UPPSALA
UNIVERSITET

Cultural Grounding in Swedish: Extending an Everyday Knowledge Benchmark for LLMs

Meriem Beloucif & **Johan Sjons**

Uppsala University

May 11, 2026

Context

- Our Swedish extension was developed as part of the SemEval 2026 shared task on culturally grounded everyday knowledge (Ousidhoum et al., 2026) (forthcoming, but it's on arxiv)

Context

- Our Swedish extension was developed as part of the SemEval 2026 shared task on culturally grounded everyday knowledge (Ousidhoum et al., 2026) (forthcoming, but it's on arxiv)
- The shared task expands the original **B**enchmark for **L**LMs on **E**veryday k**n**owledge in **D**iverse cultures and languages (BLEnD, Myung et al., 2024) from 13 to more than 30 language-culture pairs.

Introduction: a problem in principle?

- LLMs “know” a lot

Introduction: a problem in principle?

- LLMs "know" a lot
- There has been improvement, no doubt (a.k.a. more human-like output)

Introduction: a problem in principle?

- LLMs "know" a lot
- There has been improvement, no doubt (a.k.a. more human-like output)
- However, there is an outside world (perhaps) of some sorts, to which LLMs have no access

Introduction: a problem in principle?

- LLMs "know" a lot
- There has been improvement, no doubt (a.k.a. more human-like output)
- However, there is an outside world (perhaps) of some sorts, to which LLMs have no access
- Rather, LLMs rely training data that comes from a cognitive machinery (us) that we still do not understand

Introduction: a problem in principle?

- LLMs "know" a lot
- There has been improvement, no doubt (a.k.a. more human-like output)
- However, there is an outside world (perhaps) of some sorts, to which LLMs have no access
- Rather, LLMs rely training data that comes from a cognitive machinery (us) that we still do not understand
- (What about the bitter lesson? Cf. Sutton, 2019)

Introduction: a problem on top of this???

- LLMs are typically skewed towards WEIRD characteristics (Henrich et al., 2010)
- Even strong multilingual models may miss local cultural knowledge (due to what the training data is)

Current study

- We present a Swedish extension of an existing benchmark for evaluating everyday knowledge in LLMs.
- Even though Sweden itself is WEIRD, model outputs may still fail to match local norms and expectations.

Related Work

- Myung et al. (2024) introduced BLEnD, a carefully constructed benchmark for culturally grounded everyday knowledge.

Related Work

- Myung et al. (2024) introduced BLEnD, a carefully constructed benchmark for culturally grounded everyday knowledge.
  - Comprises 500 question-answer pairs, covering topics such as food, sports, holidays, work etc.
  - Spans 16 countries/regions and 13 languages.
- Myung et al. (2024) evaluated on several state-of-the-art LLMs, including GPT-4 (best-performing)
- Results indicate inverse relation between performance and language online representation

Related Work

- Myung et al. (2024) introduced BLEnD, a carefully constructed benchmark for culturally grounded everyday knowledge.
  - Comprises 500 question-answer pairs, covering topics such as food, sports, holidays, work etc.
  - Spans 16 countries/regions and 13 languages.
- Myung et al. (2024) evaluated on several state-of-the-art LLMs, including GPT-4 (best-performing)
- Results indicate inverse relation between performance and language online representation
- Put differently: LLMs performed substantially worse on underrepresented cultures and languages.

Related Work

- Myung et al. (2024) introduced BLEnD, a carefully constructed benchmark for culturally grounded everyday knowledge.
  - Comprises 500 question-answer pairs, covering topics such as food, sports, holidays, work etc.
  - Spans 16 countries/regions and 13 languages.
- Myung et al. (2024) evaluated on several state-of-the-art LLMs, including GPT-4 (best-performing)
- Results indicate inverse relation between performance and language online representation
- Put differently: LLMs performed substantially worse on underrepresented cultures and languages.
  - GPT-4 showed a performance gap of up to 57%.

Swedish resources are missing?

- Benchmarks for Swedish do exist, for example for:
  - general language understanding (e.g., SuperLim Berdicevskis et al., 2023)

Swedish resources are missing?

- Benchmarks for Swedish do exist, for example for:
  - general language understanding (e.g., SuperLim Berdicevskis et al., 2023)
  - target domains such as medicine (Hertzberg and Lokrantz, 2024)

Swedish resources are missing?

- Benchmarks for Swedish do exist, for example for:
  - general language understanding (e.g., SuperLim Berdicevskis et al., 2023)
  - target domains such as medicine (Hertzberg and Lokrantz, 2024)
  - Swedish syntax (Lundqvist, 2025; Sjons et al., 2026).

Swedish resources are missing?

- Benchmarks for Swedish do exist, for example for:
  - general language understanding (e.g., SuperLim Berdicevskis et al., 2023)
  - target domains such as medicine (Hertzberg and Lokrantz, 2024)
  - Swedish syntax (Lundqvist, 2025; Sjons et al., 2026).
- However, culturally grounded everyday knowledge remains more or less unexplored in Swedish.

Overall RQ

- Generally: Do LLMs capture everyday Swedish cultural knowledge?
- More specific RQ: How do (L)LMs perform on Swedish BLEnD?

Dataset Construction

- We followed the BLEnD methodology for data aggregation and analysis... Hence:

Dataset Construction

- We followed the BLEnD methodology for data aggregation and analysis... Hence:
- We took the original 500 BLEnD questions and translated them into Swedish using ChatGPT

Dataset Construction

- We followed the BLEnD methodology for data aggregation and analysis... Hence:
- We took the original 500 BLEnD questions and translated them into Swedish using ChatGPT
- A native Swedish speaker manually reviewed and corrected the translations
  - For instance, giving examples specific to Swedish culture (e.g., amusement park names)

Dataset Construction

- We followed the BLEnD methodology for data aggregation and analysis... Hence:
- We took the original 500 BLEnD questions and translated them into Swedish using ChatGPT
- A native Swedish speaker manually reviewed and corrected the translations
  - For instance, giving examples specific to Swedish culture (e.g., amusement park names)
- Some translations sounded anglicized or culturally unnatural

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2
- Participants were instructed to:
  - Provide up to three short answers when multiple alternatives were relevant
  - Put “I don’t know” (or similar) or “Not relevant”, where appropriate
  - Remember that their estimated answers were the truth
  - Not to use online sources to find answers

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2
- Participants were instructed to:
  - Provide up to three short answers when multiple alternatives were relevant
  - Put "I don't know" (or similar) or "Not relevant", where appropriate
  - Remember that their estimated answers were the truth
  - Not to use online sources to find answers
- We manually cleaned and aggregated the responses

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2
- Participants were instructed to:
  - Provide up to three short answers when multiple alternatives were relevant
  - Put "I don't know" (or similar) or "Not relevant", where appropriate
  - Remember that their estimated answers were the truth
  - Not to use online sources to find answers
- We manually cleaned and aggregated the responses
  - For example: Lexical variants such as "go to bed" and "sleep" were merged

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2
- Participants were instructed to:
  - Provide up to three short answers when multiple alternatives were relevant
  - Put "I don't know" (or similar) or "Not relevant", where appropriate
  - Remember that their estimated answers were the truth
  - Not to use online sources to find answers
- We manually cleaned and aggregated the responses
  - For example: Lexical variants such as "go to bed" and "sleep" were merged
- Responses such as "I don't know" were excluded from the final aggregation

Response Collection

- Five Swedish participants were recruited and paid to independently answer the Swedish questions (no interaction)
  - All Stockholm-based, native in Swedish
  - Age span 15–40
  - Males: 3, Females: 2
- Participants were instructed to:
  - Provide up to three short answers when multiple alternatives were relevant
  - Put "I don't know" (or similar) or "Not relevant", where appropriate
  - Remember that their estimated answers were the truth
  - Not to use online sources to find answers
- We manually cleaned and aggregated the responses
  - For example: Lexical variants such as "go to bed" and "sleep" were merged
- Responses such as "I don't know" were excluded from the final aggregation
- Final answers included Swedish responses, English translations, and vote counts

Method

- Models were prompted to answer the questions

Method

- Models were prompted to answer the questions
- Prompt: “Svara med ett eller två ord på svenska. Endast ord, ingen förklaring, ingen punkt” (*Answer with one or two words in Swedish, no explanation, no full stop*)

Models

- We evaluated multilingual, Swedish, and instruction-tuned LLMs:
  - **GPT-4o-mini**
  - **GPT-SW3-1.3B-Instruct** (Ekgren et al., 2022)
  - **Mistral7B** (Jiang et al., 2023)

Results

Model	Accuracy
gpt-4o	**50.80%**
gpt-sw3-1.3b-instruct	15.80%
Mistral7B	0.80%

Table: Strict exact-match accuracy against Swedish BLEnD gold answers.

Analysis

- Errors were usually not nonsense but often culturally off
  - "Where do mothers stay for a certain period after childbirth for recovery in Sweden?"
    - **Humans:** "BB" ("barnbördsavdelning"; *maternity ward*)
    - **GPT-SW3-1.3B-Instruct:** "I hemmet" (*At home*)

Exact matching? Strengths? Weaknesses?

- Simple, reproducible, and easy to interpret
  - Useful as first baseline
- However:
- Different words may express the same idea
- Small wording differences become full errors
- Measures overlap – not necessarily understanding

Takeaway

- Surface-level fluency is not the same thing as cultural grounding

Limitations and future work – specific

- Sweden is not one culture, and five participants cannot capture the whole country
- Model performance may also not be representative...
- ...however, it probably is, given previous results (Myung et al., 2024)

Limitations and future work – more generally

- Participants were told not to use Google or external sources (after all, their answers constitute gold answers)
- However, many everyday questions do not have one clear ground truth
- Example:
  - "Which team is the most popular in Sweden?"
  - According to which metric? (tv-viewers, most sold merch, social media?)
- A model may fail because it lacks cultural knowledge...
- ...or because it interprets the question differently from the participants
- Put differently: what exactly should a culturally competent model predict?

Limitations and future work – more generally

- Participants were told not to use Google or external sources (after all, their answers constitute gold answers)
- However, many everyday questions do not have one clear ground truth
- Example:
  - "Which team is the most popular in Sweden?"
  - According to which metric? (tv-viewers, most sold merch, social media?)
- A model may fail because it lacks cultural knowledge...
- ...or because it interprets the question differently from the participants
- Put differently: what exactly should a culturally competent model predict?
- Still, we have to start somewhere

Conclusion

- Cultural evaluation needs locally grounded human data

Last point (general!)

- What is models of language a model *of*?

References I

Berdicevskis, A., Bouma, G., Kurtz, R., Morger, F., Öhman, J., Adesam, Y., Borin, L., Dannélls, D., Forsberg, M., Isbister, T., Lindahl, A., Malmsten, M., Rekathati, F., Sahlgren, M., Volodina, E., Börjeson, L., Hengchen, S., and Tahmasebi, N. (2023). Superlim: A Swedish language understanding evaluation benchmark. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Ekgren, A., Cuba Gyllensten, A., Gogoulou, E., Heiman, A., Verlinden, S., Öhman, J., Carlsson, F., and Sahlgren, M. (2022). Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language*

References II

Resources and Evaluation Conference, pages 3509–3518, Marseille, France. European Language Resources Association.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Hertzberg, N. and Lokrantz, A. (2024). MedQA-SWE - a clinical question & answer dataset for Swedish. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186, Torino, Italia. ELRA and ICCL.

References III

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Lundqvist, S. (2025). Do large language models and humans follow similar learning stages?: Assessing GPT-2's order of Swedish grammar acquisition within the Processability Theory framework. Master's thesis, Uppsala University.

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., Ousidhoum, N., Camacho-Collados, J., and Oh, A. (2024). Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In Globerson, A., Mackey, L.,

References IV

Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Ousidhoum, N., Myung, J., Perez-Almendros, C., Jin, J., Keleg, A., Beloucif, M., Zhou, Y., Agerri, R., Araujo, V., Baes, N., Barry, J., Boisson, J., Chen, N. F., de Kock, C., Edwards, A., Fernandez de Landa, J., Imam, M. F., Hakami, H., Hsieh, S.-K., Imperial, J. M., Lee, R. K.-W., Liu, Z., Lyu, C., Samih, Y., Sjons, J., Tan, B., Ushio, A., Zheng, W., Oh, A., and Camacho-Collados, J. (2026). Semeval-2026 task 7: Everyday knowledge across diverse languages and cultures. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. To appear.

References V

Sjons, J., Heinat, F., and Kurfali, M. (2026). The swedish benchmark of linguistic minimal pairs. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6783–6794, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Sutton, R. (2019). The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`. Accessed: 2025-09-07.

QA

Thank you! Questions?