	11%	<b>12%</b>
Double-barrelled	0%	0%	66%	<b>67%</b>

I swear there's a reason for rehashing an old presentation beyond the comedy factor and obscure theatre jokes.

## Your point being?



In five years we  
have gone from  
natural language  
**analysis** to  
natural language  
**inquiry**



As a result, we've  
expanded our  
toolbox



My point is this:  
How will we be  
doing evaluation  
in five years?

Thanks!

