



Universiteit
Leiden
The Netherlands

Out of the box, into the Clinic? Evaluating state-of-the-art ASR for clinical applications for older adults

Dijk, B. van; Kuiper, T.; Aoulad si Ahmed, S.; Lefebvre, A.E.J.L.; Johnson, J.; Duin, J.; ... ; Yang, D.

Citation

Dijk, B. van, Kuiper, T., Aoulad si Ahmed, S., Lefebvre, A. E. J. L., Johnson, J., Duin, J., ... Spruit, M. R. (2025). Out of the box, into the Clinic?: Evaluating state-of-the-art ASR for clinical applications for older adults. *Proceedings Of The Fourth Workshop On Bridging Human-Computer Interaction And Natural Language Processing (Hci+Nlp)*, 72-78.
doi:10.18653/v1/2025.hcinlp-1.7

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](#)

Downloaded from: <https://hdl.handle.net/1887/4306648>

Note: To cite this publication please use the final published version (if applicable).

Out of the Box, into the Clinic? Evaluating State-of-the-Art ASR for Clinical Applications for Older Adults

**Bram van Dijk¹, Tiberon Kuiper³, Sirin Aoulad si Ahmed³, Armel Lefebvre¹,
Jake Johnson³, Jan Duin³, Simon Mooijaart^{3,4}, Marco Spruit^{1,2}**

¹Department of Public Health and Primary Care, Leiden University Medical Center

²Leiden Institute of Advanced Computer Science, Leiden University

³Center for Medicine for Older People, Leiden University Medical Center

⁴Department of Internal Medicine, Leiden University Medical Center

Correspondence: b.m.a.van_dijk@lumc.nl

Abstract

Voice-controlled interfaces can support older adults in clinical contexts – with chatbots being a prime example – but reliable Automatic Speech Recognition (ASR) for underrepresented groups remains a bottleneck. This study evaluates state-of-the-art ASR models on language use of older Dutch adults, who interacted with the Welzijn.AI chatbot designed for geriatric contexts. We benchmark generic multilingual ASR models, and models fine-tuned for Dutch spoken by older adults, while also considering processing speed. Our results show that generic multilingual models outperform fine-tuned models, which suggests recent ASR models can generalise well out of the box to real-world datasets. Moreover, our results indicate that truncating generic models is helpful in balancing the accuracy-speed trade-off. Nonetheless, we also find inputs which cause a high word error rate and place them in context.

1 Introduction

Although there is a surge of interest in AI-driven applications like chatbots in the health domain (Guo et al., 2024; Huo et al., 2025), tailoring them to groups underrepresented in AI research remains a challenge. Older populations are one example: because they often have different needs and preferences when interacting with AI (van Dijk et al., 2025; Klaassen et al., 2025), their involvement in system development is key for building clinically relevant systems in geriatrics, the field in healthcare concerned with the health of older adults. This population is increasing in size globally (World Health Organisation, 2023), while personnel shortages in healthcare become pressing (Eurofound, 2023); yet addressing these challenges with AI warrants systems that align well with older adults.

Voice control is a key element in chatbots in geriatrics, as older individuals may struggle with small fonts, icons, and typing text in standard interfaces

(Khamaj, 2025). Yet, implementing voice control is not obvious, as the performance of Automatic Speech Recognition (ASR) systems depends on the representation of older adults in training data, and on their articulation, speech volume and technological literacy (Klaassen et al., 2025). Moreover, evaluation of state-of-the-art ASR systems on realistic data of older adults is lagging behind.

In this short paper, we evaluate recent ASR models on older adults’ language use in interaction with Welzijn.AI. This is a new digital platform for older users, that currently features a prototype chatbot to converse about clinically relevant topics like quality of life and frailty. Audio data of 10 older Dutch adults interacting with Welzijn.AI were transcribed using generic multilingual ASR models (Whisper and Voxtral), and models fine-tuned specifically for Dutch or older Dutch populations (Whisper and Wav2Vec2). We find that i) generic multilingual models outperform fine-tuned models, and also that ii) truncating larger generic models helps striking a good balance between accuracy and speed. These findings also hold for a similar subset of the Mozilla Common Voice dataset (Ardila et al., 2020) we use as benchmark.

2 Background

Though work on evaluating ASR models for older adults exists, extrapolating findings to realistic contexts is hard. Performance of ASR models sometimes remains implicit in downstream use, for example through evaluation of overlap of linguistic features extracted from model vs. human-generated transcripts (Naffah et al., 2025), or by using only part of ASR models (like the audio encoder) in predicting cognitive impairment in older persons (Agbavor and Liang, 2024). Other work tailors ASR systems to older adults by drawing on additional databases of individual users (Xu et al., 2025), which is not always feasible in a clinical

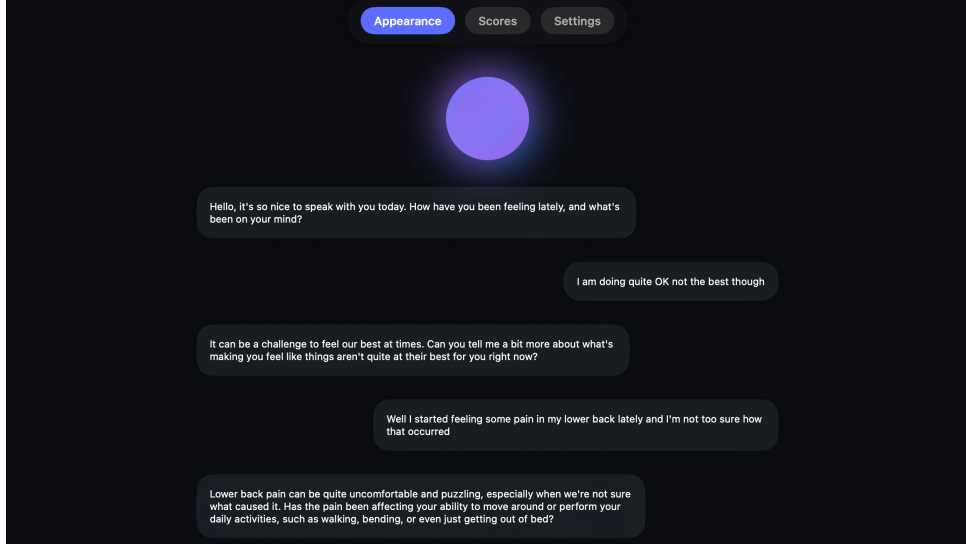


Figure 1: Interface of Welzijn.AI with an example conversation. Users press the purple button to activate the ASR functionality and start responding, after which their speech is transcribed and rendered on the screen. Chatbot responses are read out with a text-to-speech model. The ‘Scores’ button shows information extracted on quality of life and frailty, ‘Settings’ allows choosing different ASR models, and ‘Appearance’ returns the user to the conversation on display. We focus in this paper on the conversation resulting from interaction with this prototype.

context due to privacy concerns. However, [Xu et al. \(2025\)](#) also show that fine-tuning generic multilingual models with speech from older adults increases ASR performance. The work by [Shekoufandeh et al. \(2025\)](#) explores this further, by fine-tuning Whisper as recent ASR model ([Radford et al., 2023](#)) on the Dutch JASMIN-CGN dataset. This dataset includes language use of older adults in human-machine interaction settings ([Cucchiaroni and Van hamme, 2012](#)), which is potentially relevant to Welzijn.AI.

3 Materials and methods

Ten older Dutch adults (≥ 65 years) from a volunteer panel of the outpatient clinic for geriatrics at the Leiden University Medical Center (LUMC) were included. The Institutional Review Board of the LUMC approved the study and all participants provided informed consent. Participants were instructed not to share personal health information with Welzijn.AI, but rather to impersonate a peer. Participants interacted individually with the chatbot, with one experimenter standing by.

An impression of Welzijn.AI is given in [Figure 1](#). Here we focus on the system’s ability to support conversations; we refer for architectural details to [van Dijk et al. \(2025\)](#). The chatbot was driven by the meta-llama/Llama-3.3-70B-Instruct¹

model ([Grattafiori et al., 2024](#)), prompted to structure the conversation around the EQ-5D and Clinical Frailty Scale ([Brooks, 1996](#); [Rockwood et al., 2005](#)). These are validated geriatric instruments to assess quality of life and frailty, by retrieving information about mobility, mental wellbeing, physical independence, and so on. These instruments can be presented via surveys, in conversations with clinicians, or in our case, by a chatbot ([Figure 1](#)).

The chatbot was used on a laptop. Interactions took 5-10 minutes and were recorded with a handheld device. The default ASR model in Welzijn.AI was openai/whisper-large-v3, which in early testing we found to work best.

Type	Example
<i>Orthographic</i>	Hi, I am uh... feeling great today.
<i>Orthographic_clean</i>	Hi I am feeling great today
<i>Normalised</i>	hi i am feeling great today

Table 1: Examples of gold transcript types.

Hugging Face ID	Params
mistralai/Voxtral-Mini-3B-2507	4.68B
openai/whisper-large-v2	1.55B
openai/whisper-large-v3	1.55B
golesheed/whisper-native-elderly-9-dutch	1.54B
golesheed/wav2vec2-xls-r-1b-dutch-3	963M
openai/whisper-large-v3-turbo	809M
openai/whisper-medium	769M
openai/whisper-small	244M

Table 2: Models used for our ASR experiments.

¹We will use Hugging Face IDs to denote AI models.

Hugging Face ID	WER Welzijn.AI				WER Common Voice			
	Orth.	Orth_c.	Norm.	Time	Orth.	Orth_c.	Norm.	Time
mistralai/Voxtral-Mini-3B-2507	.17	.11	.09	3.75	.05	.04	.04	3.72
openai/whisper-large-v2	.19	.12	.10	4.69	.05	.04	.04	4.24
openai/whisper-large-v3	.12	.07	.06	3.41	.04	.03	.03	3.71
golesheed/whisper-native-elderly-9-dutch	.40	.22	.14	4.59	.29	.18	.07	4.07
golesheed/wav2vec2-xls-r-1b-dutch-3	.49	.37	.36	.99	.30	.19	.19	.89
openai/whisper-large-v3-turbo	.16	.10	.08	1.43	.06	.04	.04	1.47
openai/whisper-medium	.19	.13	.11	2.35	.07	.06	.06	2.48
openai/whisper-small	.26	.18	.17	1.11	.12	.10	.10	1.19

Table 3: Word Error Rate (WER) is the edit distance between prediction and reference (sum of substitutions, deletions, and insertions), divided by the length of the reference, so here denotes the average number of errors per reference word, for Orthographic (*Orth.*), Orthographic_clean (*Orth_c.*), and Normalised (*Norm.*) gold transcripts. Processing time (*Time*) in average seconds per input. Best results in bold.

Recorded user speech was after the interactions separated from chatbot responses and segmented using PyDub² and pyannote (Bredin, 2023). We obtained 199 segments with an average length of 3.4 seconds, due to the turn-taking nature of the conversation. Our sample totalled 11 min. and 15 sec., so is small, but still valuable as data from clinical contexts is challenging to obtain. Besides chatbot data, we also drew 200 random samples from the Common Voice dataset (Ardila et al., 2020) of Dutch older individuals (≥ 60 years), which totals 17 min., and 37 sec. As this data concerns written text read out loud, it should intuitively be an easier benchmark for ASR models.

For obtaining gold standard (i.e. human) reference transcriptions, segments were transcribed with the default openai/whisper-large-v3 model, and subsequently corrected by the first author. Since the Word Error Rate (WER) as standard metric in the ASR field is sensitive to fillers, capitalisation, and punctuation, and since choosing the ‘right’ reference depends on the use case, we created three types of human and model transcriptions (via rule-based postprocessing) as visible in Table 1: *orthographic* (including fillers, capitalisation, and punctuation) *orthographic_clean* (only capitalisation, as also used in Figure 1), and *normalised* (no fillers, capitalisation or punctuation). Orthographic transcription is useful in that it provides additional structure, though speech content is often arguably sufficiently preserved in normalised transcriptions, with *orthographic_clean* transcription striking a balance between the strictest and most flexible evaluation scenarios.

Table 2 shows the ASR models included in our experiments. As can be seen, we focus

on models from the Whisper model family as the current standard in the field, and include mistralai/Voxtral-Mini-3B-2507 (Liu et al., 2025) as new potential competitor. We included golesheed/whisper-native-elderly-9-dutch as the Whisper model fine-tuned for older Dutch adults by Shekoufandeh et al. (2025), and we also included golesheed/wav2vec2-xls-r-1b-dutch-3, a work-in-progress Wav2Vec2 model (Baevski et al., 2020) fine-tuned for general Dutch, as older but potentially fast contender. We note that these fine-tuned models do not output fillers, capitals or punctuation by default, hence evaluate them with normalised transcripts.

Our motivation for this set of ASR models is that they are all small (compared to Voxtral’s 24B variant for example) and open weights, hence suitable for local/private downstream applications. We take processing time into account since a good balance between accuracy and speed is key in many applications, so include models of different sizes.

Our code for the ASR pipeline is available.³ All experiments were carried out on a Macbook Pro M1 16GB using the Hugging Face ecosystem for ASR and PyTorch’s MPS GPU acceleration backend. Due to privacy restrictions we cannot share recordings nor transcripts of the interactions.

4 Results

WER and processing time per model are given in Table 3. Regarding WER, on both the Welzijn.AI and Common Voice datasets, openai/whisper-large-v3 as generic multilingual model outperforms all other models, also regarding normalised transcripts as

²<https://github.com/jiaaro/pydub>

³https://github.com/bma-vandijk/asr_pipelines

#	Model	Prediction	Reference	WER
1	openai/whisper-large-v2	hartelijk bedankt voor het kijken en tot de volgende keer <i>thank you cordially for watching and until next time</i>	gaat wel <i>it's okay</i>	5
2	golesheed/whisper-native-elderly-9-dutch	poet hem in oranje <i>loot him in orange</i>	goedemiddag <i>good afternoon</i>	4
3	openai/whisper-large-v2	ik denk dat het weer zo is als het altijd is <i>I think that it is as it always is</i>	ongeveer zoals altijd <i>roughly as usual</i>	3.33
4	openai/whisper-medium	ik ben benieuwd <i>I am curious</i>	goedemiddag <i>good afternoon</i>	3
5	openai/whisper-small	voor de wereld <i>for the world</i>	goedemiddag <i>good afternoon</i>	3
6	mistralai/Voxtral-Mini-3B-2507	groen de medaillon <i>green the medallion</i>	goedemiddag <i>good afternoon</i>	3
7	golesheed/wav2vec2-xls-r-1b-dutch-3	goede midda <i>good midda</i>	goedemiddag <i>good midday</i>	2
8	openai/whisper-large-v3-turbo	ja het is heel erg goed <i>yes it is very well</i>	weer te ingewikkeld <i>again too complicated</i>	2
9	openai/whisper-medium	bedankt voor het kijken <i>thank you for watching</i>	gaat wel <i>it's okay</i>	2
10	mistralai/Voxtral-Mini-3B-2507	het zelf gaat het <i>it self goes it</i>	hetzelfde uiteraard <i>the same of course</i>	2

Table 4: Example inputs with high WER at the sample level. English (literal) translations in italics.

most flexible evaluation scenario. In terms of processing time, on both datasets, the Wav2Vec2 model fine-tuned on Dutch language use (not specifically older language users) (golesheed/wav2vec2-xls-r-1b-dutch-3) is the fastest, but also the least accurate.

For chatbot systems like Welzijn.AI, understanding the trade-off between WER and processing time is crucial, given that in chatbots other components also impose processing time, and seconds may greatly impact the perceived quality of the experience. We visualise performance w.r.t. normalised transcripts in Figure 2. Here we see that openai/whisper-large-v3-turbo, which is essentially openai/whisper-large-v3 with a truncated decoder 1/8 its size, strikes the best balance. The ‘nearest’ improvement in WER concerns openai/whisper-large-v3, which is about three times slower, while the ‘nearest’ improvement in processing time comes from openai/whisper-small, at the expense of more than twice its WER. Figure 2 also shows that for all models, the Common Voice dataset is easier regarding WER, though not all models process these data faster. For Common Voice data the same observations hold regarding models that are the ‘nearest’ improvements in WER and processing times compared to openai/whisper-large-v3-turbo: they lead to considerable drops in speed or accuracy respectively. Our findings align with earlier work that shows that overall, larger ASR models perform better on commonly used datasets compared to smaller models (Atwany et al., 2025), though they also

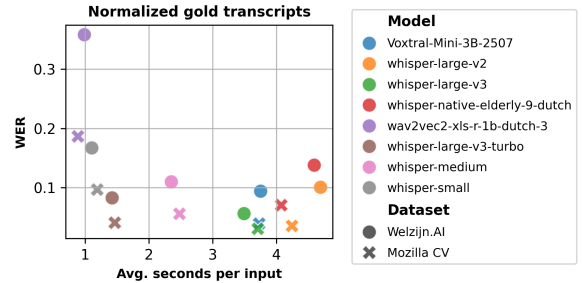


Figure 2: Overview of accuracy vs. processing time.

note that larger numbers of parameters eventually yield diminishing returns. This finding helps understand why openai/whisper-large-v3-turbo shows a relatively small performance drop while still being about 50% smaller than openai/whisper-large-v3.

4.1 Error analysis

To disclose common pitfalls in using ASR models on Dutch, we provide ten examples of the 50 predictions with highest WER (≥ 2) in Table 4, which as expected come only from Welzijn.AI data. Errors in predictions in deep neural ASR models like Whisper can be categorised in various ways. *Mishearings* could be induced by ambiguous or unclear phonemes in the input, or confusion of phonemes by the model (e.g. ‘this guy’ vs. ‘the sky’). *Hallucinations* are errors where the prediction has no semantic or phonetic relation to the reference. *Looping* is when a model keeps repeating previously recognized speech (‘Welcome to Amsterdam to Amsterdam to Amsterdam’). Furthermore, it is also known that deep neural ASR

models are sensitive to non-speech audio signals in the background caused by e.g. objects or animals (Barański et al., 2025), which is not obvious to trace in the predictions, but in live interaction settings like Welzijn.AI something to take into account.

Examples 2, 3, 6, 7 and 10 seem cases of mishearings, where 2 and 6 have at least some phonetic (but not semantic) alignment with the reference; examples 3, 7 and 10 also have some semantic overlap. Examples 1, 4, 5, 8, 9 have no clear phonetic or semantic link with the reference thus qualify as hallucinations; 1 and 9 are probably frequency effects from Whisper’s training data (video transcriptions) (Barański et al., 2025), meaning the models prioritize patterns in the training distribution over the actual audio input. All in all, our examples suggest high WER is not limited to just a few types of models, which aligns with earlier documented unpredictability in state-of-the-art ASR models across the board (Koenecke et al., 2024; Atwany et al., 2025), and this finding should inform their development and deployment in high-stakes contexts.

5 Discussion and conclusion

We evaluated state-of-the-art ASR models on transcribing language use of older adults interacting with the Welzijn.AI chatbot, which was designed for geriatrics. We included various generic multilingual models as well as models fine-tuned on language use of older Dutch adults and general Dutch. We found that on both Welzijn.AI and Common Voice data, generic multilingual models perform better than fine-tuned models, with openai/whisper-large-v3 as best model achieving WERs of .06 and .12 for normalised and orthographic transcriptions of realistic Welzijn.AI conversations. Interestingly, its truncated variant openai/whisper-large-v3-turbo struck the best balance between accuracy and processing speed, the latter being crucial in chatbot systems used in real-time. This is useful from a systems development perspective, since truncated models may perform well out of the box, without the need for training smaller architectures from scratch, or for additional data for group- or task-specific fine-tuning. Future work should further support this claim by using larger samples across settings.

To put our results in perspective, other evaluations of state-of-the-art ASR models

on the entire Dutch subset of the Common Voice dataset reported WERs of .06 by mistralai/Voxtral-Mini-3B-2507 (Liu et al., 2025), and .04 by openai/whisper-v3-large, assuming the strictest evaluation scenario (orthographic).⁴ We obtained similar results for our subset of Common Voice spoken by Dutch Adults: .04 for the same Whisper and .05 for the same Voxtral model. Hence, it seems that for generic multilingual models, changing the target population does not imply large performance degradation if the task is straightforward (reading text out loud).

Still, our Welzijn.AI data is conversational, hence results are harder to put in perspective. Though the WER on orthographic transcripts for Welzijn.AI data triples for the same Voxtral and Whisper models (.17 and .12 respectively), given the different nature of read speech and conversational language, this is not a dramatic loss of performance. Recent work has reported WERs for a variety of datasets transcribed with our best performing model openai/whisper-v3-large, as large as .32 for English speech recordings in home environments (BERSt), and .23 for English meeting recordings (AMI) (Atwany et al., 2025).

We also attempted to categorize WER error types. We saw that mishearings were as frequent as hallucinations. Hallucinations, however, are potentially more problematic for systems like Welzijn.AI, as for a user who is unable to make sense of the resulting transcription, trust will erode faster compared to a mishearing, which still has some semantic or phonetic similarity.

Strategies to improve WER and mitigate hallucinations, could include more independent language modelling components that take specific contexts into account. When the audio input for the decoder is noisy, its language prior generates a transcript based on its own distribution instead of the input, which may well be out-of-context. Mitigation could involve generating candidate predictions and evaluating their likelihoods from the perspective of a domain-specific language model, or combining the decoder’s predicted token probabilities directly with the prediction of a context-specific language model (see also Zhou and Li, 2025). Hence, some exciting work remains for making real-world impact with recent ASR models.

⁴Results of latest model reported on <https://github.com/openai/whisper>, August 6 2025.

6 Limitations

Though dedicated GPUs in high performing computing clusters have higher bandwidths and are faster, hence the default choice in experiments like ours, unified architectures (such as provided by Apple’s Silicon M series) are receiving more attention nowadays due to their benefits in terms of latency, energy consumption, and system footprint, and are recognized as efficient and competitive alternatives to dedicated GPUs (Hübner et al., 2025; Kenyon and Capano, 2022). So faster processing times could probably be attained by using dedicated GPUs or by using a more recent M-chip. Still, we anticipate a scenario where one device hosts multiple smaller AI models to do different tasks, for which our setup can provide a good lower bound.

Also, though we tried to make comparison fair for different ASR models by evaluating with different kinds of transcripts, developing further transcript normalisations to take ‘acceptable errors’ into account, e.g. writing numbers in digits or letters (‘8’, ‘eight’), were beyond the scope of the current work. This means that there can be some noise in our performance estimates.

References

- Felix Agbavor and Hualou Liang. 2024. Multilingual prediction of cognitive impairment with large language models and speech analysis. *Brain sciences*, 14(12):1292.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. *Common Voice: A Massively-Multilingual Speech Corpus*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Hanin Atwany, Abdul Waheed, Rita Singh, Monojit Choudhury, and Bhiksha Raj. 2025. Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models. *arXiv preprint arXiv:2502.12414*.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Mateusz Barański, Jan Jasiński, Julitta Bartolewska, Stanisław Kacprzak, Marcin Witkowski, and Konrad Kowalczyk. 2025. Investigation of whisper asr hallucinations induced by non-speech audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th INTERSPEECH Conference (INTERSPEECH 2023)*, pages 1983–1987. ISCA.
- Richard Brooks. 1996. EuroQol: the current state of play. *Health policy*, 37(1):53–72.
- Catia Cucchiaroni and Hugo Van hamme. 2012. The JASMIN speech corpus: recordings of children, non-natives and elderly people. In *Essential Speech and Language Technology for Dutch: Results by the STEVIN programme*, pages 43–59. Springer.
- Eurofound. 2023. Measures to tackle labour shortages: Lessons for future policy. <https://www.eurofound.europa.eu/en/publications/2023/measures-tackle-labour-shortages-lessons-future-policy>. (accessed 26 March 2025).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large Language Models for Mental Health Applications: Systematic Review. *JMIR mental health*, 11(1):e57400.
- Paul Hübner, Andong Hu, Ivy Peng, and Stefano Markidis. 2025. Apple vs. Oranges: Evaluating the Apple Silicon M-Series SoCs for HPC Performance and Efficiency. *arXiv preprint arXiv:2502.05317*.
- Bright Huo, Amy Boyle, Nana Marfo, Wimonchat Tangamornsuksan, Jeremy P Steen, Tyler McKechnie, Yung Lee, Julio Mayol, Stavros A Antoniou, Arun James Thirunavukarasu, and 1 others. 2025. Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, 8(2):e2457879–e2457879.
- Connor Kenyon and Collin Capano. 2022. Apple Silicon Performance in Scientific Computing. In *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–10. IEEE.
- Abdelrahman Khamaj. 2025. AI-enhanced chatbot for improving healthcare usability and accessibility for older adults. *Alexandria Engineering Journal*, 116:202–213.
- Willemijn Klaassen, Bram van Dijk, and Marco Spruit. 2025. A Review of Challenges in Speech-Based Conversational AI for Elderly Care. *Studies in health technology and informatics*, 327:858–862.

- Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681.
- Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.
- Ava Naffah, Valeria A Pfeifer, and Matthias R Mehl. 2025. Spoken Language Analysis in Aging Research: The Validity of AI-Generated Speech to Text Using OpenAI’s Whisper. *Gerontology*, 71(5):417–423.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Kenneth Rockwood, Xiaowei Song, Chris MacKnight, Howard Bergman, David B Hogan, Ian McDowell, and Arnold Mitnitski. 2005. A global clinical measure of fitness and frailty in elderly people. *CMAJ*, 173(5):489–495.
- Golshid Shekoufandeh, Paul Boersma, and Antal van den Bosch. 2025. Improving the inclusivity of dutch speech recognition by fine-tuning Whisper on the JASMIN-CGN corpus. *arXiv preprint arXiv:2502.17284*.
- Bram MA van Dijk, Armel Ejl Lefebvre, and Marco R Spruit. 2025. Welzijn. AI: Developing responsible conversational AI for the care of older people through stakeholder involvement. *Maturitas*, 199(108616).
- World Health Organisation. 2023. Mental health of older adults. <https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults>. (accessed 29 April 2025).
- Haiying Xu, Haoze Liu, Mingshi Li, Siyu Cai, Guangxuan Zheng, Yuhuang Jia, Jinghua Zhao, and Yong Qin. 2025. EchoVoices: Preserving Generational Voices and Memories for Seniors and Children. *arXiv preprint arXiv:2507.15221*.
- Shilin Zhou and Zhenghua Li. 2025. Improving Contextual ASR via Multi-grained Fusion with Large Language Models. *arXiv preprint arXiv:2507.12252*.