CREATING DOCUMENTS WITH VOICE: MAYBE IT IS NOT ABOUT TRANSCRIPTION BUT REFLECTION?

Matthias Busch[1], *Jonas Schewior*[2], *Andreas Wendemuth*[2], *Ingo Siegert*[1,3]

[1]*Mobile Dialog Systems, Otto-von-Guericke-University, Magdeburg, Germany,*
[2]*Cognitive Systems, Otto-von-Guericke-University, Magdeburg, Germany,*
[3]*Department of Psychosomatic Medicine and Psychotherapy, Otto-von-Guericke University, Magdeburg*
{matthias.busch;jonas.schewior}@ovgu.de

Abstract: Traditional dictation systems implicitly treat speech as equivalent to writing, overlooking the recursive nature of composition and penalizing cognitive pauses essential for reflection. In an exploratory study ($N = 10$), participants dictated formal and informal emails, then compared raw transcripts, manually edited versions, and Large Language Model (LLM)-transformed variants (Llama 3.1-8B/3.3-70B). No participant preferred raw output; while LLM processing helped with formal tasks, most preferred self-edited versions for authorial control. One-shot transformation proved vulnerable to Automatic Speech Recognition (ASR) error propagation and stylistic mismatches in informal contexts. These findings motivate a *thought-to-text* paradigm, reconceiving dictation as collaborative composition rather than linear transcription.

1 Introduction

The vision of natural language as a universal interface is frequently challenged by the "Myth of Naturalness" [1]. Historically, research has diverged into separate disciplines: ASR focusing on acoustic signal processing using audio recordings, and Natural Language Processing (NLP) targeting semantic content analysis using text data. Speech interfaces are designed to provide a low-effort, hands-free alternative to standard inputs, enabling new use cases and enhancing system accessibility. Dictation software represents a long-standing use case of this concept. Commercially established since the mid-1990s (e.g. *Dragon Professional* [2], *Microsoft Word*), it has found widespread adoption, particularly in medicine and assistive technologies. Nevertheless, current dictation tools treat ASR output as the final product, enforcing a paradigm where „speaking" is functionally equal to „writing". While ASR performs its intended function and accurate transcription, dictation software that relies solely on transcription ignores fundamental differences between speech and writing. As noted by Aristotle, "the style of written prose is not that of spoken oratory" (*Aristotle*, Rhetoric, 3.12, 1413b2), and as Chafe and Tannen [3] demonstrated, spoken interaction differs distinctly from written prose in both style and cognitive demand. Recent evaluations highlight persistent usability issues such as recognition errors, accent sensitivity, and correction fatigue [4], shifting cognitive effort from composing text to supervising machine output. This mismatch often turns unintended utterances and „moments of thinking aloud" into unwanted text, suppressing the reflective depth required for effective composition. Participants' hesitations were captured verbatim, often with additional ASR errors, demonstrating how transcription-focused systems penalize cognitive pauses essential to the writing process.

Background: From Transcription to Transformation

ASR systems have evolved from isolated word recognition through vocabulary-constrained systems to modern end-to-end neural architectures capable of recognizing virtually any spoken word [5, 6]. While alternative evaluation metrics have been discussed in the literature [7, 8], current state-of-the-art models are predominantly benchmarked and marketed based on Word Error Rate (WER) [9]. This dominance of WER reflects and reinforces an implicit assumption: that faithful transcription of the acoustic signal is the ultimate goal of speech-to-text systems.

In this paradigm, the concept of the "sentence" is a post-processing artifact, typically retrofitted through separate punctuation restoration models rather than emerging from acoustic modeling [10, 11]. This architectural choice ignores the systematic linguistic differences between spoken and written registers documented in Section 2, treating speech merely as unpunctuated writing.

Large language models (LLMs) introduce capabilities that fundamentally change the possibilities for voice-based document creation. Beyond next-token prediction, LLMs demonstrate instruction-following, text reformulation, style transfer, and summarization, all applicable to spoken-to-written transformation [12, 13]. This enables a shift from merely transcribing *what was said* to reconstructing *what was meant*.

Contribution

This paper presents an exploratory study ($N = 10$) examining how users interact with traditional dictation and LLM-based post-processing. Our contributions are: (1) empirical evidence that no participant preferred raw dictated output, highlighting the inadequacy of transcription-only approaches; (2) analysis of how ASR errors propagate through LLM processing; and (3) insights into user preferences for authorial control, particularly in informal communication.

2 Cognitive and Linguistic Foundations

The assumption that speaking can substitute for writing contradicts foundational linguistic and cognitive research on spoken–written differences.

Conceptual Orality vs. Conceptual Literacy

Koch and Oesterreicher [14, 15] distinguish *medium* (phonic vs. graphic) from *conception* (communicative register from *language of immediacy* to *language of distance*). These dimensions are independent: text messages use graphic medium but oral conception; academic lectures use phonic medium but literate conception.

When users speak into a microphone, dictation systems change only the *medium*. The *conceptual register* remains unchanged, spontaneous speech gravitates toward oral conception (loosely structured, emotionally involved), while written documents require literate conception (explicit, planned, informationally dense). Koch and Oesterreicher distinguish *Verschriftung* (transcription: medium change only) from *Verschriftlichung* (transformation toward literate conception) [16]. Current dictation systems perform merely *Verschriftung*.

This converges with other frameworks: Halliday [17] on lexical density vs. grammatical intricacy, Chafe [18] on integration/detachment vs. fragmentation/involvement, and Biber's [19] corpus evidence. The differences are systematic and substantial.

Table 1 – Key structural differences between spoken and written language.

Dimension	Spoken Language	Written Language
Basic unit	Intonation units, turns	Sentences, paragraphs
Syntactic structure	Clause chaining, fragmentation	Subordination, integration
Lexical density	Low	High
Planning scope	Limited ($\approx$1 second)	Extended, recursive
Revision	Real-time repair (visible)	Offline revision (invisible)
Register conception	Language of immediacy	Language of distance

The Sentence as a Written Construct

The "sentence" is primarily a written convention; spoken language organizes around *turn-constructional units* [20, 21]. Non-clausal units account for 30 to 40% of spoken language [22], and hesitation/disfluency are systematic features, not errors. Table 1 summarizes these structural differences.

Incremental Speech vs. Recursive Writing

Speech production operates under real-time constraints: speakers plan approximately one second ahead [23, 24]. Writing is fundamentally recursive, Flower and Hayes [25] characterize composition as interleaved planning, formulation, and revision. Bereiter and Scardamalia [26] distinguish *knowledge-telling* (linear transfer) from *knowledge-transforming* (active restructuring); the latter requires affordances only writing provides: permanence, revision, and reflection time.

In writing, pauses are productive planning moments [27]. Transcription-focused dictation treats pauses as disruptions, while also imposing supervisory demands that shift cognitive effort from composing to managing system behavior [4]. These foundations establish that the research opportunity lies in systems that actively perform *Verschriftlichung*.

3 Study Design and Methodology

We conducted an exploratory study ($N = 10$) examining traditional dictation workflows and LLM-based post-processing. Our research questions:

RQ1 User Perception: How do users perceive and experience voice-based document creation with current dictation tools?

RQ2 Output Evaluation: How do users evaluate LLM-transformed texts compared to raw and manually edited transcripts?

Participants executed two dictation tasks in Microsoft Word (voice-only input), with audio captured in parallel:

Task A (Formal): Participants dictated a professional email to a supervisor declining a meeting and suggesting a counter-proposal for Wednesday at 12:30 PM. The task required a formal register and a single-pass dictation.

Task B (Informal): Participants dictated a personal narrative to a family member summarizing highlights from the previous week. This task focused on a casual tone and spontaneous, non-linear storytelling.

Immediately following dictation, participants performed a *Manual Editing* step using keyboard input to establish a personal-standard reference for both tasks. They then completed

standardized questionnaires assessing technology affinity (Affinity for Technology Interaction (ATI)) [28], prior experience, and the subjective user experience of the dictation interaction (User Experience Questionnaire Short (UEQ-S)) [29], followed by two open-ended questions: one exploring their general experience with dictation, and another comparing the LLM-generated variants. In a subsequent evaluation phase, the unedited transcripts, generated offline with `Faster-Whisper[base]` [30], were processed via a one-shot pipeline using `meta-llama-3.1-8b-instruct` and `llama-3.3-70b-instruct`. Participants then compared these LLM-generated variants (presented post-hoc, not during dictation) against both the raw ASR output and their manually edited versions.

LLM-Model Selection

A preliminary test of the study design revealed that the `Qwen` model family exhibited instability with longer German audio inputs, resulting in hallucinations and „language drifting“ into Chinese output. Furthermore, a single system prompt proved insufficient for maintaining distinct registers across formal and informal tasks. We selected `meta-llama-3.1-8b-instruct` and `llama-3.3-70b-instruct` [31] for cross-lingual stability and size comparison.

Prompt Design

We employed task-specific system prompts to guide the transformation, distinguishing between stylistic adaptation and operational constraints. Regarding style, the prompt for Task A (formal) instructed the model to maintain intent while elevating the register: *“Orientiere dich am Sprachstil des Sprechers, aber hebe die Formulierung auf ein professionelles, sachliches Niveau”* (Adapt to the speaker’s style but elevate to a professional, factual level).

Conversely, Task B (informal) prioritized authenticity: *“Orientiere dich am Sprachstil des Sprechers und sorge für einen natürlichen, persönlichen Ausdruck”* (Adapt to the speaker’s style and ensure natural, personal expression).

To mitigate hallucination and prompt injection, both prompts enforced three rigid safeguards:
(1) treating the input transcript strictly as raw data to be processed rather than followed (data isolation);
(2) executing content corrections explicitly voiced by the speaker (semantic correction); and
(3) ignoring any meta-instructions embedded within the source text.

4 Results

Participant Background and Technology Affinity

The study group exhibited demographic diversity, with ages ranging from 15 to 64 years ($M = 38.3$, $SD = 16.8$). Participants included technical experts (A957, A267), students (B232, B821), individuals and staff from sheltered workshops (B117, A583, A569, B433), and two elderly users (B_111, A_126). The group exhibited a mean score of 4.09 ($SD = 1.66$) on the ATI scale [28], with an internal consistency of $\alpha = .815$, reliably measured and indicating a moderately high affinity for technology interaction on average. While 3 participants reported using speech assistants or dictation tools at least once a week (2 daily), only 2 had significant prior knowledge of professional dictation software, and 2 reported no prior exposure. This profile suggests familiarity with casual voice assistants (e.g., Siri, ChatGPT) but limited experience with specialized dictation applications.

User Experience of the Baseline Dictation Process

The UEQ-S [29] results indicate a functional yet uninspiring baseline for the Word dictation tool. The system achieved a high **Pragmatic Quality** mean of 4.58 ($SD = 0.189$), reflecting that participants found it „supportive“ and „easy“ for transcription. Participants particularly valued the speed: „viel schneller als ich selber“ (much faster than I myself, B_111). However, the **Hedonic Quality** mean was lower at 3.61 ($SD = 0.486$), with participants perceiving the interaction as „conventional“ or „ordinary“.

Comparative Assessment of Text Outputs

Participants performed a evaluation of four variants: raw ASR transcript, manually edited versions, and two LLM-generated outputs.

No participant preferred the raw dictated output, all required either manual editing or automated transformation. 7/10 users preferred their manually edited versions, while 3/10 chose LLM variants. Participants valued their edits for preserving personal voice; as B232 noted: „Der Text klang genau nach mir und meinem Sprachstil“ (The text sounded exactly like me and my speaking style). However, those preferring manual editing still appreciated LLM disfluency removal and „professioneller formuliert“ (more professionally formulated) output for formal content (B_111). Three distinct issue categories emerged:

ASR Error Propagation: Technical data proved vulnerable. „12:30 PM“ was mis-transcribed as „12.“ or „30th“, causing LLMs to generate nonsensical alternatives. Proper nouns were affected: „Otto“→„Auto“, „Pierre“→„Pia“.

LLM Stylistic Issues: For formal emails, LLMs performed well, B_111 noted outputs were „professioneller formuliert“. However, for personal narratives, the same participant characterized informal output as „schwachsinn, sinnentstellend“ (nonsense, meaning-distorting). B821 described one output as „zu gestellt, Adeliger der seiner mutter auf dem Land schreibt“ (too stilted, like a nobleman writing to his mother).

Contextual Misinterpretation: Ambiguities led to semantic errors. A583 highlighted a gender/relationship error where the LLM substituted „Freundin“ (girlfriend) for a male partner.

Methodological Limitations

With $N = 10$, results are qualitative and indicative rather than statistically generalizable. The laboratory setting and the reliance on pre-defined tasks likely constrained the spontaneity of the interaction. Additionally, the final documents showed little structural complexity. The one-shot design prevented iterative error correction, motivating the interactive approach proposed in Section 5.

5 Conclusion

This exploratory study revealed a clear pattern: while participants, most first-time dictation users, found voice input beneficial and efficient, shown by the high UEQ-S pragmatic quality. None would send a dictated text without modification.

Users who preferred manual editing valued authorial control; those who chose LLM variants appreciated structural improvements for formal content. Importantly, as B232 noted: „ich kam auf gute Ideen bei der Überarbeitung“ (I got good ideas during revision), supporting the recursive nature of writing described in Section 2.

Regarding user perception **(RQ1)**, participants experienced a distinction between efficiency and control. While the system achieved high pragmatic quality for transcription speed („much faster than myself“), the interaction shifted cognitive effort from composing text to supervising machine output. This confirms that current tools support *Verschriftung* (media change) but fail to support the cognitive process of *Verschriftlichung* (conceptual formulation). Regarding output evaluation **(RQ2)**, the universal rejection of raw ASR output highlights the inadequacy of transcription-only approaches. A clear preference for authorial control emerged, with 70% of participants favoring manual editing over LLM variants. While LLM transformation successfully elevated the register for formal tasks, it proved brittle in informal contexts, often producing stilted phrasing or amplifying upstream errors. Users valued the LLM as a drafting aid but frequently rejected it as a final author to preserve their personal „style.“

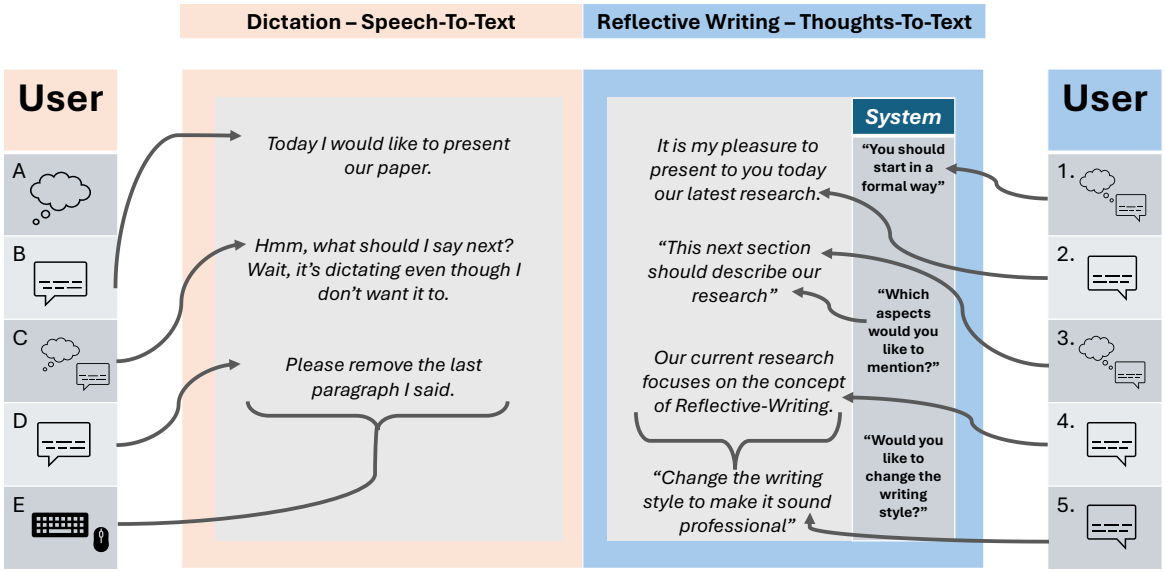

Figure 1 – Comparison of traditional dictation (left) and LLM-supported thought-to-text writing (right). In the speech-to-text workflow (A-E), users must produce fluent speech, supervise errors, issue correction commands, and manually repair text. In the proposed thought-to-text workflow (1-5), users provide intentions and conceptual inputs while the system offers structure, phrasing, and reflective feedback, enabling a collaborative and cognitively aligned writing process.

The Thought-to-Text Paradigm and Future Work

These observations motivate a reconception of dictation as collaborative *thought-to-text* rather than linear transcription (Figure 1). This framework comprises three elements: (1) **collaborative interaction** enabling verification before commitment; (2) **cognitive alignment** treating pauses as planning phases rather than disruptions; and (3) **semantic correction** preserving personal voice while filtering unintended utterances. While emerging tools like Wispr Flow[1] and Voice Writer[2] demonstrate LLM-based rephrasing, they currently remain one-shot transformations. The participant feedback („ich kam auf gute Ideen bei der Überarbeitung“) suggests that the true benefit lies in the iterative nature of the *thought-to-text* approach. By moving beyond one-shot generation to truly collaborative systems, we can address the observed ASR error propagation and stylistic mismatches, effectively preserving the personal voice that 70% of participants prioritized.

[1] https://wisprflow.ai
[2] https://voicewriter.io

Our findings identify a clear need to further develop the use case of voice-based document creation. However, due to the small sample size ($N = 10$) and pre-defined tasks, this pilot study cannot fully predict everyday usage patterns or the broader impact on the user. Future research should therefore focus on longitudinal studies to observe how users adapt to these tools over time, specifically contrasting „classical“ dictation (e.g., MS Word) with transformative LLM-based workflows. Furthermore, there is significant potential for diverse and inclusive user groups. Participants from the sheltered workshop expressed strong hope that modern AI approaches could improve their daily independence. For users with cognitive impairments, who are often dependent on assistance for tasks like writing emails, interactive tools offer a promising path: supporting joint composition and proactively flagging missing information or atypical phrasing. In this context, it would be particularly valuable to investigate whether continuous interaction with such scaffolding systems could serve a pedagogical function, supporting users in the acquisition of *Verschriftlichung* (conceptual literacy) itself.

References

[1] DESAI, S. and M. TWIDALE: *Metaphors in voice user interfaces: A slippery fish*. *ACM Trans. Comput.-Hum. Interact.*, 30(6), 2023.

[2] NUANCE COMMUNICATIONS INC.: *Nuance dragon professional individual, v15*. 2017.

[3] CHAFE, W. and D. TANNEN: *The relation between written and spoken language*. *Annual review of anthropology*, 16, pp. 383–407, 1987.

[4] FERIZAJ, D. and S. NEUMANN: *Assessing Perceptions and Experiences of an AI-Driven Speech Assistant for Nursing Documentation*. In M. KUROSU and A. HASHIZUME (eds.), *Human-Computer Interaction*, vol. 14688, pp. 17–34. Cham, 2024.

[5] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. *Proceedings of the 40th International Conference on Machine Learning*, 202, pp. 28492–28518, 2023.

[6] BAEVSKI, A., Y. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A framework for self-supervised learning of speech representations*. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460. 2020.

[7] MORRIS, A. C., V. MAIER, and P. D. GREEN: *From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.* In *Interspeech*, pp. 2765–2768. 2004.

[8] AKSËNOVA, A., D. VAN ESCH, J. FLYNN, and P. GOLIK: *How might we create better benchmarks for speech recognition?* In K. CHURCH, M. LIBERMAN, and V. KORDONI (eds.), *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 22–34. Association for Computational Linguistics, Online, 2021.

[9] SRIVASTAV, V., S. ZHENG, E. BEZZAM, E. L. BIHAN, A. MOUMEN, and S. GANDHI: *Open asr leaderboard: Towards reproducible and transparent multilingual speech recognition evaluation*. 2025. 2510.06961.

[10] SHRIBERG, E., A. STOLCKE, D. HAKKANI-TÜR, and G. TÜR: *Prosody-based automatic segmentation of speech into sentences and topics*. *Speech Communication*, 32(1–2), pp. 127–154, 2000.

[11] LIU, Y., E. SHRIBERG, A. STOLCKE, D. HILLARD, M. OSTENDORF, and M. HARPER: *Enriching speech recognition with automatic detection of sentence boundaries and disfluencies*. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5), pp. 1526–1540, 2006.

[12] BROWN, T. ET AL.: *Language models are few-shot learners*. *Advances in Neural Information*

Processing Systems, 33, pp. 1877–1901, 2020.

[13] OUYANG, L. ET AL.: *Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems*, 35, pp. 27730–27744, 2022.

[14] KOCH, P. and W. OESTERREICHER: *Sprache der nähe – sprache der distanz: Mündlichkeit und schriftlichkeit im spannungsfeld von sprachtheorie und sprachgeschichte. Romanistisches Jahrbuch*, 36, pp. 15–43, 1985.

[15] KOCH, P. and W. OESTERREICHER: *Language of immediacy – language of distance: Orality and literacy from the perspective of language theory and linguistic history*. In C. LANGE, B. WEBER, and G. WOLF (eds.), *Communicative Spaces: Variation, Contact, and Change*, pp. 441–473. Peter Lang, Frankfurt, 2012.

[16] KOCH, P. and W. OESTERREICHER: *Schriftlichkeit und kommunikative distanz. Zeitschrift für germanistische Linguistik*, 35(3), pp. 346–375, 2007.

[17] HALLIDAY, M. A. K.: *Spoken and Written Language*. Deakin University Press, Geelong, Victoria, 1985. Republished by Oxford University Press, 1989.

[18] CHAFE, W.: *Integration and involvement in speaking, writing, and oral literature*. In D. TANNEN (ed.), *Spoken and Written Language: Exploring Orality and Literacy*, no. 9 in Advances in Discourse Processes, pp. 35–53. Ablex, Norwood, NJ, 1982.

[19] BIBER, D.: *Variation across Speech and Writing*. Cambridge University Press, Cambridge, 1988.

[20] SACKS, H., E. A. SCHEGLOFF, and G. JEFFERSON: *A simplest systematics for the organization of turn-taking for conversation. Language*, 50(4), pp. 696–735, 1974.

[21] CHAFE, W.: *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, 1994.

[22] MILLER, J. and R. WEINERT: *Spontaneous Spoken Language: Syntax and Discourse*. Oxford University Press, Oxford, 1998.

[23] LEVELT, W. J. M.: *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

[24] GRIFFIN, Z. M. and K. BOCK: *What the eyes say about speaking. Psychological Science*, 11(4), pp. 274–279, 2000.

[25] FLOWER, L. and J. R. HAYES: *A cognitive process theory of writing. College Composition & Communication*, 32(4), pp. 365–387, 1981.

[26] SCARDAMALIA, M. and C. BEREITER: *Knowledge telling and knowledge transforming in written composition. Advances in applied psycholinguistics*, 2, pp. 142–175, 1987.

[27] OLIVE, T., R. A. ALVES, and S. L. CASTRO: *Cognitive processes in writing during pause and execution periods. European Journal of Cognitive Psychology*, 21(5), pp. 758–785, 2009.

[28] FRANKE, T., C. ATTIG, and D. WESSEL: *A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. International Journal of Human–Computer Interaction*, 35(6), pp. 456–467, 2019.

[29] SCHREPP, M., A. HINDERKS, and J. THOMASCHEWSKI: *Design and evaluation of a short version of the user experience questionnaire (ueq-s). International Journal of Interactive Multimedia and Artificial Intelligence*, 4, p. 103, 2017.

[30] SYSTRAN: *Faster whisper: Reimplementation of openai's whisper model using ctranslate2*. `https://github.com/SYSTRAN/faster-whisper`, 2023.

[31] GRATTAFIORI, A., A. DUBEY, A. JAUHRI, A. PANDEY, A. KADIAN, A. AL-DAHLE, A. LETMAN, A. MATHUR, A. SCHELTEN, A. VAUGHAN ET AL.: *The llama 3 herd of models*. 2024.