SACAIR 2024
The South African Conference for Artificial Intelligence Research

ALEX AGYEMANG AND TIM SCHLIPPE

AI IN EDUCATION: AN ANALYSIS OF LARGE LANGUAGE MODELS FOR TWI AUTOMATIC SHORT ANSWER GRADING

Bloemfontein, South Africa
December 4, 2024

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

AGENDA

Introduction	1
Related Work	2
Experimental Setup	3
Experiments and Results	4
Conclusion and Future Work	5

1

INTRODUCTION

MOTIVATION: UN Sustainable Development Goal 4

SUSTAINABLE DEVELOPMENT GOALS

Image Sources: United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021); OpenClipart-Vectors/154119/Pixabay.

MOTIVATION: Teacher Shortage Worldwide

UNESCO DATA REVEALS A GLOBAL DECLINE OF INTEREST IN PURSUING A TEACHING CAREER.

Image and Text Sources: United Nations (02/2024)

MOTIVATION: Teacher Shortage in Sub-Saharan Africa

UNESCO DATA REVEALS A GLOBAL DECLINE OF INTEREST IN PURSUING A TEACHING CAREER.

Image and Text Sources: United Nations (02/2024); LinkedIn Post by Mastercard Foundation (02/2024)

AI IN EDUCATION: Potential

AUTOMATIC

SHORT

ANSWER

GRADING

Image Source: OpenClipart-Vectors/154119/Pixabay.

AUTOMATIC SHORT ANSWER GRADING

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

Graphic Source: Custom Depiction.

IDEA: LLMS FOR TWI AUTOMATIC SHORT ANSWER GRADING

GPT-4o

Claude 3

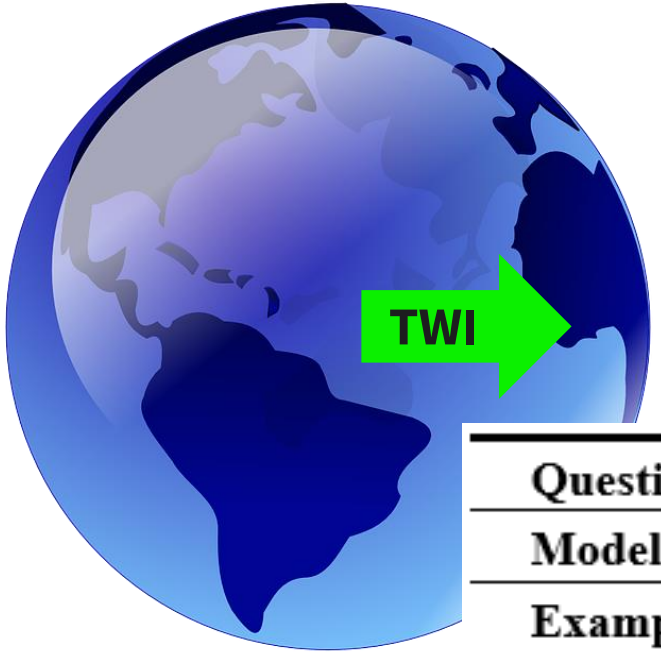
Llama 3

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

ZERO/FEW-SHOT CAPABILITIES: EASY TO USE WITH PROMPTING

Graphic Source: Custom Depiction.

IDEA: LLMS FOR TWI AUTOMATIC SHORT ANSWER GRADING



GPT-4o

Claude 3

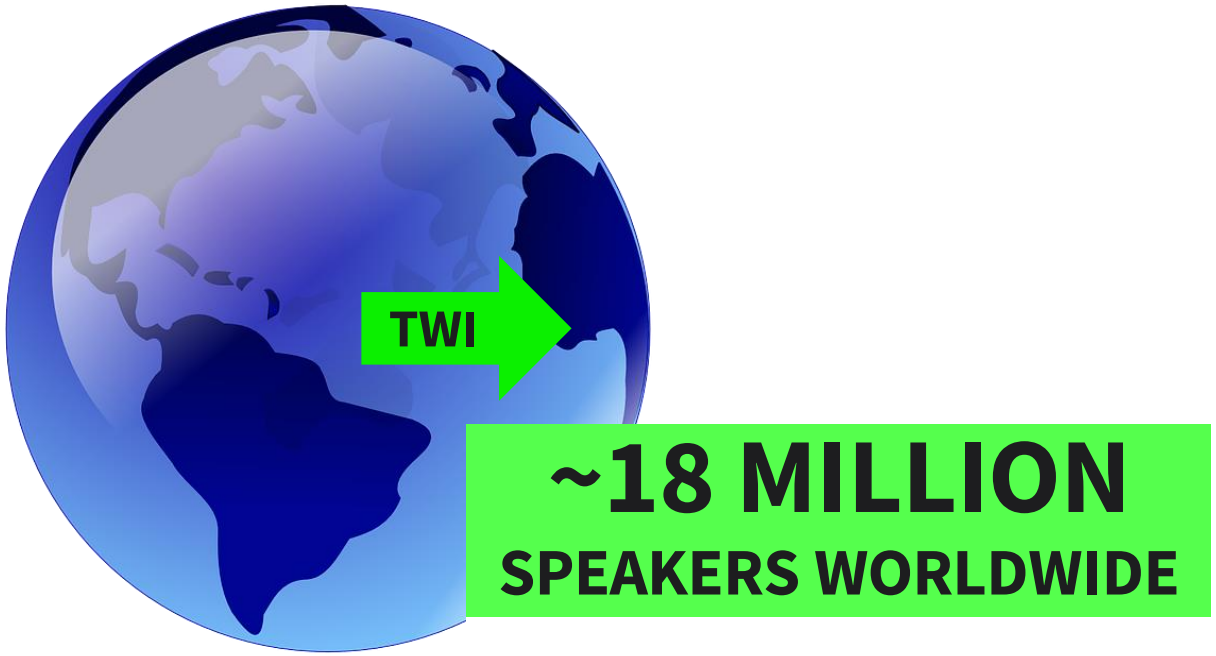
Llama 3

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

ZERO/FEW-SHOT CAPABILITIES: EASY TO USE WITH PROMPTING

Graphic Source: Custom Depiction.

THE LANGUAGE TWI



Text Source: Yakubu (2024)

THE LANGUAGE TWI

~18 MILLION
SPEAKERS WORLDWIDE

- **GHANA**
- **COTE D'IVOIRE**
- **BENIN**
- **SURINAME**
- **JAMAICA**
- **UNITED STATES**

Text Source: Yakubu (2024)

2

RELATED WORK

RELATED WORK: LLMs on African Languages

- ✓ No known publications explicitly describe the use of LLMs for Twi natural language processing tasks

RELATED WORK: LLMs on African Languages

- ✓ No known publications explicitly describe the use of LLMs for Twi natural language processing tasks
- ✓ But the use LLMs for other African languages, e.g.

  Ojo et al. (2024) analyzed 4 LLMs (mT0, Aya, LLaMA 2, GPT-4) on six tasks (topic classification, sentiment classification, machine translation, summarization, question answering, and named entity recognition) across 60 African languages.
  ➔ All LLMs performed worse on African languages compared to high-resource languages.
  ➔ LLMs were only evaluated from scratch, no few-shot learning.

RELATED WORK: LLMs on African Languages

- ✓ No known publications explicitly describe the use of LLMs for Twi natural language processing tasks
- ✓ But the use LLMs for other African languages, e.g.

  Ojo et al. (2024) analyzed 4 LLMs (mT0, Aya, LLaMA 2, GPT-4) on six tasks (topic classification, sentiment classification, machine translation, summarization, question answering, and named entity recognition) across 60 African languages.
  ➔ All LLMs performed worse on African languages compared to high-resource languages.
  ➔ LLMs were only evaluated from scratch, no few-shot learning.

  Sikiru et al. (2024) analyzed 3 LLMs (GPT-4.0, GPT-3.5, Bard) on answering basic financial questions in English and Yoruba.
  ➔ GPT-4.0 outperformed GPT-3.5 and Bard.
  ➔ LLMs show potential for financial queries but need improvement to better support Yoruba.

REL[illegible]RK: LLMs on African Languages

LLMS SEEM TO HAVE POTENTIAL

✓ No known publ[illegible]he use of LLMs for Twi natural language processing tasks

✓ But the use LLMs for other African lang[illegible]

Ojo et al. (2024) analyzed 4 LLMs (mT0, Aya, LLaMA 2, GP[illegible]ification, sentiment classification, machine translation, summarization, questi[illegible]d entity recognition) across 60 African languages.
➔ All LLMs performed worse on African languages compared to high-resource languages.
➔ LLMs were only evaluated from scratch, no few-shot learning.

Sikiru et al. (2024) analyzed 3 LLMs (GPT-4.0, GPT-3.5, Bard) on answering basic financial questions in English and Yoruba.
➔ GPT-4.0 outperformed GPT-3.5 and Bard.
➔ LLMs show potential for financial queries but need improvement to better support Yoruba.

REL[illegible]RK: LLMs on African Languages

LLMS SEEM TO HAVE POTENTIAL

NO AUTOMATIC SHORT ANSWER GRADING

- ✓ No known publ[illegible]he use of LLMs for Twi natural language processing tasks
- ✓ But [illegible]r African lang[illegible]

Ojo et al. (2024) a[illegible] LaMA 2, GP[illegible]sification, sentiment classification, na[illegible]ation, questi[illegible]d entity recognition) across 60 African lan[illegible]
➔ All LLMs performed worse on African language[illegible]languages.
➔ LLMs were only evaluated from scratch, no few-shot learn[illegible]

Sikiru et al. (2024) analyzed 3 LLMs (GPT-4.0, GPT-3.5, Bard) on answering basic fina[illegible]stions in English and Yoruba.
➔ GPT-4.0 outperformed GPT-3.5 and Bard.
➔ LLMs show potential for financial queries but need improvement to better support Yoruba.

REL[illegible]RK: LLMs on African Languages

LLMS SEEM TO HAVE POTENTIAL

NO AUTOMATIC SHORT ANSWER GRADING

NO FEW-SHORT LEARNING

✓ No known publ[illegible]he use of LLMs for Twi natural language processing tasks

✓ But [illegible]r African lang[illegible]

Ojo et al. (2024) a[illegible]LaMA 2, GP[illegible]ification, se[illegible]ification, fina[illegible]ation, questi[illegible]d e[illegible]60 African lang[illegible]

➔ All LL[illegible]n languages[illegible]languages.

➔ LLMs were only e[illegible]w-shot learn[illegible]

Sikiru et al. (2024) analyzed 3 LLMs (GPT-4.[illegible]ing basic fina[illegible]stions in English and Yoruba.

➔ GPT-4.0 outperformed GPT-3.5 and Bard.

➔ LLMs show potential for financial queries but need improvement to bett[illegible]

REL[illegible]RK: LLMs on African Languages

✓ No known publ[illegible] the use of LLMs for Twi natural language processing tasks

✓ But [illegible] African lang[illegible]

Ojo et al. (2024) a[illegible] LaMA 2, GPT[illegible]ification, se[illegible]ification, ina[illegible]ation, questi[illegible]d e[illegible]60 African lang[illegible]

➔ All LLM[illegible]n languages[illegible] languages.

➔ [illegible]were only ev[illegible]w-shot learning[illegible]

S[illegible]LMs (GPT-4.0[illegible]ing basic fina[illegible]stions in English a[illegible]

➔ GPT-4.0 outperforme[illegible]

➔ LLMs show potential for financia[illegible]ment to bette[illegible]

LLMS SEEM TO HAVE POTENTIAL

NO AUTOMATIC SHORT ANSWER GRADING

NO FEW-SHORT LEARNING

NO METHOD TO IMPROVE LLM OUTPUT

EXPERIMENTAL SETUP

EXPERIMENTAL SETUP

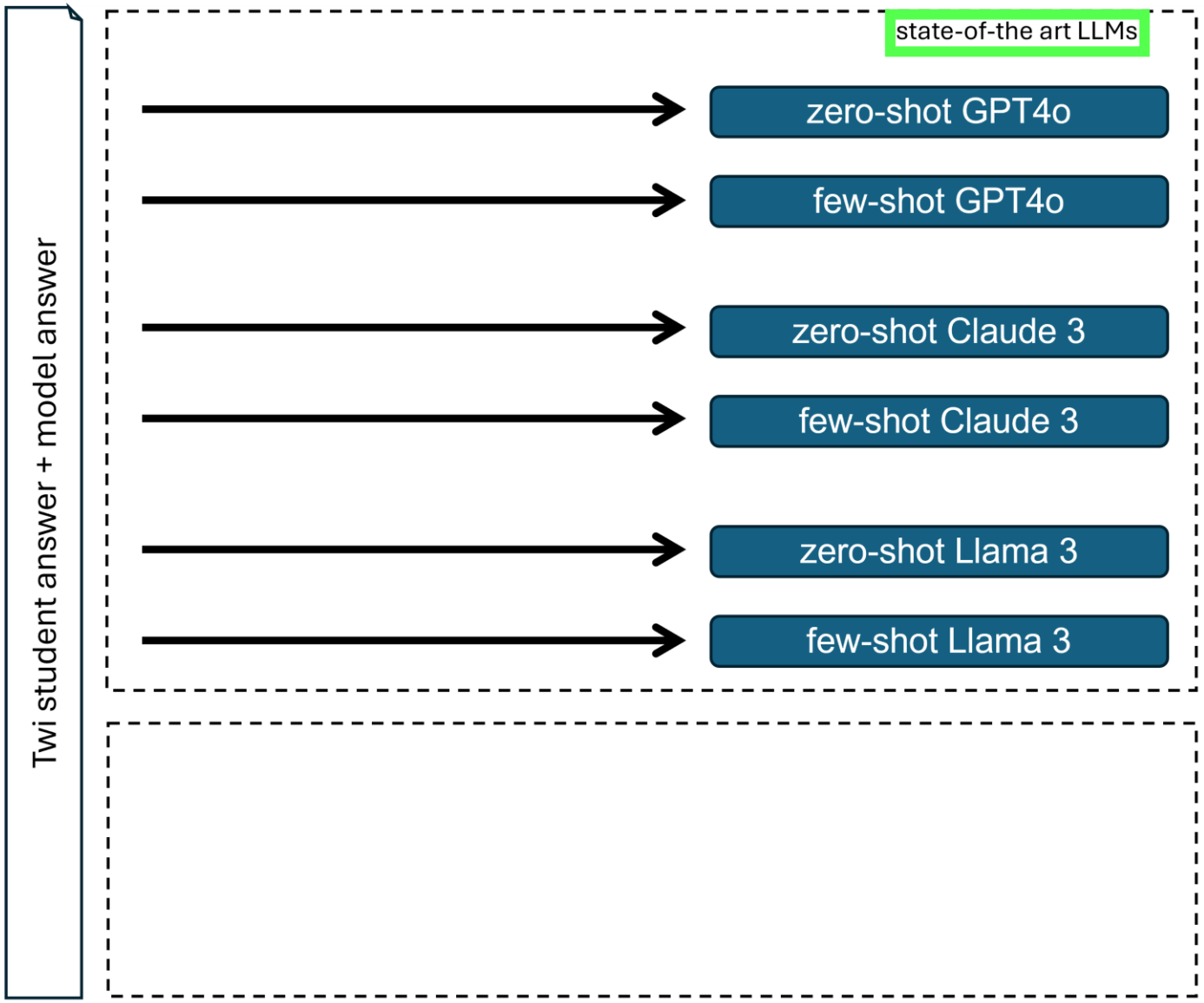
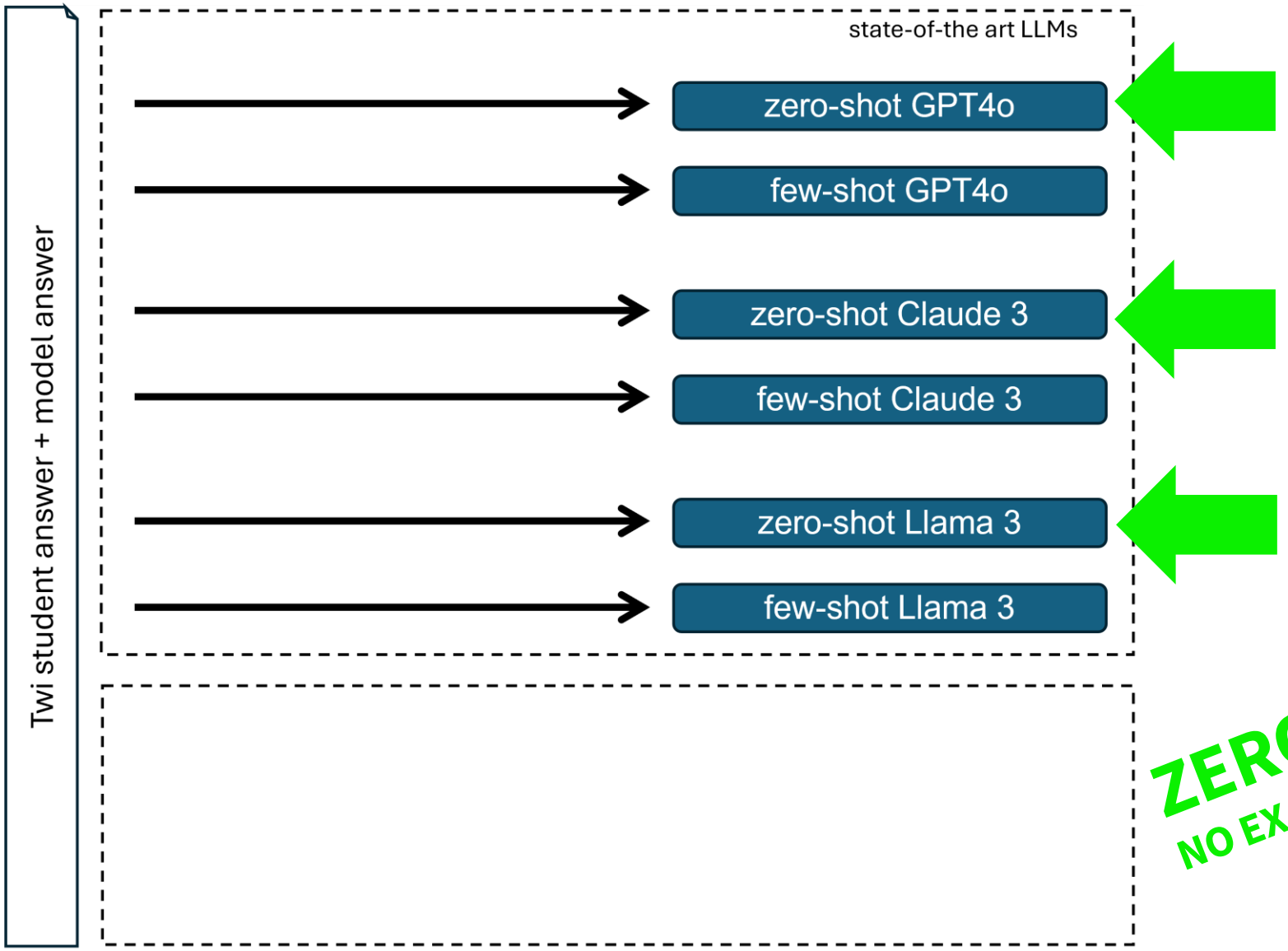


Image Source: Agyemang & Schlippe (2024).

EXPERIMENTAL SETUP



ZERO-SHOT
NO EXAMPLES TO LERN

Image Source: Agyemang & Schlippe (2024).

EXPERIMENTAL SETUP

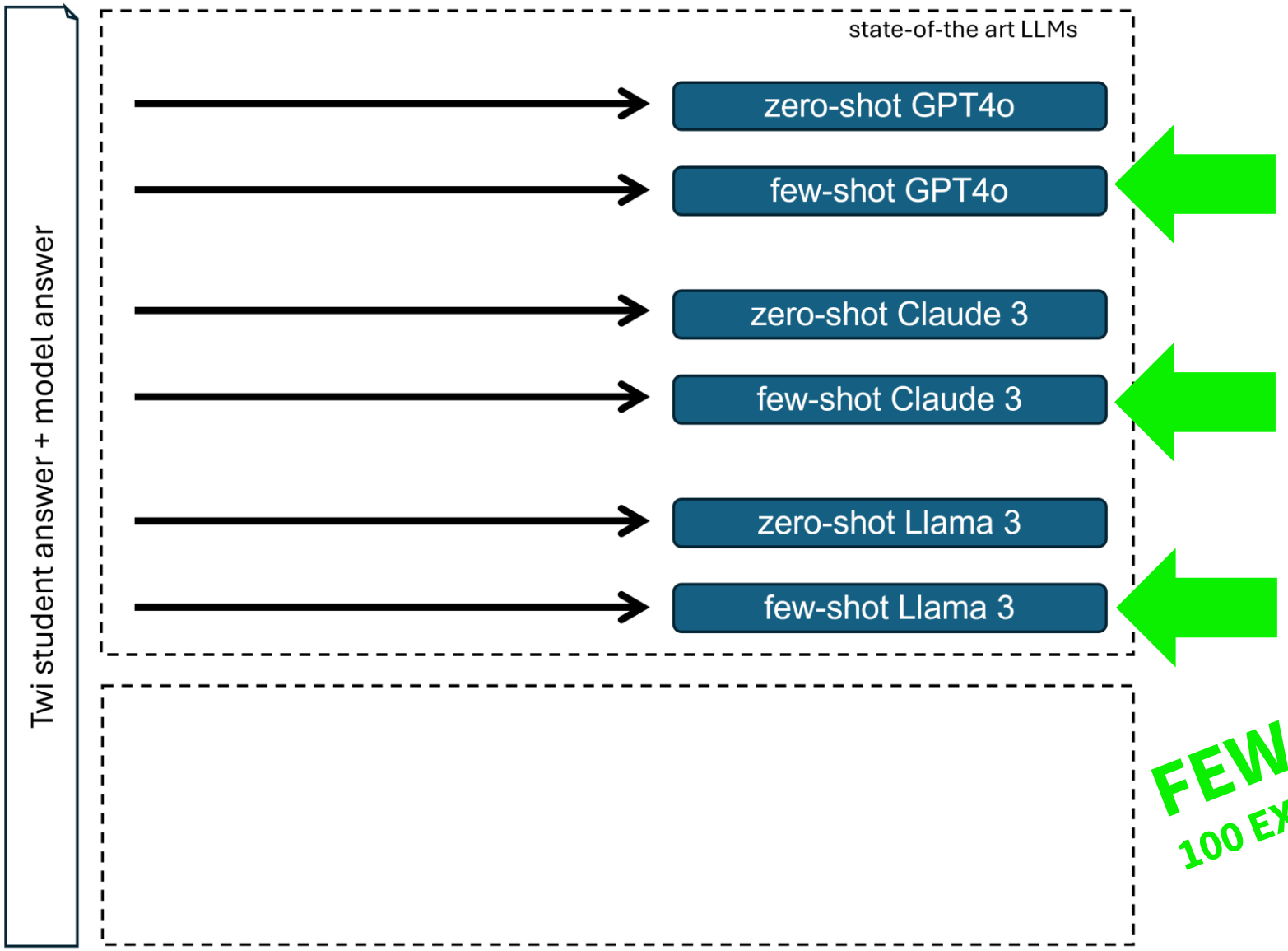


Image Source: Agyemang & Schlippe (2024).

EXPERIMENTAL SETUP

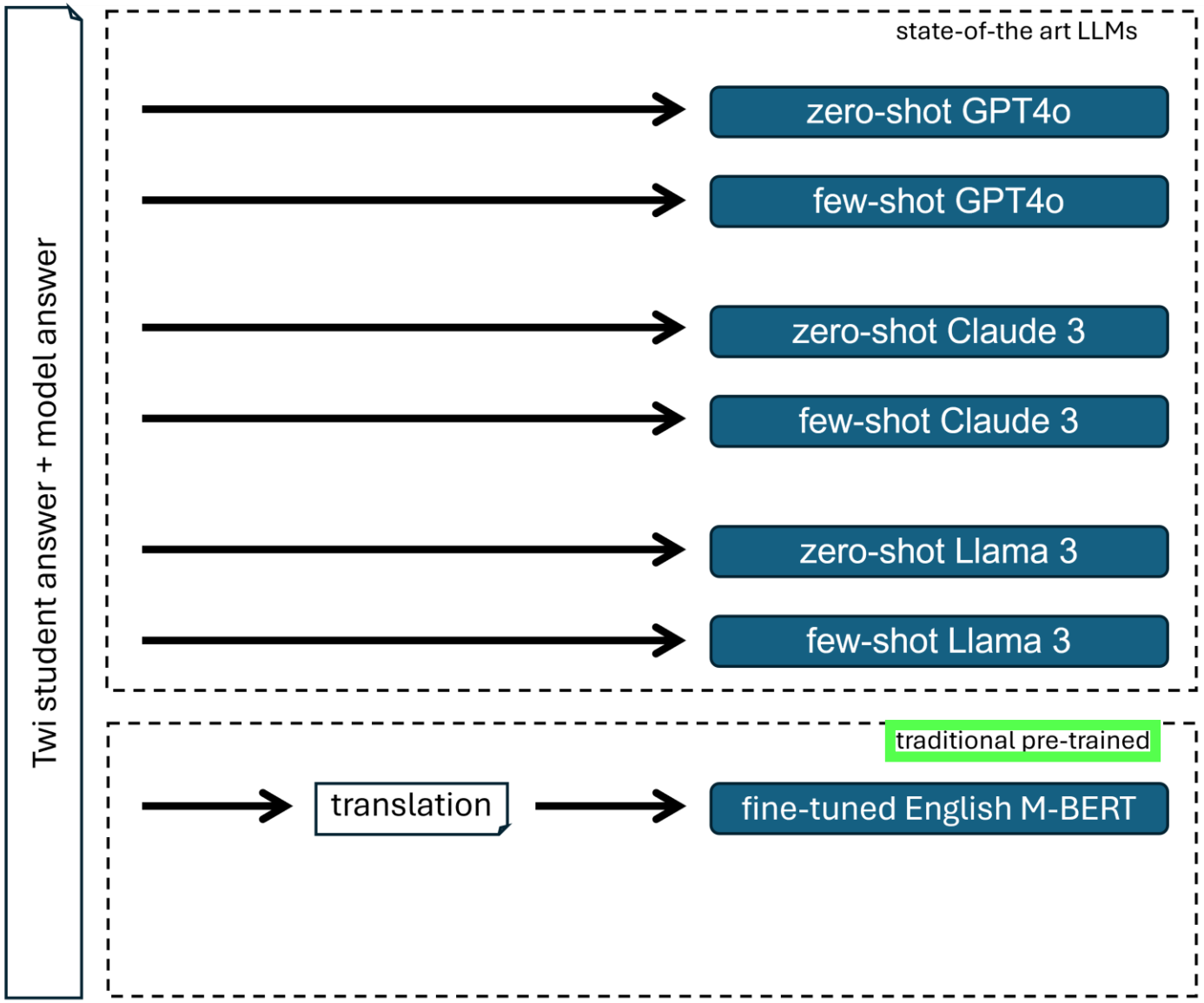


Image Source: Agyemang & Schlippe (2024).

EXPERIMENTAL SETUP

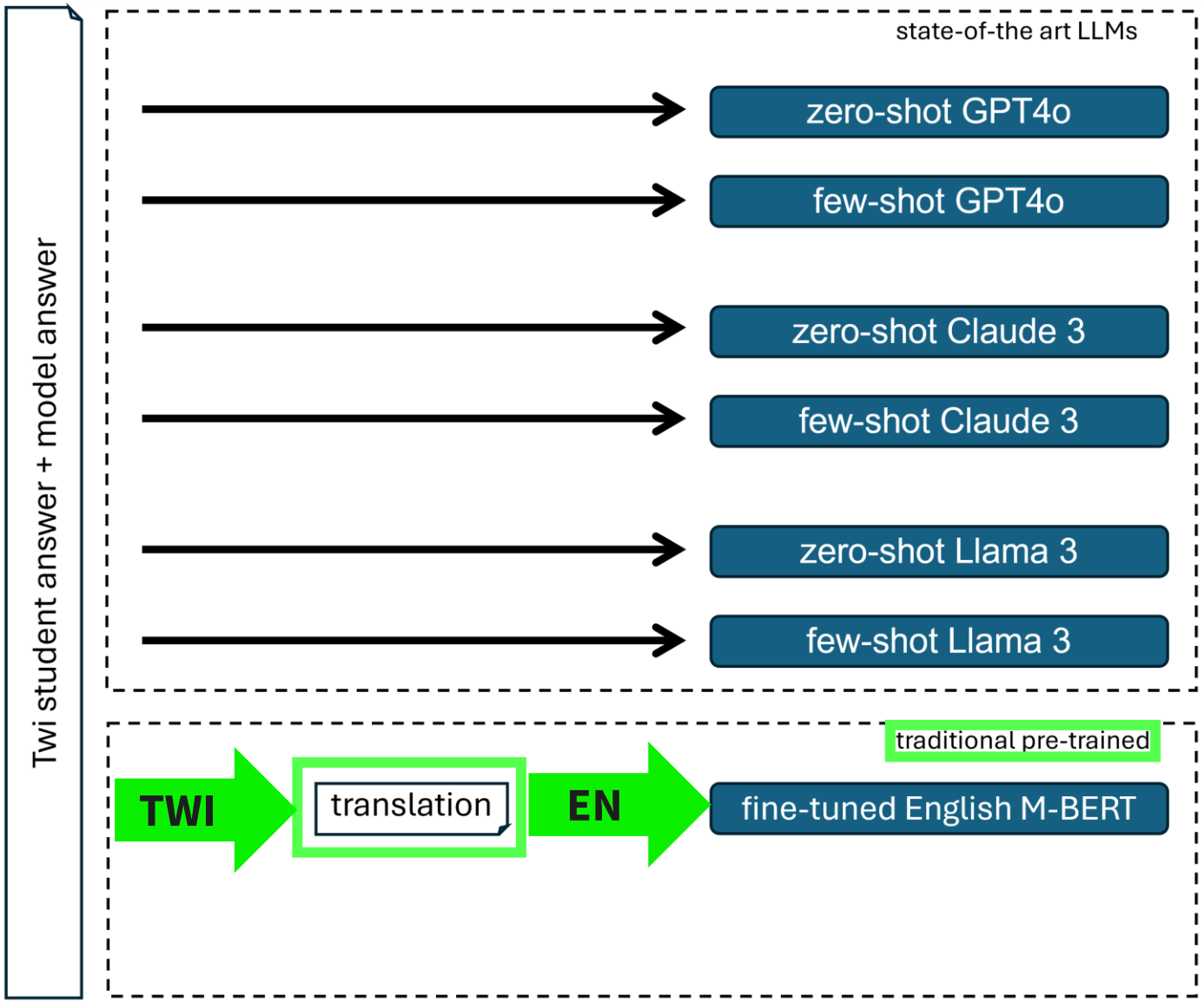


Image Source: Agyemang & Schlippe (2024).

EXPERIMENTAL SETUP

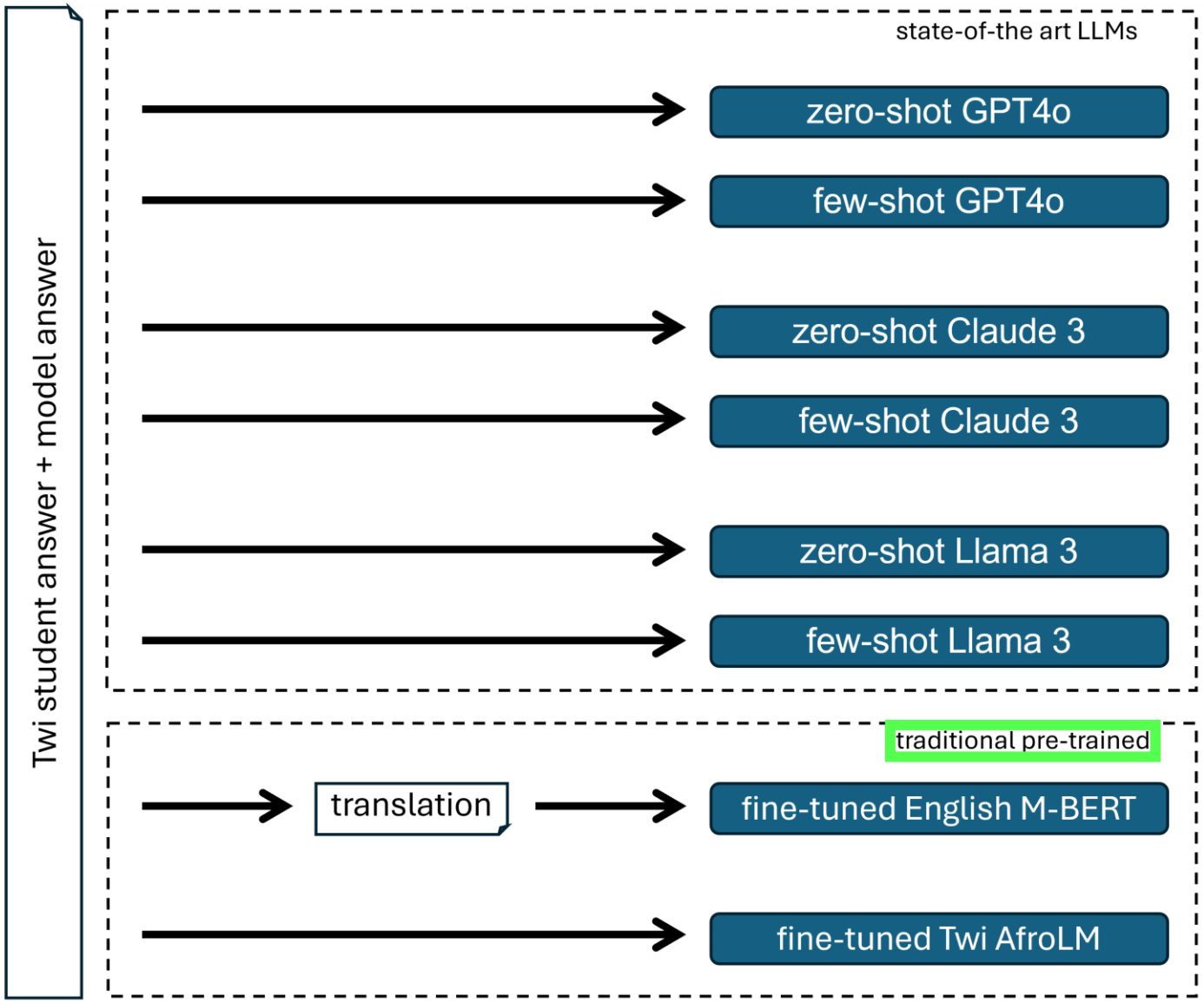
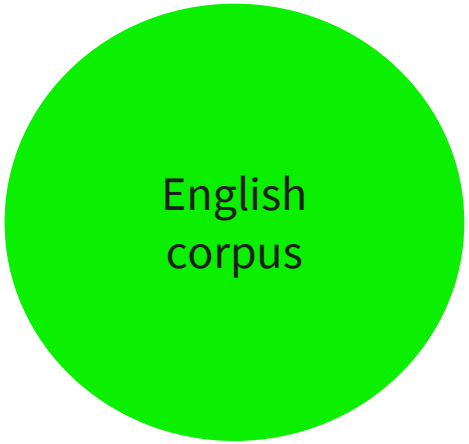


Image Source: Agyemang & Schlippe (2024).

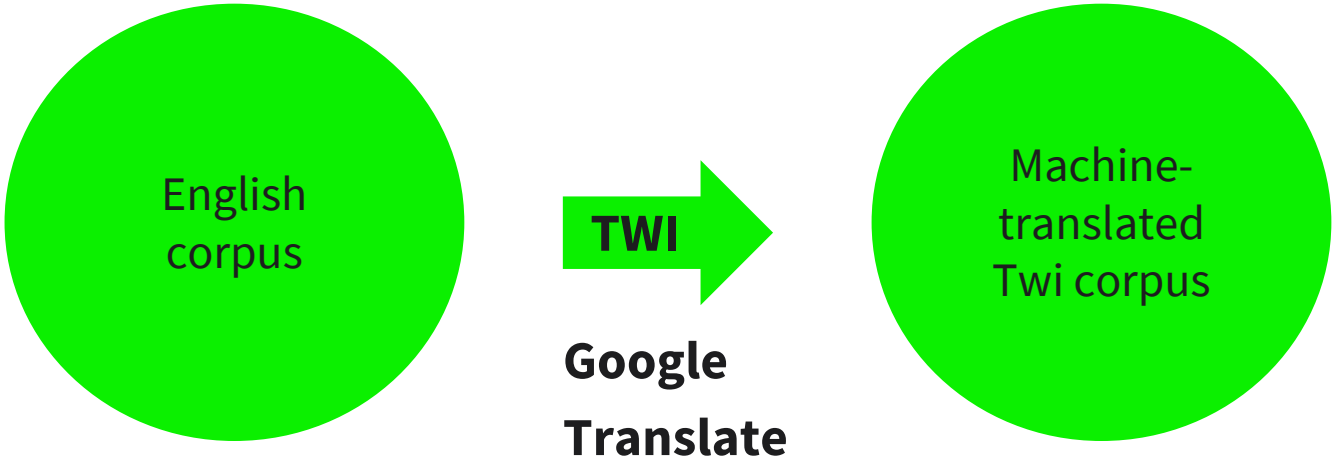
DATA



TWI AUTOMATIC SHORT ANSWER GRADING CORPUS BASED ON THE ENGLISH BENCHMARK CORPUS OF THE UNIVERSITY OF NORTH TEXAS

Graphic Source: Custom Depiction.

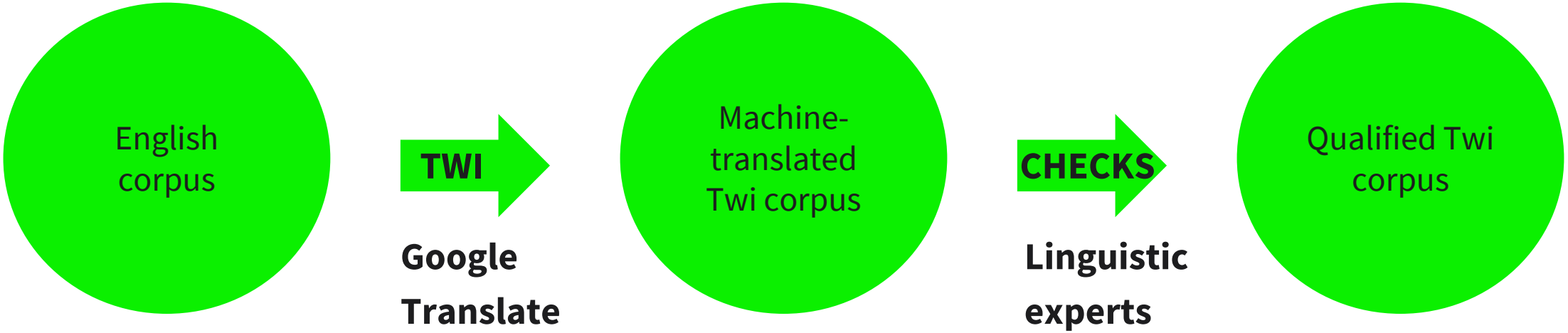
DATA



TWI AUTOMATIC SHORT ANSWER GRADING CORPUS BASED ON THE ENGLISH BENCHMARK CORPUS OF THE UNIVERSITY OF NORTH TEXAS

Graphic Source: Custom Depiction.

DATA



TWI AUTOMATIC SHORT ANSWER GRADING CORPUS BASED ON THE ENGLISH BENCHMARK CORPUS OF THE UNIVERSITY OF NORTH TEXAS

Graphic Source: Custom Depiction.

EXPERIMENTS AND RESULTS

PROMPT for zero-shot

Please grade the student answer, given the question and the model answer below. You can assign points between 0 and 5, where a completely correct answer received 5 points and a completely false answer receives 0 points. Decimal places are also possible. Please provide the question, model answer, student answer, and the number of points in a table format.

<test set in csv format>

Image Source: Agyemang & Schlippe (2024).

PROMPT for zero-shot

Please grade the student answer, given the question and the model answer below. You can assign points between 0 and 5, where a completely correct answer received 5 points and a completely false answer receives 0 points. Decimal places are also possible. Please provide the question, model answer, student answer, and the number of points in a table format.

<test set in csv format>

CSV
STUDENT ANSWER
MODEL ANSWER
EXAM QUESTION

Image Source: Agyemang & Schlippe (2024).

PROMPT for few-shot

Prompt for teaching how to grade:

> Here are examples of exam questions (question), model answers (desired_answer), student answers (student_answer), and corresponding grades (score_avg) in each line in csv format:
>
> <examples from training set in csv format>
>
> From these examples, please learn how to grade, i.e. assign scores for the student answers. The reason is that in the next prompt you will be provided with exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format, and you need to include the corresponding grades (score_avg) in each line.

Prompt for instructing to grade:

> Thanks for learning how to grade based on the examples provided. Now you will be provided with new exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format in Twi, and you need to include the corresponding grades (score_avg) in each line. Grade all the questions provided within the range of 0 to 5. you can assign decimal values. Display your output in table format with the headings "question", "desired_answer", "student_answer", and "score_avg".
>
> <test set in csv format>

CSV
STUDENT ANSWER
MODEL ANSWER
EXAM QUESTION
GRADER'S SCORE

Image Source: Agyemang & Schlippe (2024).

PROMPT for few-shot

Prompt for teaching how to grade:

> Here are examples of exam questions (question), model answers (desired_answer), student answers (student_answer), and corresponding grades (score_avg) in each line in csv format:
>
> <examples from training set in csv format>
>
> From these examples, please learn how to grade, i.e. assign scores for the student answers. The reason is that in the next prompt you will be provided with exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format, and you need to include the corresponding grades (score_avg) in each line.

Prompt for instructing to grade:

> Thanks for learning how to grade based on the examples provided. Now you will be provided with new exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format in Twi, and you need to include the corresponding grades (score_avg) in each line. Grade all the questions provided within the range of 0 to 5. you can assign decimal values. Display your output in table format with the headings "question", "desired_answer", "student_answer", and "score_avg".
>
> <test set in csv format>

CSV
STUDENT ANSWER
MODEL ANSWER
EXAM QUESTION

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

BEST AUTOMATIC SHORT ANSWER GRADING PERFORMANCE: AfroLM

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	$Lang_{Prompt}$	$Lang_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$GPT\text{-}4o_{zero-shot}$	TW	TW	2.36
TW-EN translation + $GPT\text{-}4o_{zero-shot}$	EN	EN	2.51
$GPT\text{-}4o_{zero-shot}$	EN	TW	2.15
$GPT\text{-}4o_{few-shot}$	EN	TW	1.53
Claude 3.5 $Sonnet_{zero-shot}$	EN	TW	1.01
Claude 3.5 $Sonnet_{few-shot}$	EN	TW	1.00
$LLaMa\ 3_{zero-shot}$	EN	TW	1.48
$LLaMa\ 3_{few-shot}$	EN	TW	1.11

2ND-BEST AUTOMATIC SHORT ANSWER GRADING PERFORMANCE: TW-EN translation + M-BERT

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
GPT-4o$_{zero-shot}$	TW	TW	2.36
TW-EN translation + GPT-4o$_{zero-shot}$	EN	EN	2.51
GPT-4o$_{zero-shot}$	EN	TW	2.15
GPT-4o$_{few-shot}$	EN	TW	1.53
Claude 3.5 Sonnet$_{zero-shot}$	EN	TW	1.01
Claude 3.5 Sonnet$_{few-shot}$	EN	TW	1.00
LLaMa 3$_{zero-shot}$	EN	TW	1.48
LLaMa 3$_{few-shot}$	EN	TW	1.11

0.75 POINTS

HUMAN GRADER VARIABILITY

Image Source: Agyemang & Schlippe (2024); Schlippe & Sawatzki (2022).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	$Lang_{Prompt}$	$Lang_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$GPT\text{-}4o_{zero-shot}$	TW	TW	2.36
TW-EN translation + $GPT\text{-}4o_{zero-shot}$	EN	EN	2.51
$GPT\text{-}4o_{zero-shot}$	EN	TW	2.15
$GPT\text{-}4o_{few-shot}$	EN	TW	1.53
Claude 3.5 $Sonnet_{zero-shot}$	EN	TW	1.01
Claude 3.5 $Sonnet_{few-shot}$	EN	TW	1.00
$LLaMa\ 3_{zero-shot}$	EN	TW	1.48
$LLaMa\ 3_{few-shot}$	EN	TW	1.11

IS LLMS' AUTOMATIC SHORT ANSWER GRADING PERFORMANCE BETTER IF LangPrompt = TW or EN?
IF LangtoGrade = TW or EN?

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	$Lang_{Prompt}$	$Lang_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
GPT-4o$_{zero-shot}$	TW	TW	2.36
TW-EN translation + GPT-4o$_{zero-shot}$	EN	EN	2.51
GPT-4o$_{zero-shot}$	EN	TW	2.15
GPT-4o$_{few-shot}$	EN	TW	1.53
Claude 3.5 Sonnet$_{zero-shot}$	EN	TW	1.01
Claude 3.5 Sonnet$_{few-shot}$	EN	TW	1.00
LLaMa 3$_{zero-shot}$	EN	TW	1.48
LLaMa 3$_{few-shot}$	EN	TW	1.11

BEST LLM'S AUTOMATIC SHORT ANSWER GRADING PERFORMANCE WITH GTP-4o: LangPrompt = EN and LangtoGrade = TW

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

IMPACT OF PROVIDING EXAMPLES TO LLMS: ZERO-SHOT VS. FEW SHOT

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF SINGLE LLMS

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MAE
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

IMPACT OF PROVIDING EXAMPLES TO LLMS: FEW SHOT !!!

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF COMBINED LLMS

Model Combination	MAE
Pre-trained LMs (AfroLM + M-BERT)	**0.64**
$\text{LLMs}_{zero-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.26
$\text{LLMs}_{few-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.10
Pre-trained LMs + $\text{LLMs}_{few-shot}$	0.99
Pre-trained LMs + $\text{LLMs}_{few-shot}$ + $\text{LLMs}_{zero-shot}$	1.15

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MER
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
GPT-4o$_{zero-shot}$	TW	TW	2.36
TW-EN translation + GPT-4o$_{zero-shot}$	EN	EN	2.51
GPT-4o$_{zero-shot}$	EN	TW	2.15
GPT-4o$_{few-shot}$	EN	TW	1.53
Claude 3.5 Sonnet$_{zero-shot}$	EN	TW	1.01
Claude 3.5 Sonnet$_{few-shot}$	EN	TW	1.00
LLaMa 3$_{zero-shot}$	EN	TW	1.48
LLaMa 3$_{few-shot}$	EN	TW	1.11

IMPACT OF COMBINING OUTPUTS?

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF COMBINED LLMS

Model Combination	MAE
Pre-trained LMs (AfroLM + M-BERT)	**0.64**
$\text{LLMs}_{zero-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.26
$\text{LLMs}_{few-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.10
Pre-trained LMs + $\text{LLMs}_{few-shot}$	0.99
Pre-trained LMs + $\text{LLMs}_{few-shot}$ + $\text{LLMs}_{zero-shot}$	1.15

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MER
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

IMPACT OF COMBINING OUTPUTS: IMPROVEMENT

Image Source: Agyemang & Schlippe (2024).

RESULTS: MEAN ABSOLUTE ERROR RATES OF COMBINED LLMS

Model Combination	MAE
Pre-trained LMs (AfroLM + M-BERT)	**0.64**
$\text{LLMs}_{zero-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.26
$\text{LLMs}_{few-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.10
Pre-trained LMs + $\text{LLMs}_{few-shot}$	0.99
Pre-trained LMs + $\text{LLMs}_{few-shot}$ + $\text{LLMs}_{zero-shot}$	1.15

Model	Lang_{Prompt}	$\text{Lang}_{toGrade}$	MER
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
$\text{GPT-4o}_{zero-shot}$	TW	TW	2.36
TW-EN translation + $\text{GPT-4o}_{zero-shot}$	EN	EN	2.51
$\text{GPT-4o}_{zero-shot}$	EN	TW	2.15
$\text{GPT-4o}_{few-shot}$	EN	TW	1.53
$\text{Claude 3.5 Sonnet}_{zero-shot}$	EN	TW	1.01
$\text{Claude 3.5 Sonnet}_{few-shot}$	EN	TW	1.00
$\text{LLaMa 3}_{zero-shot}$	EN	TW	1.48
$\text{LLaMa 3}_{few-shot}$	EN	TW	1.11

IMPACT OF COMBINING OUTPUTS: IMPROVEMENT

Image Source: Agyemang & Schlippe (2024).

5

CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.
- However, training and fine-tuning such systems can rather be done by experts.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.
- However, training and fine-tuning such systems can rather be done by experts.
- Providing LLMs with a few examples improves performance.

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.
- However, training and fine-tuning such systems can rather be done by experts.
- Providing LLMs with a few examples improves performance.

Future Work

- Investigate weighted combination of LLM outputs

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.
- However, training and fine-tuning such systems can rather be done by experts.
- Providing LLMs with a few examples improves performance.

Future Work

- Investigate weighted combination of LLM outputs
- Explainable AI

CONCLUSION AND FUTURE WORK

Conclusion

- We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi.
- Lack of a Twi corpus ➔ translated and validated the English University of North Texas benchmark corpus to create the first Twi ASAG dataset.
- Traditional pre-trained models AfroLM and M-BERT outperform individual LLMs in terms of MAE, particularly when their outputs are combined.
- Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance.
- However, training and fine-tuning such systems can rather be done by experts.
- Providing LLMs with a few examples improves performance.

Future Work

- Investigate weighted combination of LLM outputs
- Explainable AI
- Collect a real Twi ASAG corpus and analyze the Twi ASAG Performance in exams in other subjects.

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

THANK YOU

Tim Schlippe
tim.schlippe@iu.org

REFERENCES

Literature

- **Quarterly Labour Force (2022)**: *Survey – Quarter 2: 2022*
- **Lorraine, M., Molapo, R. (2014):** *South Africa's Challenges of Realising Her Socio-Economic Rights.* Mediterranean Journal of Social Sciences 5
- **Nkomo, S. (2017)**: *Public Service Delivery in South Africa Councillors and Citizens Critical Links in Overcoming Persistent Inequities.* Tech. Rep. 42, Afrobarometer
- **Sustainable Development Goals (2019):** *Country Report 2019 – South Afrcia.* Tech. Rep. ISBN 978-0-621-47619-4, Statistics South Africa
- **United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021)**: https://www.un.org/sustainabledevelopment/sustainable-development-goals
- **South African Government (2022)**: *National Departments*: https://www.gov.za/about-government/government-system/national-departments
- **Mabokela, K. R. & Schlippe, T. (2022)**: *AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa*, The South African Conference for Artificial Intelligence Research (SACAIR 2022), Stellenbosch, South Africa

Images

- **Images provided by OpenClipart-Vectors/429883177/Shutterstock**. (https://www.shutterstock.com/image-photo/kiev-ukraine-may-30-2016-collection-429883177 [last access: 11/29/2022])

REFERENCES

Literature

- **Kiritchenko, S., Mohammad, S.M. (2018)**: *Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems*. ArXiv abs/1805.04508
- **Kaur, C., Sharma, A. (2020)**: *Sentiment Analysis of Tweets on Social Issues using Machine Learning Approach*. International Journal of Advanced Trends in Computer Science and Engineering 9, 6303–6311 (08 2020).
- **Balahur, A. and Turchi, M. (2014)**: *Comparative Experiments using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis*. Comput. Speech Lang., 28:56–75.
- **Nguyen, P. X. V., Hong, T. V. T., Nguyen, K. V., and Nguyen, N. L.-T. (2018)**: *Deep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus*. 5th NAFOSTED Conference on Information and Computer Science (NICS).
- **Kumar, A. and Sharan, A., (2020)**: *Deep Learning-Based Frameworks for Aspect-Based Sentiment Analysis*, pages 139–158. Springer Singapore.
- **Rakhmanov, O. (2020)**: *A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments*. Procedia Computer Science, 178:194–204.
- **Kolchyna, O., Souza, T. T. P., Treleaven, P. C., and Aste, T. (2015)**: *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*. arXiv: Computation and Language.

REFERENCES

Literature

- **Kotelnikova, A., Paschenko, D., Bochenina, K., and Kotelnikov, E. (2021)**. *Lexicon-based Methods vs. BERT for Text Sentiment Analysis*. In AIST.
- **Lin, Z., Jin, X., Xu, X., Wang, Y., Tan, S., and Cheng, X. (2014)**: *Make It Possible: Multilingual Sentiment Analysis without Much Prior Knowledge*. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), volume 2, pages 79–86.
- **Vilares, D., Alonso Pardo, M., and Gómez-Rodríguez, C. (2017)**: *Supervised Sentiment Analysis in Multilingual Environments*. Information Processing Management, 53, 05.