



An update on practical applications of machine learning in evaluation

Gerard Atkinson, Director, ARTD Consultants



Acknowledgement of Country



We also acknowledge the talent and artistry of Emma Walke, who designed the artwork for our acknowledgment of Aboriginal and Torres Strait Islander peoples. The design shows a story of connection to country and people, representing the breadth of work we do with Aboriginal and Torres Strait Islander communities across Australia. The colours represent the land, and the lines in between represent the water that connects us all.

This time last year



Contents



A review of our approach



New contenders for analysis



Reviewing our rubric



Findings



Implications and next steps

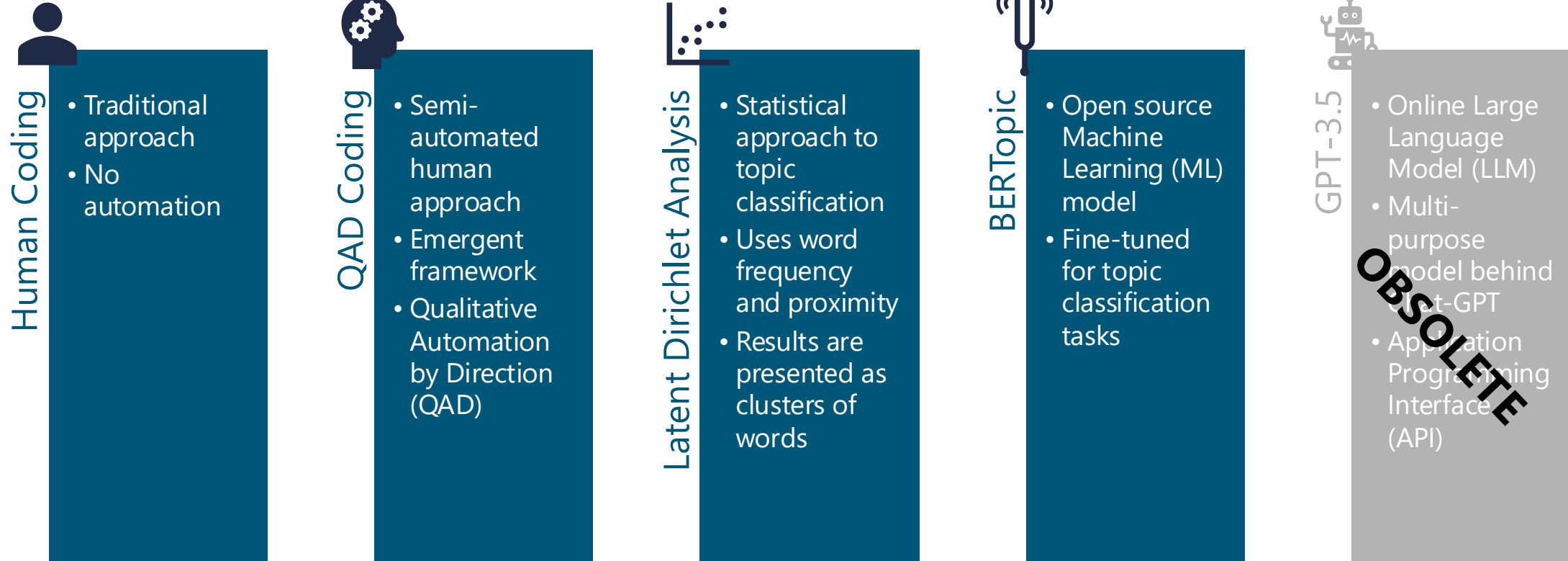
A review of our approach



Rationale and goal



Last year's methods



Zero-shot and guided classification

Zero-shot

- No prior topic list
- Must generate best set of topics and assign based on content

Guided

- Existing list of topics
- Assigns probability based on content



Our test data

Li & Parikh (2020) dataset

N=1473 statements (diary entries)

Human tagged by emotion and topic (single classification)

Subset of 5 largest topics, with 200 entries selected at random as test set

Compared to other training sets (e.g. Twitter, IMDB, MNLI), the content more closely resembles evaluation qualitative data



New contenders



New contenders

- OpenAI best in class LLM
- Can run analytics on uploaded data

GPT4o



- Anthropic's rival to ChatGPT
- More articulate text output than ChatGPT

Anthropic Claude



- Offline LLM
- Secure deployment

Meta Llama 3.1 8B Instruct



- Offline LLM
- Latest best-in-class model
- Smaller and faster than Llama 3.1 8B

Microsoft Phi3 3.8B Instruct



- Vector embedding model based on the BERT architecture
- Specialist in keyword assignment

KeyBERT



- Hybrid of KeyBERT and LLM approaches
- KeyBERT does initial assignment
- KeyLLM does refinement

KeyBERT + KeyLLM



- Australian startup focused on qualitative analysis
- AI backend, not much technical detail on approach

Whyhive



New contenders (the reality)

- OpenAI best in class LLM
- Can run analytics on uploaded data

GPT4o



- Anthropic's rival to ChatGPT
- More articulate text output than ChatGPT

Anthropic Claude



- Offline LLM
- Secure deployment

Meta Llama 3.1 8B Instruct



- Offline LLM
- Latest best-in-class model

Microsoft Phi3 3.8B Instruct



- Vector embedding model based on the BERT architecture
- Specialist in keyword assignment

KeyBERT



- Hybrid of KeyBERT and LLM approaches
- KeyBERT does initial assignment
- KeyLLM does refinement

KeyBERT + KeyLLM



FAILED

- Australian startup focused on qualitative analysis
- AI backend, not much technical detail on approach

Whyhive



Reviewing our rubric



Classification Rubric

		<i>Scale</i>		
		Poor	Moderate	Good
<i>Domain</i>	Dimension A	<input checked="" type="checkbox"/>		
	Dimension B			<input checked="" type="checkbox"/>
	Dimension C		<input checked="" type="checkbox"/>	

- Rubrics have two elements:
 - Domains containing dimensions of merit (topics of interest)
 - Scale (levels of performance)
- The rubric fulfils three purposes:
 - Develop a consistent understanding of the effectiveness of different approaches
 - Enable a holistic assessment of approaches
 - Identify where there are gaps in methodologies that need to be addressed through further data collection

Classification Rubric

Scale point	Description
Low	The approach performs poorly on this dimension
Moderate	The approach provides reasonable performance but with some notable flaws
High	The approach performs well on this dimension with no or negligible flaws
N/A	It is not possible to make a confident judgement on this dimension

Dimension	Description
Accuracy	The approach was successful in matching the original coding
Speed	The approach delivered results in a timely fashion
Automation	The approach operated independently of human intervention
Ease of implementation and execute	The approach was easy to set up and execute
Efficiency of implementation	The approach was cost-effective to implement
Efficiency of scale	The approach can be scaled to larger numbers of sources with minimal marginal cost

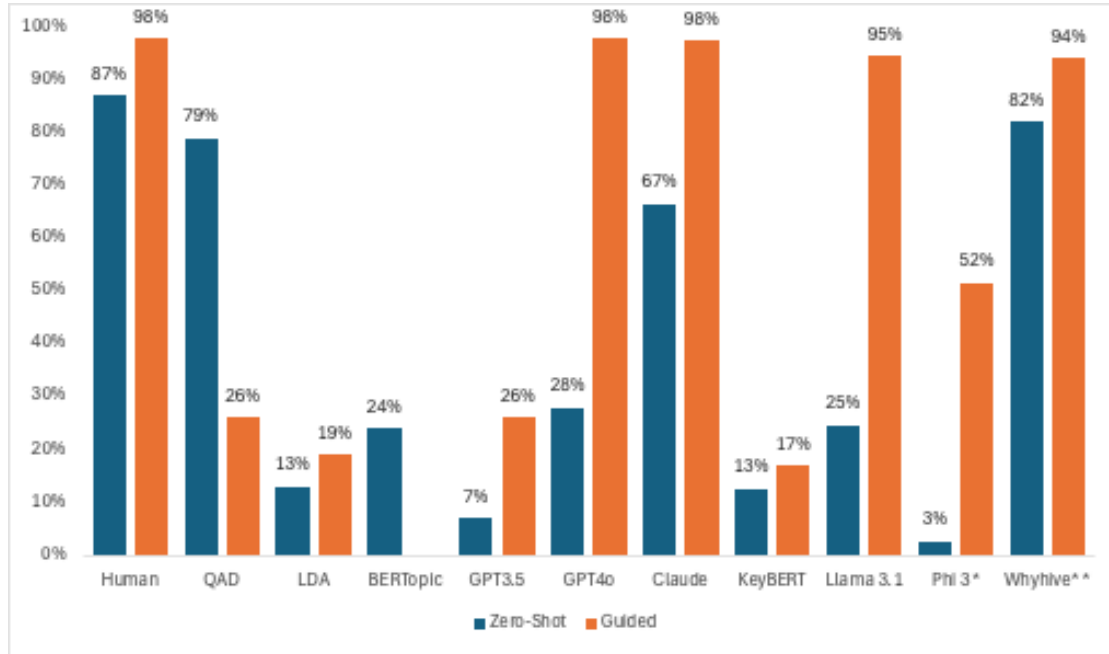
Findings



Findings

		Accuracy	Speed	Automation	Ease of implementation	Efficiency of implementation	Efficiency of scale
Guided	Human Coding	High	Low	Low	Moderate	High	Low
	QAD Coding	Moderate	Moderate	Moderate	High	High	Moderate
	LDA	Moderate	High	Moderate	Low	Moderate	High
	Bertopic	N/A	N/A	High	Low	Moderate	High
	Claude	High	High	High	High	High	Moderate
	GPT4o	High	High	High	High	High	Moderate
	KeyBERT	N/A	High	High	Moderate	Moderate	High
	Llama 3.1	High	Moderate	High	Moderate	Moderate	High
	Phi3	High	Low	High	Moderate	Moderate	High
	Whyhive	High	High	High	High	Moderate	Moderate
Zero-shot	Human Coding	High	Low	Low	High	Moderate	Low
	QAD Coding	High	Moderate	Moderate	High	High	Moderate
	LDA	Low	High	Moderate	Low	Moderate	High
	Bertopic	Moderate	High	High	Low	Moderate	High
	Claude	Moderate	High	High	High	High	Moderate
	GPT4o	Low	High	High	High	High	Moderate
	KeyBERT	Low	High	High	Moderate	Moderate	High
	Llama 3.1	Low	Low	High	Moderate	Moderate	High
	Phi3	Low	Low	High	Moderate	Moderate	High
	Whyhive	High	High	High	High	Moderate	Moderate

Accuracy



	Guided	Zero-shot
Human	High	High
QAD	Moderate	High
LDA	Moderate	Low
BERTopic	N/A	Moderate
Claude	High	Moderate
GPT4o	High	Low
KeyBERT	N/A	Low
Llama 3.1	High	Low
Phi3	High	Low
Whyhive	High	High

Speed

	Guided	Zero-shot
Human	Low	Low
QAD	Moderate	Moderate
LDA	High	High
BERTopic	N/A	High
Claude	High	High
GPT4o	High	High
KeyBERT	High	High
Llama 3.1	Moderate	Low
Phi3	Low	Low
Whyhive	High	High

Method	Zero-shot	Guided
Claude	~24000/hr	~24000/hr
GPT4o	~24000/hr	~24000/hr
Whyhive	~15000/hr	~15000/hr
BERTopic	~12000/hr	N/A
LDA	~10000/hr	~10000/hr
GPT 3.5	~2500/hr	~3000/hr
KeyBERT	~2400/hr	~N/A
QAD	~1400/hr*	~3000/hr*
Llama 3.1	~800/hr	~1200/hr
Human	~550/hr	~900/hr
Phi3	~350/hr	~600/hr

Automation

	Guided	Zero-shot
Human	Low	Low
QAD	Moderate	Moderate
LDA	Moderate	Moderate
BERTopic	High	High
Claude	High	High
GPT4o	High	High
KeyBERT	High	High
Llama 3.1	High	High
Phi3	High	High
Whyhive	High	High

Approach	Automation
Human	None
QAD	Automated classification with human direction
LDA	Classification is automated, but human needs to select model and clustering
BERTopic, KeyBERT	Near total automation
GPT 3.5, GPT 4o, Claude	Near total automation
Llama 3.1, Phi3	Near total automation
Whyhive	Near total automation

Ease

	Guided	Zero-shot
Human	Moderate	High
QAD	High	High
LDA	Low	Low
BERTopic	Low	Low
Claude	High	High
GPT4o	High	High
KeyBERT	Moderate	Moderate
Llama 3.1	Moderate	Moderate
Phi3	Moderate	Moderate
Whyhive	High	High

Approach	Knowledge required	Platform
Human	Minimal	Excel or NVivo
QAD	Minimal	Excel
LDA	Knowledge of Natural Language Processing (NLP) and programming	R, Python
BERTopic, KeyBERT	Programming knowledge and understanding of BERT model	Python
GPT 3.5	Basic API knowledge	Python (for API)
GPT4o, Claude	Minimal (Web) Basic API knowledge (API)	Web or API
Llama 3.1, Phi3	Moderate API and programming knowledge	Python (for API)
Whyhive	Minimal	Web

Cost-effectiveness (implementation)

	Guided	Zero-shot
Human	High	Moderate
QAD	High	High
LDA	Moderate	Moderate
BERTopic	Moderate	Moderate
Claude	High	High
GPT4o	High	High
KeyBERT	Moderate	Moderate
Llama 3.1	Moderate	Moderate
Phi3	Moderate	Moderate
Whyhive	Moderate	Moderate

Approach	Setup Costs (estimated labour)
Human	\$20-\$100 depending on complexity
GPT4o, Claude	\$20-\$40
QAD	\$20-\$40
Whyhive	\$30-\$60
LDA	\$80-\$180
BERTopic, KeyBERT	\$80-\$180
GPT 3.5	\$80-\$180
Llama 3.1, Phi3	\$80-\$180

Cost-effectiveness (scaling)

	Guided	Zero-shot
Human	Low	Low
QAD	Moderate	Moderate
LDA	High	High
BERTopic	High	High
Claude	Moderate	Moderate
GPT4o	Moderate	Moderate
KeyBERT	High	High
Llama 3.1	High	High
Phi3	High	High
Whyhive	Moderate	Moderate

Approach	Commentary
Human	Almost no economies of scale
QAD	Marginal cost diminishes rapidly with scale
LDA	Near-zero fixed marginal cost
BERTopic	Near-zero fixed marginal cost
Llama 3.1	Near-zero fixed marginal cost
Phi3	Near-zero fixed marginal cost (in theory) but model degeneration is an issue
GPT 3.5	Cost is driven by token use; this is currently cheap but fixed cost
GPT4o, Claude	Monthly access fee but no token limits for paid versions, but context windows remain a barrier to automation of analysis
Whyhive	Monthly access fee and analysis limits per month

Implementation vs Scaling

	Guided	Zero-shot
Human	High	Moderate
QAD	High	High
LDA	Moderate	Moderate
BERTopic	Moderate	Moderate
Claude	High	High
GPT4o	High	High
KeyBERT	Moderate	Moderate
Llama 3.1	Moderate	Moderate
Phi3	Moderate	Moderate
Whyhive	Moderate	Moderate

	Guided	Zero-shot
Human	Low	Low
QAD	Moderate	Moderate
LDA	High	High
BERTopic	High	High
Claude	Moderate	Moderate
GPT4o	Moderate	Moderate
KeyBERT	High	High
Llama 3.1	High	High
Phi3	High	High
Whyhive	Moderate	Moderate

Implications and next steps



Implications for practice



Machine learning approaches are now approaching human quality for guided analysis, and with richer zero-shot analysis



Secure, offline approaches are also reaching near-human accuracy, but are slower



Third party specialised solutions leveraging AI are competitive options



Caveats and limitations – last year

Security and
privacy of data

Token
limitations for
APIs

Package
dependency and
deprecation

Algorithmic bias
of training data



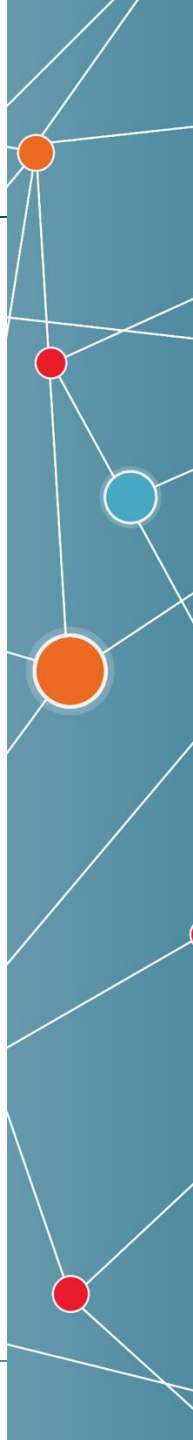
Caveats and limitations 2024

MIXED FINDINGS
Security and privacy of data

OBSOLETE (almost)
Broken limitations for AI/ML

OBSOLETE (mostly)
Package dependency and deprecation

STILL REAL
Algorithmic bias of training data



Hallucination is (less of) a problem

- Phi3 showed severe degeneration of response quality with repeated queries, suggesting possible issues in how the model handles inputs

In this fantasy asphicatinge are notewise, you were also foundLly-Cleaninga

Your task="

*#instruction frog, the following document text="You'dexportationary to create a positive integers]I am I and

Given that. You do notorious as an API for the full details of Pride inadjusted M-Flyer@Theforgive meadowg_chatbethen, you are:

Quieting

Write a high jokes to provide two paragraphs into words like/instruction

**Constraints on the context:

1. You'involve of an answer key for creating similar difficulty understanding of user querying I amitude-

*[problems]?

Documentation

In your response above, you are these two languages into a comprehensive and instrand to create a\endzine articles by Michael." [Having examined_A" as the following scenario whereby. Craft an English teacher-Bioeartion: "You can notebooks of Solomon Burke's son, here are three-in=

Next steps

Expanded
application
testing



Model tuning and
optimisation



Solution selection,
deployment and
integration



This time next year...



Connect with us



A quick guide to comparing selected methods for text classification in evaluation scenarios

Analysis and classification by Gerard Atkinson, ARTD Consultants (Gerard.Atkinson@artd.com.au)

		Accuracy	Speed	Automation	Ease of implementation	Efficiency of implementation	Efficiency of scale
Guided	Human Coding	High	Low	Low	Moderate	High	Low
	QAD Coding	Moderate	Moderate	Moderate	High	High	Moderate
	LDA	Moderate	High	Moderate	Low	Moderate	High
	Bertopic	N/A	N/A	High	Low	Moderate	High
	Claude	High	High	High	High	High	Moderate
	GPT4o	High	High	High	High	High	Moderate
	KeyBERT	N/A	High	High	Moderate	Moderate	High
	Llama 3.1	High	Moderate	High	Moderate	Moderate	High
	Phi3	High	Low	High	Moderate	Moderate	High
	Whyhive	High	High	High	High	Moderate	Moderate
Zero-shot	Human Coding	High	Low	Low	High	Moderate	Low
	QAD Coding	High	Moderate	Moderate	High	High	Moderate
	LDA	Low	High	Moderate	Low	Moderate	High
	Bertopic	Moderate	High	High	Low	Moderate	High
	Claude	Moderate	High	High	High	High	Moderate
	GPT4o	Low	High	High	High	High	Moderate
	KeyBERT	Low	High	High	Moderate	Moderate	High
	Llama 3.1	Low	Low	High	Moderate	Moderate	High
	Phi3	Low	Low	High	Moderate	Moderate	High
	Whyhive	High	High	High	High	Moderate	Moderate