Information Asymmetry Across Language Varieties: A Case Study on Cantonese-Mandarin and Bavarian-German QA

Renhao Pei[3,4,*], Siyao Peng[1,2,*], Verena Blaschke[1,2], Robert Litschko[1,2], Barbara Plank[1,2]

[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Germany
[3]ELLIS Institute Finland
[4]University of Turku, Finland

April 1, 2026

Outline

Introduction

On Wikipedia, information about the same topic vary a lot across languages:

Introduction

On Wikipedia, information about the same topic vary a lot across languages:

- High-resource languages have more wikipedia pages and more detailed content

Introduction

On Wikipedia, information about the same topic vary a lot across languages:

- High-resource languages have more wikipedia pages and more detailed content
- But some **regional or local information** is found **only** in Wikipedia pages in local languages

Introduction

Lead section (Bavarian): Es Hiatamadl is a traditionella Voiksdanz, wo in Östareich und Bayern weit vabroadd is. En Danz gibts in vaschiedna Variantn. Da Nama kimmt vom Danzliadl "Koa Hiatamadl mog i net". Da Danz hod de sejm Wuazln wia da Strohschneida, wo iwa ganz Eiropa vabroadd wor. *The Hiatamadl is a traditional folk dance widespread throughout Austria and Bavaria. There are various variations of the dance. The name comes from the dance tune "Koa Hiatamadl mog i net (I don't like shepherd girls)." The dance has the same roots as the Strohschneider (straw cutter), which was widespread throughout Europe.*

Lead section (German): Hiatamadl ist der Name eines alpenländischen Volkstanzes, der sich in Österreich und Bayern großer Beliebtheit erfreut. *Hiatamadl is the name of an Alpine folk dance that is very popular in Austria and Bavaria.*

Introduction

Lead section (Bavarian): Es Hiatamadl is a traditionella Voiksdanz, wo in Östareich und Bayern weit vabroadd is. En Danz gibts in vaschiedna Variantn. Da Nama kimmt vom Danzliadl "Koa Hiatamadl mog i net". Da Danz hod de sejm Wuazln wia da Strohschneida, wo iwa ganz Eiropa vabroadd wor. *The Hiatamadl is a traditional folk dance widespread throughout Austria and Bavaria. There are various variations of the dance. The name comes from the dance tune "Koa Hiatamadl mog i net (I don't like shepherd girls)." The dance has the same roots as the Strohschneider (straw cutter), which was widespread throughout Europe.*

Lead section (German): Hiatamadl ist der Name eines alpenländischen Volkstanzes, der sich in Österreich und Bayern großer Beliebtheit erfreut. *Hiatamadl is the name of an Alpine folk dance that is very popular in Austria and Bavaria.*

Question 1: Woher kommt der Name des Tanzes Hiatamadl? *Where does the name of the dance Hiatamadl come from?*

Question 2: Aus welchem Tanz hat sich das Hiatamadl entwickelt? *From which dance did the Hiatamadl develop?*

Introduction

Lead section (Bavarian): Es Hiatamadl is a traditionella Voiksdanz, wo in Östareich und Bayern weit vabroadd is. En Danz gibts in vaschiedna Variantn. Da Nama kimmt vom Danzliadl "Koa Hiatamadl mog i net". Da Danz hod de sejm Wuazln wia da Strohschneida, wo iwa ganz Eiropa vabroadd wor. *The Hiatamadl is a traditional folk dance widespread throughout Austria and Bavaria. There are various variations of the dance. The name comes from the dance tune "Koa Hiatamadl mog i net (I don't like shepherd girls)." The dance has the same roots as the Strohschneider (straw cutter), which was widespread throughout Europe.*
Lead section (German): Hiatamadl ist der Name eines alpenländischen Volkstanzes, der sich in Österreich und Bayern großer Beliebtheit erfreut. *Hiatamadl is the name of an Alpine folk dance that is very popular in Austria and Bavaria.*
Question 1: Woher kommt der Name des Tanzes Hiatamadl? *Where does the name of the dance Hiatamadl come from?* **Answer 1**: Der Name des Tanzes kommt vom Tanzlied "Koa Hiatamadl mog i net". *The name of the dance comes from the dance tune "Koa Hiatamadl mog i net".*
Question 2: Aus welchem Tanz hat sich das Hiatamadl entwickelt? *From which dance did the Hiatamadl develop?*

Table: Sample WiLoVa-QA annotations for `bar-deu` page "Hiatamadl."

Introduction

Lead section (Bavarian): Es Hiatamadl is a traditionella Voiksdanz, wo in Östareich und Bayern weit vabroadd is. En Danz gibts in vaschiedna Variantn. Da Nama kimmt vom Danzliadl "Koa Hiatamadl mog i net". Da Danz hod de sejm Wuazln wia da Strohschneida, wo iwa ganz Eiropa vabroadd wor. *The Hiatamadl is a traditional folk dance widespread throughout Austria and Bavaria. There are various variations of the dance. The name comes from the dance tune "Koa Hiatamadl mog i net (I don't like shepherd girls)." The dance has the same roots as the Strohschneider (straw cutter), which was widespread throughout Europe.*

Lead section (German): Hiatamadl ist der Name eines alpenländischen Volkstanzes, der sich in Österreich und Bayern großer Beliebtheit erfreut. *Hiatamadl is the name of an Alpine folk dance that is very popular in Austria and Bavaria.*

Question 1: Woher kommt der Name des Tanzes Hiatamadl? *Where does the name of the dance Hiatamadl come from?* **Answer 1**: Der Name des Tanzes kommt vom Tanzlied "Koa Hiatamadl mog i net". *The name of the dance comes from the dance tune "Koa Hiatamadl mog i net".*

Question 2: Aus welchem Tanz hat sich das Hiatamadl entwickelt? *From which dance did the Hiatamadl develop?*

Answer 2: Das Hiatamadl hat dieselben Wurzeln wie der Strohschneider, der in ganz Europa verbreitet war. *The Hiatamadl has the same roots as the Strohschneider, which was widespread throughout Europe.*

Table: Sample WiLoVa-QA annotations for `bar-deu` page "Hiatamadl."

Introduction: LLM cross-lingual information retrieval

- Besides Wikipedia, Large Language Models (LLMs) are also commonly used channels for information

Introduction: LLM cross-lingual information retrieval

- Besides Wikipedia, Large Language Models (LLMs) are also commonly used channels for information
- Cross-lingual retrieval for information-seeking questions remains challenging for LLMs:
  - LLMs struggle to share knowledge when the question is proposed in a language different from the context language (Goldman et al., 2025)

Introduction: LLM cross-lingual information retrieval

- Besides Wikipedia, Large Language Models (LLMs) are also commonly used channels for information
- Cross-lingual retrieval for information-seeking questions remains challenging for LLMs:
  - LLMs struggle to share knowledge when the question is proposed in a language different from the context language (Goldman et al., 2025)
- Research Question: How well can LLMs answer questions based on knowledge found in local-language Wikipedia pages but absent from standard-language pages?

Introduction: local vs. standard

- To address this question, we conduct a case study on two pairs of higher-resource (*standard*) versus lower-resource (*local*) language varieties:
  - Mandarin Chinese (ISO 639-3: `cmn`) vs. Cantonese (`yue`)
  - German (`deu`) vs. Bavarian (`bar`)

Introduction: local vs. standard

- To address this question, we conduct a case study on two pairs of higher-resource (*standard*) versus lower-resource (*local*) language varieties:
  - Mandarin Chinese (ISO 639-3: `cmn`) vs. Cantonese (`yue`)
  - German (`deu`) vs. Bavarian (`bar`)
- Both pairs are spoken in overlapping regions, where the standard language prevails in formal communication, while the local variety serves as a marker of regional and cultural identity

Introduction: local vs. standard

- To address this question, we conduct a case study on two pairs of higher-resource (*standard*) versus lower-resource (*local*) language varieties:
  - Mandarin Chinese (ISO 639-3: `cmn`) vs. Cantonese (`yue`)
  - German (`deu`) vs. Bavarian (`bar`)
- Both pairs are spoken in overlapping regions, where the standard language prevails in formal communication, while the local variety serves as a marker of regional and cultural identity
- We create WiLoVa-QA (Wikipedia Local Variety Question Answering), a QA dataset for assessing LLMs' capability in retrieving information that...
  - appears on a Wikipedia page in the **local** language variety
  - but is absent from its **standard** variety counterpart

Outline

WiLoVa-QA dataset: preprocessing

- Aligning Wikipedia pages:
  - Extract and align Wikipedia pages for Cantonese-Mandarin (`yue-cmn`) and Bavarian-German (`bar-deu`)
  - Covering the majority of the local pages, 76.2% (111.6K/146.3K) of Cantonese and 90.4% (24.6K/27.2K) of Bavarian

WiLoVa-QA dataset: preprocessing

- Aligning Wikipedia pages:
  - Extract and align Wikipedia pages for Cantonese-Mandarin (`yue-cmn`) and Bavarian-German (`bar-deu`)
  - Covering the majority of the local pages, 76.2% (111.6K/146.3K) of Cantonese and 90.4% (24.6K/27.2K) of Bavarian
- Local-heavy filtering:
  - Only keep Wikipedia pairs for which both the *lead section* (introductory paragraphs) and the whole document in the local language edition are longer
  - After filtering, 5.38% (6.2K/111.6K) and 4.07% (1.0K/24.6K) of `yue-cmn` and `bar-deu` pairs remained

WiLoVa-QA dataset: preprocessing

- Aligning Wikipedia pages:
  - Extract and align Wikipedia pages for Cantonese-Mandarin (`yue-cmn`) and Bavarian-German (`bar-deu`)
  - Covering the majority of the local pages, 76.2% (111.6K/146.3K) of Cantonese and 90.4% (24.6K/27.2K) of Bavarian
- Local-heavy filtering:
  - Only keep Wikipedia pairs for which both the *lead section* (introductory paragraphs) and the whole document in the local language edition are longer
  - After filtering, 5.38% (6.2K/111.6K) and 4.07% (1.0K/24.6K) of `yue-cmn` and `bar-deu` pairs remained
- Removing page pairs are not ideal for our annotation, such as:
  - *translated* (the local page seems to be a direct translation of the standard page)
  - *structural mismatch* (either the local page or the standard page is a disambiguation page or a page without a lead section)

WiLoVa-QA dataset: preprocessing

- Aligning Wikipedia pages:
  - Extract and align Wikipedia pages for Cantonese-Mandarin (`yue-cmn`) and Bavarian-German (`bar-deu`)
  - Covering the majority of the local pages, 76.2% (111.6K/146.3K) of Cantonese and 90.4% (24.6K/27.2K) of Bavarian
- Local-heavy filtering:
  - Only keep Wikipedia pairs for which both the *lead section* (introductory paragraphs) and the whole document in the local language edition are longer
  - After filtering, 5.38% (6.2K/111.6K) and 4.07% (1.0K/24.6K) of `yue-cmn` and `bar-deu` pairs remained
- Removing page pairs are not ideal for our annotation, such as:
  - *translated* (the local page seems to be a direct translation of the standard page)
  - *structural mismatch* (either the local page or the standard page is a disambiguation page or a page without a lead section)
- As a result, 137 `yue-cmn` and 94 `bar-deu` article pairs were selected as containing distinctive content in the local page for QA annotation

WiLoVa-QA dataset: lead-section annotation

- Two master's students in computational linguistics are recruited as annotators, one fluent in Cantonese and Mandarin and one fluent in Bavarian and German

WiLoVa-QA dataset: lead-section annotation

- Two master's students in computational linguistics are recruited as annotators, one fluent in Cantonese and Mandarin and one fluent in Bavarian and German
- Information-asymmetry annotation framed as QA task:
  - For each filtered Wikipedia article pair, we ask annotators to identify 2-3 pieces information present only on lead section of local page but absent from the standard

WiLoVa-QA dataset: lead-section annotation

- Two master's students in computational linguistics are recruited as annotators, one fluent in Cantonese and Mandarin and one fluent in Bavarian and German
- Information-asymmetry annotation framed as QA task:
  - For each filtered Wikipedia article pair, we ask annotators to identify 2-3 pieces information present only on lead section of local page but absent from the standard
- Annotators are instructed to:
  - formulate a *question* in the standard language that addresses the information
  - *answer* the question in both the standard language and the local variety
  - *highlight* texts on the local section from which the answer was drawn

WiLoVa-QA dataset: lead-section annotation

- Two master's students in computational linguistics are recruited as annotators, one fluent in Cantonese and Mandarin and one fluent in Bavarian and German
- Information-asymmetry annotation framed as QA task:
  - For each filtered Wikipedia article pair, we ask annotators to identify 2-3 pieces information present only on lead section of local page but absent from the standard
- Annotators are instructed to:
  - formulate a *question* in the standard language that addresses the information
  - *answer* the question in both the standard language and the local variety
  - *highlight* texts on the local section from which the answer was drawn
- Result: 294 `yue-cmn` and 178 `bar-deu` lead section QA annotations

WiLoVa-QA dataset: document-level annotation

- Furthermore, annotators are aksed to skim through the whole Wikipedia page of both language varieties and find 1 piece of information from the local page that is:
  - not on the standard page
  - not already annotated at the lead section level

WiLoVa-QA dataset: document-level annotation

- Furthermore, annotators are aksed to skim through the whole Wikipedia page of both language varieties and find 1 piece of information from the local page that is:
  - not on the standard page
  - not already annotated at the lead section level
- Result 80 `yue-cmn` and 46 `bar-deu` document-level QA annotations

Constructing WiLoVa-QA Dataset: Statistics

language	Bavarian	German	Cantonese	Mandarin
ISO 639-3	`bar`	`deu`	`yue`	`cmn`
#speakers	13.7M	134.0M	85.7M	1.2B
(#rank)[1]	(#92)	(#12)	(#25)	(#2)
#wiki-pages[2]	27.2K	3,053.0K	146.3K	1,501.5K
#aligned-pages	24.6K		111.6K	
#local-heavy-pages	1,038		6,173	
#inspected-pages	497		229	
#qa-pages	94		137	
#lead section-qa	178		294	
#doc-qa	46		80	

Table: Statistics of the WiLoVa-QA dataset and included language varieties.

[1] According to The Ethnologue 200 (Eberhard et al., 2025) `https://www.ethnologue.com/insights/ethnologue200/`.

[2] As of September 23, 2025.

Outline

Experiments: set-up

State-of-the-art LLMs' are evaluated for QA, where answers are absent from the Wikipedia page in the prompting languages (Mandarin Chinese and German)

Experiments: set-up

State-of-the-art LLMs' are evaluated for QA, where answers are absent from the Wikipedia page in the prompting languages (Mandarin Chinese and German)

- Models: Llama3.1-8B/70B, Qwen2.5-7B/72B and gpt-oss-20B/120B

Experiments: set-up

State-of-the-art LLMs' are evaluated for QA, where answers are absent from the Wikipedia page in the prompting languages (Mandarin Chinese and German)

- Models: Llama3.1-8B/70B, Qwen2.5-7B/72B and gpt-oss-20B/120B
- Evaluation metrics:
  - Standard Natural Language Generation (NLG) evaluation metrics measuring the similarity between the predicted and reference answers at the **lexical, character**, and **semantic** levels: **ROUGE-L** (Lin, 2004) , **chrF++** (Popović, 2015) , and **BERTScore** (Zhang et al., 2020)

Experiments: set-up

State-of-the-art LLMs' are evaluated for QA, where answers are absent from the Wikipedia page in the prompting languages (Mandarin Chinese and German)

- Models: Llama3.1-8B/70B, Qwen2.5-7B/72B and gpt-oss-20B/120B
- Evaluation metrics:
  - Standard Natural Language Generation (NLG) evaluation metrics measuring the similarity between the predicted and reference answers at the **lexical, character**, and **semantic** levels: **ROUGE-L** (Lin, 2004) , **chrF++** (Popović, 2015) , and **BERTScore** (Zhang et al., 2020)
  - **LLM-as-a-judge** (LLMaJ, Chiang and Lee 2023; Zheng et al. 2023), with the judge model (gpt-oss-20b) takes the question, the reference answer, and the generated answer as input and determines whether the generated answer is correct
  - To verify the reliability of LLMaJ evaluation, 50 QA instances have been manually inspected and only one misjudgment is found

Experiments: WiLoVa-QA

Context scenarios given to the LLMs:

- (1a) *question-only*: asking the question without any additional context information
- (1b) *+standard*: providing lead section text from the Wikipedia article in the standard language as context
- (1c) *+local*: providing lead section text from the Wikipedia article in the local language

Experiments: WiLoVa-QA

Context scenarios given to the LLMs:

- (1a) *question-only*: asking the question without any additional context information
- (1b) *+standard*: providing lead section text from the Wikipedia article in the standard language as context
- (1c) *+local*: providing lead section text from the Wikipedia article in the local language
- (1d) *+standard+local*: providing lead section texts from both the standard and the local from Wikipedia articles
- (1e) *+local (translated)*: providing translated lead section translated by Google Cloud Translation
  (Only for `yue`→`cmn` due to the lack of high-quality `bar`→`deu` translators)

Experiments: WiLoVa-QA results (lead section)

Model	Context	yue-cmn					bar-deu				
		ROUGE-L	chrF++	BERTScore	BS-diff	LLMaJ	ROUGE-L	chrF++	BERTScore	BS-diff	LLMaJ
Llama3.1-70B	question-only	20.60	12.47	26.57	/	14.63	17.89	20.94	23.75	/	24.16
	+standard	20.31	13.12	21.12	-5.45	18.37	21.04	23.24	27.15	+3.40	25.28
	+local	42.20	31.48	54.23	+27.66	**85.03**	**28.77**	**27.33**	**31.13**	**+7.38**	**75.84**
	+standard+local	45.30	33.35	52.95	+26.38	82.65	28.16	26.57	30.02	+6.27	73.03
	+local (translated)	**56.47**	**41.81**	**55.86**	**+29.29**	82.99	/	/	/	/	
Qwen2.5-72B	question-only	17.05	11.35	23.17	/	16.33	18.56	23.39	26.96	/	17.42
	+standard	19.58	12.73	23.83	+0.66	20.75	22.91	28.52	29.63	+2.67	20.22
	+local	52.49	38.79	55.85	+32.68	**88.10**	35.03	36.82	40.28	+13.32	**81.46**
	+standard+local	53.54	40.42	54.94	+31.77	85.03	**37.13**	**40.90**	**44.65**	**+17.69**	76.97
	+local (translated)	**58.03**	**42.74**	**55.92**	**+32.75**	86.39	/	/	/	/	
gpt-oos-20b	question-only	17.32	12.14	24.45	/	15.65	14.42	20.86	21.89	/	22.47
	+standard	19.59	13.54	26.88	+2.43	21.09	18.01	23.85	24.98	+3.09	25.28
	+local	48.06	35.60	**50.74**	**+26.29**	**88.44**	**26.33**	**29.30**	**31.04**	**+9.15**	**70.79**
	+standard+local	47.15	35.07	48.29	+23.84	86.39	25.14	27.84	29.59	+7.70	68.54
	+local (translated)	**50.69**	**37.15**	50.56	+26.11	85.37	/	/	/	/	
gpt-oos-120b	question-only	20.91	14.41	26.24	/	20.41	17.75	23.85	25.98	/	23.03
	+standard	23.58	16.18	27.98	+1.74	32.65	20.33	26.35	28.06	+2.08	26.97
	+local	48.34	35.56	50.13	+23.89	**89.46**	**28.90**	32.41	34.81	+8.83	**73.60**
	+standard+local	48.21	35.34	49.01	+22.77	86.39	28.69	**33.44**	**35.03**	**+9.05**	73.03
	+local (translated)	**51.33**	**37.73**	**51.12**	**+24.88**	86.39	/	/	/	/	/

Table: Lead section QA performance, where the key information required to answer the questions is contained in the *+local* context. **Bold** indicates the highest score and Blue highlights the winner between *+local* and *+standard*, and all other settings higher than the winner. BS-diff measures the difference in BERTScore compared to the *question-only* baseline.

- *+local* substantially improves the QA performance
- *+local (translated)* tends to further improve scores on ROUGE-L, chrF++, and BERTScore

Experiments: WILOVA-QA results (document-level)

Model	Context	yue-cmn	bar-deu
Llama3.1-70B	question-only	24.48	22.39
	+standard	-0.49	**+8.61**
	+local	+0.42	+1.63
	+standard+local	+0.68	+5.59
	+local (translated)	**+1.62**	/
Qwen2.5-72B	question-only	25.83	25.36
	+standard	+2.00	+4.02
	+local	**+3.25**	**+7.54**
	+standard+local	+1.55	+6.64
	+local (translated)	+0.19	/
gpt-oss-20b	question-only	22.44	22.59
	+standard	-1.13	+3.95
	+local	+0.72	**+6.32**
	+standard+local	+1.24	+3.09
	+local (translated)	**+3.59**	/
gpt-oss-120b	question-only	23.23	23.38
	+standard	+3.05	+6.15
	+local	+1.72	+5.75
	+standard+local	**+3.95**	**+7.51**
	+local (translated)	+1.68	/

Table: Document-level BERTScore difference from *question-only* baseline; **bold** indicates best.

- While the crucial information can only be found in outside lead sections, providing lead section context, e.g. *+standard*, *+local*, or *+local (translated)* tends to slightly improves performance

Experiments: ECLeKTic

- ECLeKTic (Goldman et al., 2025) is a comparable QA evaluation dataset based on Wikipedia in 12 high-resource languages:
  - QA pairs about information in Wikipedia articles in only 1 language, and not in the other 11 languages
  - Questions generated using Gemini (unlike our manual QA annotations)
  - Questions and answers machine-translated into the other 11 languages for evaluation

Experiments: ECLeKTic

- ECLeKTic (Goldman et al., 2025) is a comparable QA evaluation dataset based on Wikipedia in 12 high-resource languages:
  - QA pairs about information in Wikipedia articles in only 1 language, and not in the other 11 languages
  - Questions generated using Gemini (unlike our manual QA annotations)
  - Questions and answers machine-translated into the other 11 languages for evaluation
- We evaluate on ECLeKTic to compare retrieving information from another standard language (e.g., French) versus from a local variety (i.e., Cantonese or Bavarian)

Experiments: ECLeKTic

- We assess our LLM selection on a subset of 333 ECLeKTic items of 10 languages, where the answer is absent from `cmn`/`deu` pages

Experiments: ECLeKTic

- We assess our LLM selection on a subset of 333 ECLeKTic items of 10 languages, where the answer is absent from `cmn`/`deu` pages
- Similar context scenarios given to LLMs:
  - (2a) *question-only*: without any additional context information
  - (2c) *+context*: providing context text from the source Wikipedia article in another language (French, Japanese etc.)
  - (2e) *+context (translated)*: providing the translated context text (others→`cmn`/`deu`)

Experiments: ECLeKTic

- We assess our LLM selection on a subset of 333 ECLeKTic items of 10 languages, where the answer is absent from `cmn`/`deu` pages
- Similar context scenarios given to LLMs:
  - (2a) *question-only*: without any additional context information
  - (2c) *+context*: providing context text from the source Wikipedia article in another language (French, Japanese etc.)
  - (2e) *+context (translated)*: providing the translated context text (others→`cmn`/`deu`)
- As ECLeKTic addresses Wikipedia pages absent in the target languages, it lacks counterparts to (1b) and (1d)

Experiments: ECLeKTic results

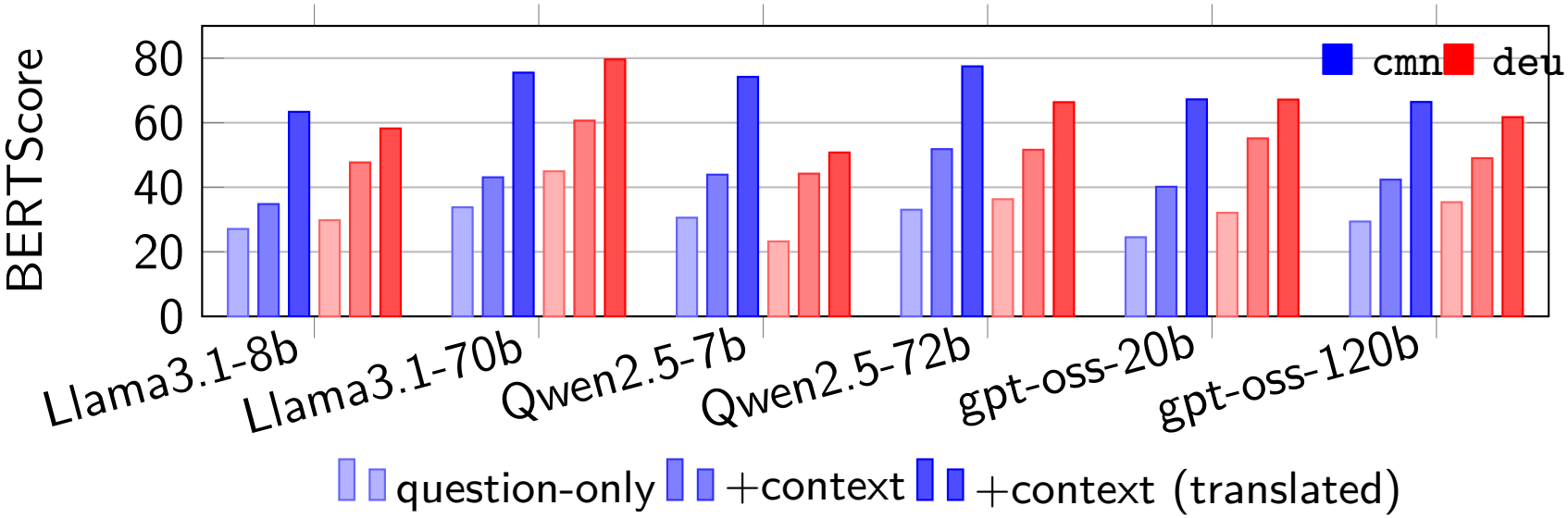

Figure: QA performance measured by BERTScore on the ECLeKTic dataset. Results compare context types of *question-only*, *+context*, and *+context (translated)* across models.

Experiments: ECLeKTic results

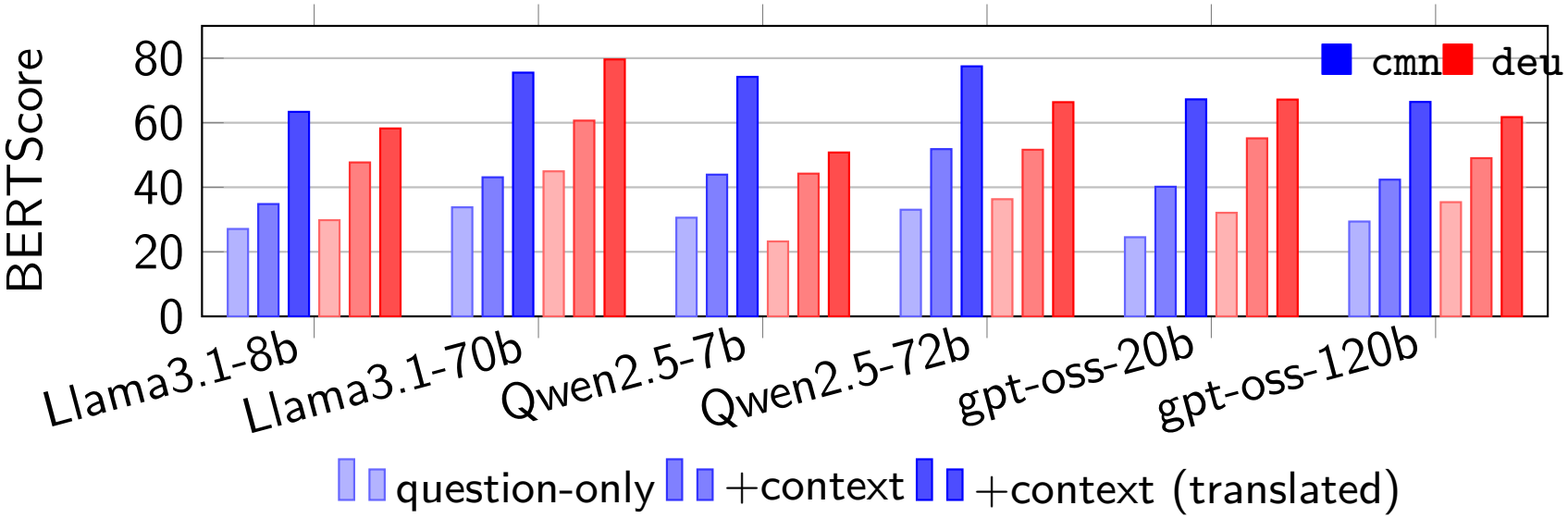

Figure: QA performance measured by BERTScore on the ECLeKTic dataset. Results compare context types of *question-only*, *+context*, and *+context (translated)* across models.

- Same trend as observed with WiLoVa-QA (lead section):
  - Performance improves noticeably when *+context* (containing the key information but in different languages)
  - Gains are even larger when the context is translated into the prompt language

Experiments: ECLeKTic results

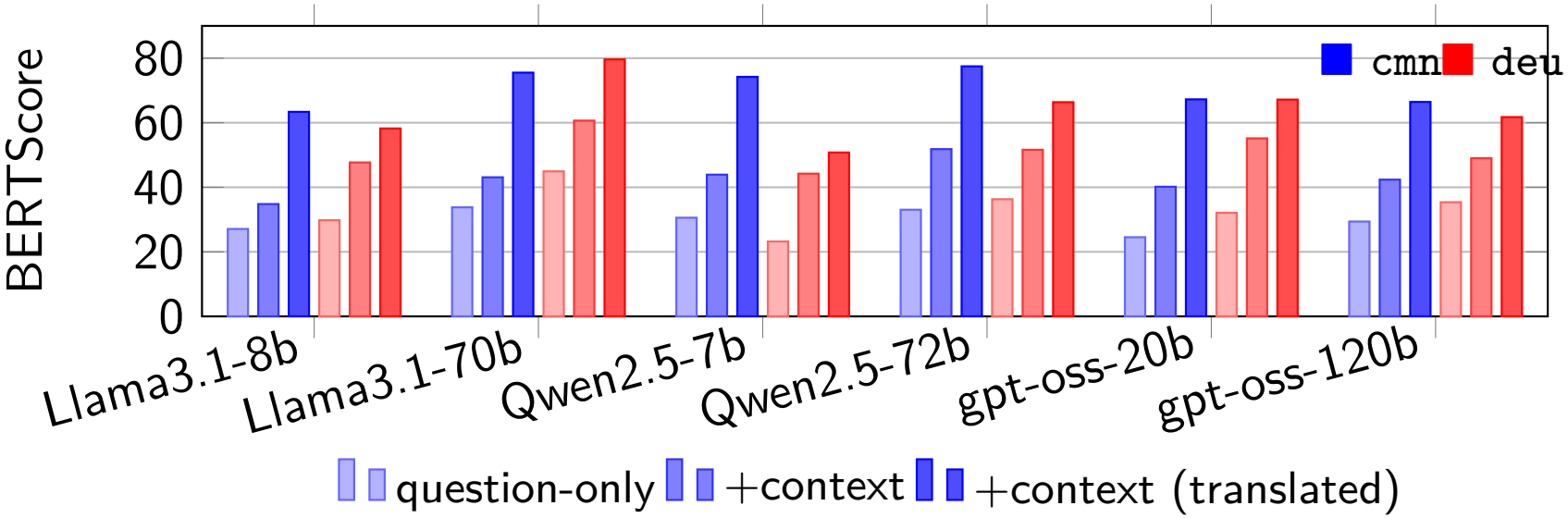

Figure: QA performance measured by BERTScore on the ECLeKTic dataset. Results compare context types of *question-only*, *+context*, and *+context (translated)* across models.

- Same trend as observed with WiLoVa-QA (lead section):
  - Performance improves noticeably when *+context* (containing the key information but in different languages)
  - Gains are even larger when the context is translated into the prompt language
- Compared with WiLoVa-QA, the performance on ECLeKTic is higher in general, indicating that the low-resource `yue` and `bar` pose greater challenges for LLMs

Experiments: stratified evaluation by topic and region

Stratified evaluation to investigate whether QA is more challenging for specific topics or regions:

Experiments: stratified evaluation by topic and region

Stratified evaluation to investigate whether QA is more challenging for specific topics or regions:

- 14 topic categories assigned to articles: *history*, *sport*, ***geography***, ***entertainment-art***, *animals-plants*, *politics-government*, *food*, *transportation*, *science-math-technology*, *linguistics*, *culture-customs*, *education*, *business*, and *other* (2 most frequent topic categories in bold)

Experiments: stratified evaluation by topic and region

Stratified evaluation to investigate whether QA is more challenging for specific topics or regions:

- 14 topic categories assigned to articles: *history*, *sport*, ***geography***, ***entertainment-art***, *animals-plants*, *politics-government*, *food*, *transportation*, *science-math-technology*, *linguistics*, *culture-customs*, *education*, *business*, and *other* (2 most frequent topic categories in bold)
- Each article also annotated for region relevance:
  - *local region*: discuss concepts and cultures of Cantonese- and Bavarian-speaking regions
  - *standard region*: the information is general to the *standard* language region (China, Germany etc.)
  - *others*: pertinent to other specific regions (e.g. Japan)

Experiments: stratified evaluation by topic and region

Stratified evaluation to investigate whether QA is more challenging for specific topics or regions:

- 14 topic categories assigned to articles: *history*, *sport*, ***geography***, ***entertainment-art***, *animals-plants*, *politics-government*, *food*, *transportation*, *science-math-technology*, *linguistics*, *culture-customs*, *education*, *business*, and *other* (2 most frequent topic categories in bold)
- Each article also annotated for region relevance:
  - *local region*: discuss concepts and cultures of Cantonese- and Bavarian-speaking regions
  - *standard region*: the information is general to the *standard* language region (China, Germany etc.)
  - *others*: pertinent to other specific regions (e.g. Japan)
- In both `yue-cmn` and `bar-deu` samples, less than half are *local region* articles: 49 (35.8%) and 35 (37.2%)

Stratified evaluation by topic and region: results

Topic/Region	Context	yue-cmn	bar-deu
geography	question-only	25.16	28.17
	+standard	22.99	31.08
	+local	**57.02**	37.23
	+standard+local	56.34	**43.67**
entertainment-art	question-only	24.00	22.79
	+standard	28.26	27.59
	+local	56.46	40.42
	+standard+local	**57.18**	**44.07**
local region	question-only	26.04	26.86
	+standard	26.26	31.65
	+local	**59.55**	43.46
	+standard+local	58.57	**49.19**
all articles (avg)	question-only	23.17	26.96
	+standard	23.83	29.63
	+local	**55.85**	40.28
	+standard+local	54.94	**44.65**

Table: Lead section BERTScore results of Qwen2.5-72B (overall best performing model) on top 2 topics and *local* articles.

Stratified evaluation by topic and region: results

Topic/Region	Context	yue-cmn	bar-deu
geography	question-only	25.16	28.17
	+standard	22.99	31.08
	+local	**57.02**	37.23
	+standard+local	56.34	**43.67**
entertainment-art	question-only	24.00	22.79
	+standard	28.26	27.59
	+local	56.46	40.42
	+standard+local	**57.18**	**44.07**
local region	question-only	26.04	26.86
	+standard	26.26	31.65
	+local	**59.55**	43.46
	+standard+local	58.57	**49.19**
all articles (avg)	question-only	23.17	26.96
	+standard	23.83	29.63
	+local	**55.85**	40.28
	+standard+local	54.94	**44.65**

Table: Lead section BERTScore results of Qwen2.5-72B (overall best performing model) on top 2 topics and *local* articles.

- Results by topic and region match the dataset's average

Stratified evaluation by topic and region: results

Topic/Region	Context	yue-cmn	bar-deu
geography	question-only	25.16	28.17
	+standard	22.99	31.08
	+local	**57.02**	37.23
	+standard+local	56.34	**43.67**
entertainment-art	question-only	24.00	22.79
	+standard	28.26	27.59
	+local	56.46	40.42
	+standard+local	**57.18**	**44.07**
local region	question-only	26.04	26.86
	+standard	26.26	31.65
	+local	**59.55**	43.46
	+standard+local	58.57	**49.19**
all articles (avg)	question-only	23.17	26.96
	+standard	23.83	29.63
	+local	**55.85**	40.28
	+standard+local	54.94	**44.65**

Table: Lead section BERTScore results of Qwen2.5-72B (overall best performing model) on top 2 topics and *local* articles.

- Results by topic and region match the dataset's average
- Better performance for *local region*, indicating that *local region*-relevant articles are **not** more challenging for LLMs than other articles

Outline

Conclusion

- We introduce WiLoVa-QA, a Wikipedia-based QA dataset that addresses information asymmetry across *standard* versus *local* language varieties

Conclusion

- We introduce WiLoVa-QA, a Wikipedia-based QA dataset that addresses information asymmetry across *standard* versus *local* language varieties
- LLMs consistently fail to answer questions derived from local Wikipedia pages in closed-book QA, suggesting that such knowledge is systematically underrepresented in models

Conclusion

- We introduce WiLoVa-QA, a Wikipedia-based QA dataset that addresses information asymmetry across *standard* versus *local* language varieties
- LLMs consistently fail to answer questions derived from local Wikipedia pages in closed-book QA, suggesting that such knowledge is systematically underrepresented in models
- Providing the relevant local Wikipedia context can substantially improve the performance, indicating that LLMs can utilize that knowledge when it is explicitly supplied, even in a low-resource language variants

Conclusion

- We introduce WiLoVa-QA, a Wikipedia-based QA dataset that addresses information asymmetry across *standard* versus *local* language varieties
- LLMs consistently fail to answer questions derived from local Wikipedia pages in closed-book QA, suggesting that such knowledge is systematically underrepresented in models
- Providing the relevant local Wikipedia context can substantially improve the performance, indicating that LLMs can utilize that knowledge when it is explicitly supplied, even in a low-resource language variants
- Translating into the standard variant further boosts performance, highlighting the importance of context language for knowledge integration

Thank you!

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. The ethnologue 200: What are the top 200 most spoken languages? Ethnologue Insights. Online; accessed 2025-09-23.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.