The **Be**yond **Tra**nscription **C**hallenge: End-to-End Long Clinical Audio Summarization

Jiyun Chun
Computer Science and Engineering
The Ohio State University
Columbus, United States
chun.203@osu.edu

Samuele Cornell
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, USA
samuele.cornell@ieee.org

Syed-Amad Hussain
Abigail Wexner Research Institute
Nationwide Children's Hospital
Columbus, Ohio
https://orcid.org/0000-0002-7529-8644

Leibny Paola Garcia Perera
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, United States
https://orcid.org/0000-0002-7449-5726

Ahmed Hassoon
Data Science and AI Institute
Johns Hopkins University
Baltimore, United States
https://orcid.org/0000-0002-4930-3331

Markus Müller
IAR
Karlsruhe Institute of Technology
Karlsruhe, Germany
https://orcid.org/0009-0000-1701-0019

Thomas Schaaf
Health Information Systems
Solventum
Pittsburgh, United States
https://orcid.org/0000-0002-9569-4759

Andrew Perrault
Computer Science and Engineering
The Ohio State University
Columbus, United States
perrault.17@osu.edu

***Abstract*—Recent open large audio language models perform well on short audio but struggle with long, multi-speaker conversations. Clinical documentation is a useful setting for evaluating this ability—in doctor–patient conversations, symptoms, medical history, examination findings, clinical assessments, and treatment plans are distributed across the conversation. In this setting, end-to-end models that generate SOAP notes directly from audio often add unsupported information or omit relevant details, and cascaded ASR-then-summarization pipelines are commonly reported to outperform them. We introduce the Beyond Transcription Challenge (BeTraC) to address this gap. Participating systems must generate SOAP notes directly from long doctor–patient conversation audio without intermediate transcription. BeTraC includes a Lightweight Track for a single end-to-end model with at most 6 billion parameters and a Heavyweight Track that allows multi-model systems, tools, and agent-based architectures with at most 36 billion parameters in total. Training data, in the form of synthetic doctor-patient conversation audio and SOAP pairs (Synth-DoPaCo), was provided to the teams. Official rankings used Open Medical Concept F1 on held-out Synth-DoPaCo, while post-competition evaluation assessed Faithfulness, Coverage, Structure, and Conciseness with an LLM judge and tested generalization to recorded OSCE interviews. Results showed strong transfer from synthetic to real data under both metrics. The two metrics agreed closely for high-scoring systems but diverged for low-scoring ones, exposing systematic blind spots in the automated metric.**

***Index Terms*—Clinical audio summarization, clinical note generation, long-form audio understanding, audio-language models, synthetic data.**

I. Introduction

Recent open large audio language models (LALMs) [1]–[7] have advanced rapidly in speech recognition, audio question answering, and voice based interaction [8]–[11], suggesting that they may be able to understand long spoken interactions in real world settings. However, existing long audio benchmarks show that performance declines sharply as input duration increases [9], [11]. Clinical documentation is an important application of this technology. In doctor patient conversations, symptoms, medical history, examination findings, clinical assessments, and treatment plans are often distributed across several minutes of conversation. Generating a useful ***S**ubjective, **O**bjective, **A**ssessment, **P**lan* (SOAP) note therefore requires speech recognition, long context understanding, clinical information extraction, and structured generation. Compared with other long audio summarization tasks such as meeting summarization [12]–[14], clinical documentation is more structured and therefore easier to evaluate. This structure has also motivated related shared tasks. MEDIQA Chat 2023 [15] addressed clinical note generation from text based doctor patient conversations and synthetic conversation generation from clinical notes.

Clinical audio documentation typically relies on an ASR to text cascade. An ASR model transcribes the doctor patient conversation, and a separate LLM converts the transcript into a SOAP note. Although cascades generally outperform models that generate SOAP notes directly from audio, both approaches remain limited for the small open weight models considered in this challenge. Cascades also make it difficult to separate audio understanding from text summarization. In addition, ASR errors propagate to the final note, while transcription removes clinically meaningful nonlexical cues such as prosody,

hesitation, stress, coughing, and breathing difficulty. Faithfully using these cues remains a longer term goal because current audio synthesis cannot fully reproduce them. BeTraC therefore focuses on narrowing the gap in generating clinically grounded SOAP notes with end to end models.

This gap leads to the main research question of the **Beyond Transcription Challenge (BeTraC)**[1]: can open-weight LALM systems be constructed that generate faithful and clinically useful SOAP notes from doctor-patient conversation audio without using intermediate transcription? In particular, if many clinical claims generated by existing end-to-end audio-to-SOAP systems are not well supported by the actual conversation, can we reduce these hallucinations? BeTraC is designed as an end-to-end clinical audio reasoning challenge and enforces a no transcription rule. During inference, no textual representation of the audio, including a transcript, summary, or entity list, may pass between components. Only the final model may produce text derived from the audio.

Because privacy regulations such as HIPAA severely restrict the public release of real doctor patient conversation datasets [16]–[20], BeTraC uses Synth-DoPaCo [21], an open synthetic dataset developed with clinicians for training and testing, together with automated evaluation. The dataset is generated entirely with open weight models. LLM role play produces the conversation transcript, the transcript is used to generate a SOAP note, and speech synthesis renders the conversation as audio. The official leaderboard is based on Open Medical Concept F1 on the Synth-DoPaCo test set. This metric combines MeSH keyword matching with named entity recognition using scispaCy [22].

These design choices raise questions about external validity because top systems may overfit the synthetic data or the Concept F1 metric. We examine these concerns through two independent post competition evaluations. We reevaluate the systems on 275 human acted doctor patient conversations, most of which are OSCE interviews [20], to assess overfitting to synthetic data, and evaluate every note with an LLM as a judge pipeline developed with clinicians [21] to assess overfitting to the metric. The results show strong transfer from synthetic to real data, with system rankings on synthetic and real conversations closely aligned under both metrics. Concept F1 and the LLM judge also agree for high scoring systems but diverge for low scoring systems, revealing systematic blind spots in the automated metric. Overall, BeTraC provides a standardized training set and benchmark for long form audio reasoning, faithful clinical documentation, hallucination analysis, and LALM evaluation. Our code will be released publicly.

II. Background

A. Synth-DoPaCo

BeTraC uses the Synth-DoPaCo dataset (Table I) [21] to evaluate the ability of LALMs to generate structured clinical documentation from long doctor–patient conversations. Synth-DoPaCo is a fully synthetic dataset of 8,800 first-visit encounters (∼1,300 hours), each paired with an audio recording and a reference SOAP note. The dataset was developed to address the lack of resources for long-form audio understanding and was generated entirely using open-weight models with permissive licenses.

TABLE I
STATISTICS AND RELEASE SETTINGS OF THE BETRAC DATASETS.

Split	#Conv.	Hrs	Release
Train [21]	7,200	∼1,087	Audio + reference SOAP notes
Dev [21]	400	∼59	Audio + reference SOAP notes
Test [21]	600	∼101	Audio only*
OSCE [20]	272	∼55	Audio only*
Realistic	3	∼0.4	Audio only*

*Reference SOAP notes are withheld for blind evaluation.

B. Data Generation Pipeline

Synth-DoPaCo generates each example using a three stage pipeline built entirely with open weight models. It consists of persona driven dialogue generation using Gemma 3 27B IT [23], multi speaker audio synthesis with realistic acoustics and non speech events, and reference note generation from clinical facts extracted from the ground truth transcript using Kimi K2 Thinking [24]. Using only these facts helps limit unsupported clinical information. The pipeline is described fully in [21]. A 10 minute conversation contains only ˜1,200 text tokens but roughly ten times as many audio tokens, which contributes to the long context difficulty faced by end to end systems.

Three distinct open-weight models play roles in this paper: Gemma 3 27B IT generates the dialogues, Kimi K2 Thinking generates the reference SOAP notes, and, for the post-competition analysis (§ III-E), Gemma 4 31B IT [25] serves as the LLM judge.

III. The BeTraC Challenge

A. Task Input, Output, and Submission Format

Each BeTraC test instance consists of a doctor-patient conversation audio. Participating systems receive the raw audio as input and generate a SOAP note containing the clinical information from the conversation. Each audio file must be processed independently without using information from other test files. The target output is a SOAP note with information organized into Subjective, Objective, Assessment, and Plan sections. The Subjective section contains patient reported symptoms and medical history. The Objective section contains observations and examination findings. The Assessment section contains the clinical assessment. The Plan section contains treatment and follow up plans.

Following the challenge timeline, teams were required to submit brief system descriptions by June 24, 2026 to receive the test audio. Teams then had approximately one week to generate and submit SOAP notes, with submissions due on July 1, 2026. For each team that did so, the organizers provided

[1]https://betrac.github.io/

them with 875 test audios in a random order, without reference notes. This consisted of 600 Synth-DoPaCo test set audios and 275 out-of-distribution (OOD) test audios, which are described in § III-E. Teams were instructed not to modify their systems between receiving the test audio and submitting their generated SOAP notes. Each submission included the raw model output for every test audio file, a script that extracts the final SOAP note from the raw output, and the extracted SOAP notes. We required that the raw output contain every token generated by the model, including chain of thought, tags, delimiters, and any other generated content. Heavyweight systems that use multiple models or tools were required to submit a complete execution trace containing the inputs and outputs of all model and tool calls.

B. *Tracks and Eligibility*

All submitted systems had to use approved open-weight models meeting four eligibility criteria: publicly available weights, a reproducible (pinned) checkpoint, a published model card or technical report, and a forward-pass parameter count within the track budget. Proprietary or API-only models were disallowed anywhere in the system, including for generating synthetic training data such as SOAP notes or TTS audio.

The defining constraint is the *no-transcription rule*, which governs both what may pass between components at inference and how the final model may be trained.

At inference. No textual representation of the spoken content (e.g., transcript, paraphrase, summary, keyword or clinical-entity list, or draft note) may pass between separate models or components, and only the final model may emit text derived from the audio. This excludes conventional ASR cascades. ASR-pretrained encoders may still be used, but only as non-decoding acoustic feature extractors.

At training. Multi-task training and distillation are permitted, and the final model's chain-of-thought (CoT) is allowed and unconstrained in form, i.e., emergent transcript-like text is not a violation. The one restriction: the model must not have been trained with a transcript of the input audio as a CoT target, prefix, or supervision signal, nor distilled from a transcript-aligned teacher [26].

Prohibited. Post-hoc learned rewriting of the note, cross-file context, test-label use, and output-level ensembling (including self-consistency). Only deterministic, non-learned post-processing (formatting, header insertion) may edit the final note.

BeTraC has two tracks that differ in architecture and total forward-pass parameters. The **Lightweight (LW)** track requires a single end-to-end model invoked exactly once per file, with no separately executed preprocessing or tools. Learned audio components (encoders, diarization, VAD) are allowed only as integrated submodules and count toward the 6B total parameter budget. The **Heavyweight (HW)** track permits multi-model pipelines, agents, tools, and retrieval (medical ontologies, drug databases, guidelines, PubMed) under a 36B-parameter budget that sums all learned components in the forward pass. Audio tools (diarization, speaker embedding, VAD) may run as separate steps but must return only non-linguistic outputs (timestamps, speaker IDs, acoustic features).

C. *Baselines*

We provided per-track baselines that generate SOAP notes directly from audio without a separate ASR stage. The baselines are zero-shot: the models are prompted to produce a SOAP note with no fine-tuning or prompt optimization. The **Lightweight** baseline is Qwen2.5-Omni-3B [1]; the **Heavyweight** baseline pool comprised Qwen2.5-Omni-3B/7B and Qwen3-Omni-30B-A3B (Instruct and Thinking) [5].

For reference, we also report ineligible cascaded systems that transcribe before summarizing: Whisper-large-v3 with Qwen3-30B-A3B reaches a development-set C-F1 of 0.288 [27], [28], and Qwen3-ASR-1.7B with Qwen3-30B-A3B reaches 0.281 [27], [29]. Because they pass an audio-derived transcript between components, these systems are ineligible for the BeTraC ranking.

D. *Competition Evaluation*

The official evaluation compares the SOAP notes generated by each system with the reference SOAP notes (produced by the Synth-DoPaCo Kimi K2 pipeline). Teams were provided with reference SOAP notes for the training set and code to compute these metrics.

*1) **Open Medical Concept F1**:* Open Medical Concept F1 (C-F1) is the primary ranking metric. The default matcher combines keyword matching over a curated MeSH vocabulary [30] with biomedical named-entity recognition, and the resulting concept identifiers are compared as sets between the generated and reference notes. For dialogue i, let C_i^{pred} and C_i^{ref} be the concept-identifier sets of the generated and reference notes and $O_i = |C_i^{\text{pred}} \cap C_i^{\text{ref}}|$ their overlap. Precision and recall are

$$P_i = \frac{\left|C_i^{\text{pred}} \cap C_i^{\text{ref}}\right|}{\left|C_i^{\text{pred}}\right| + \epsilon}, \quad (1) \qquad R_i = \frac{\left|C_i^{\text{pred}} \cap C_i^{\text{ref}}\right|}{\left|C_i^{\text{ref}}\right| + \epsilon}, \quad (2)$$

where $\epsilon = 10^{-6}$.

The C-F1 score for dialogue i and the final macro-averaged score over all N evaluation dialogues are computed as

$$F_i = \frac{2P_iR_i}{P_i + R_i + \epsilon} \qquad \text{and} \qquad \text{Concept F1} = \frac{1}{N}\sum_{i=1}^{N} F_i. \quad (3)$$

Concept precision and concept recall are also reported as macro averages over all evaluation dialogues. Because the extracted concept identifiers are represented as sets, repeated occurrences of the same concept within a note are counted only once.

*2) **ROUGE**:* ROUGE measures textual overlap between each generated SOAP note and its reference note [31], [32]. We report ROUGE 2 F1, ROUGE 3 F1, and ROUGE L F1 as secondary metrics. ROUGE 2 and ROUGE 3 measure bigram

TABLE II

ID AND OOD EVALUATION RESULTS, ORDERED BY ID C-F1. BOLDFACE INDICATES THE BEST PRIMARY SUBMISSION WITHIN EACH TRACK AND EVALUATION SET FOR EACH METRIC. † INDICATES A SCORE THAT EXCEEDS THE CORRESPONDING REFERENCE-NOTE SCORE. GRAY ± VALUES INDICATE 95% CONFIDENCE INTERVAL HALF-WIDTHS. HALLUCINATION IS REPORTED AS A PER-CLAIM PERCENTAGE, WITH LOWER VALUES BEING BETTER.

System	Automated metrics														LLM-as-a-Judge									
	C-F1		C-P		C-R		R-2		R-3		R-L		Words		Faith.		Cov.		Struct.		Conc.		Halluc. (%)	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Reference notes	–	–	–	–	–	–	–	–	–	–	–	–	355.1	287	4.85 ±0.03	4.85 ±0.05	4.66 ±0.05	4.27 ±0.10	4.89	4.78	4.39	4.64	0.20	0.17
Baseline																								
Qwen2.5-Omni-3B	0.309 ±0.005	0.251 ±0.008	0.343	0.261	0.295	0.262	0.116	0.079	0.050	0.029	0.189	0.144	817.4	935.2	2.92 ±0.07	1.96 ±0.08	3.63 ±0.08	3.03 ±0.11	2.95	2.67	2.28	2.27	8.91	19.72
Qwen2.5-Omni-7B	0.326 ±0.006	0.304 ±0.009	0.379	0.324	0.296	0.301	0.140	0.103	0.061	0.039	0.224	0.180	448.2	722.1	3.75 ±0.08	2.75 ±0.11	4.04 ±0.07	3.70 ±0.10	2.97	2.77	3.00	2.60	5.13	14.52
Qwen3-Omni-30B-Instruct	0.318 ±0.004	0.293 ±0.006	0.267	0.241	0.403	0.392	0.146	0.120	0.064	0.048	0.231	0.203	563.1	547.1	2.69 ±0.05	2.42 ±0.07	4.80† ±0.03	4.69† ±0.06	4.87	4.51	4.09	3.88	19.02	21.84
Qwen3-Omni-30B-Thinking	0.285 ±0.004	0.265 ±0.006	0.239	0.214	0.363	0.361	0.133	0.116	0.055	0.044	0.219	0.196	509.8	479.7	2.54 ±0.05	2.03 ±0.06	4.76† ±0.03	4.73† ±0.05	4.86	4.44	4.03	3.89	17.78	24.39
Lightweight Track																								
TalTech	**0.543** ±0.006	0.505 ±0.010	0.579	**0.562**	**0.518**	0.471	0.366	0.317	0.250	0.194	0.434	0.401	323.4	240.4	**4.71** ±0.05	**4.31** ±0.11	**4.49** ±0.06	3.85 ±0.10	**4.88**	**4.86**†	4.16	4.55	0.25	1.25
NTT-HI-CS	0.540 ±0.006	**0.510** ±0.008	**0.587**	0.555	0.507	**0.479**	**0.381**	**0.348**	**0.265**	**0.219**	**0.452**	**0.435**	310.1	247.2	4.58 ±0.06	3.91 ±0.12	4.37 ±0.06	3.64 ±0.08	4.80	4.73	4.37	4.41	0.24	1.20
KUSLP	0.515 ±0.006	0.465 ±0.008	0.557	0.509	0.488	0.437	0.367	0.330	0.253	0.206	0.433	0.422	318.9	248.8	4.59 ±0.06	3.66 ±0.13	4.46 ±0.06	3.80 ±0.09	4.72	4.73	4.11	4.44	0.38	3.89
JYVA	0.501 ±0.006	0.455 ±0.008	0.543	0.458	0.473	0.461	0.351	0.312	0.243	0.195	0.412	0.376	323.0	303.7	4.61 ±0.06	3.73 ±0.12	4.25 ±0.07	**4.04** ±0.09	4.16	3.25	3.29	3.13	**0.22**	2.09
KIT-ISL-AI4LT	0.495 ±0.006	0.462 ±0.008	0.546	0.524	0.461	0.423	0.360	0.319	0.250	0.201	0.432	0.407	291.5	220.3	4.62 ±0.05	4.06 ±0.11	4.37 ±0.06	3.95 ±0.10	4.83	4.84†	**4.56**†	**4.77**†	0.27	**0.73**
DirectSense	0.434 ±0.006	0.448 ±0.008	0.541	0.482	0.369	0.428	0.249	0.241	0.153	0.134	0.331	0.321	240.9	246.1	4.13 ±0.07	3.53 ±0.12	3.93 ±0.07	3.96 ±0.10	3.93	4.28	3.45	3.75	2.29	4.25
ASLP	0.397 ±0.006	0.382 ±0.007	0.393	0.345	0.413	0.442	0.222	0.192	0.134	0.103	0.305	0.282	437.5	407.8	2.18 ±0.08	1.19 ±0.06	3.30 ±0.08	3.03 ±0.09	3.92	4.08	3.81	3.68	18.82	34.10
IASP	0.362 ±0.006	0.347 ±0.007	0.406	0.381	0.335	0.329	0.130	0.113	0.080	0.066	0.181	0.158	743.0	728.1	2.19 ±0.08	1.47 ±0.06	2.70 ±0.08	2.50 ±0.07	2.64	3.08	1.19	1.19	18.75	35.31
Heavyweight Track																								
TalTech	**0.563** ±0.006	**0.556** ±0.008	**0.601**	0.571	**0.536**	**0.549**	**0.401**	**0.381**	**0.280**	**0.242**	**0.468**	**0.468**	312.2	257.1	4.80 ±0.04	**4.63** ±0.08	4.50 ±0.06	4.14 ±0.10	**4.91**†	4.80†	**4.41**†	4.59	**0.08**†	**0.30**
KUSLP	0.544 ±0.006	0.530 ±0.009	0.596	**0.580**	0.509	0.494	0.393	0.374	0.277	0.241	0.459	**0.468**	310.6	248.0	**4.81** ±0.04	4.53 ±0.09	4.32 ±0.06	3.67 ±0.09	4.82	4.84†	4.34	**4.76**†	0.17†	0.42
NTT-HI-CS	0.539 ±0.006	0.517 ±0.009	0.579	0.561	0.511	0.485	0.377	0.348	0.261	0.218	0.448	0.440	315.1	246.9	4.47 ±0.06	3.60 ±0.12	4.38 ±0.06	3.65 ±0.10	4.73	4.68	4.30	4.34	0.41	2.06
ASLP	0.506 ±0.006	0.501 ±0.008	0.504	0.481	0.518	0.532	0.341	0.295	0.227	0.168	0.415	0.405	365.6	333.7	4.36 ±0.07	4.41 ±0.10	**4.69**† ±0.05	4.38† ±0.10	4.79	**4.85**†	4.31	3.96	0.94	0.79
Slate Lab	0.475 ±0.010	0.462 ±0.011	0.510	0.465	0.467	0.475	0.300	0.294	0.206	0.182	0.357	0.370	471.5	365.2	4.51 ±0.08	4.26 ±0.11	4.08 ±0.12	**4.45**† ±0.13	3.91	4.16	3.07	3.89	2.15	1.77

and trigram overlap, respectively. ROUGE L is based on the longest common subsequence. Each score is first computed for each dialogue and then macro averaged over the evaluation set.

E. Post-competition Evaluation

We conducted additional post competition analysis with out-of-domain human-acted doctor patient conversations and by using LLM-as-a-judge SOAP note evaluation (Table I). These results did not affect the outcome of the competition.

The out-of-distribution (OOD) conversations came from two sources. First, we used 272 human-roleplayed doctor–patient conversation audios in the format of Objective Structured Clinical Examinations (OSCEs) focusing on respiratory cases [20], which provide ground-truth transcripts and are designed for high audio quality and few disfluencies. Second, we used three real-world human-acted conversations recorded at the JSALT 2025 workshop, including a speaker with a substantial accent and captured with a single microphone. For both, transcripts were produced semi-automatically, and reference SOAP notes were generated with the same Kimi K2 Thinking pipeline as Synth-DoPaCo. Neither source was provided as BeTraC training data; the JSALT recordings, moreover, are private and were never released, so they cannot have been seen by any LALM, giving a contamination-free OOD test.

For LLM-as-a-judge evaluation, we use the two-stage pipeline developed for Synth-DoPaCo, with Gemma 4 31B IT as the judge model [25] (distinct from the Kimi K2 ground truth note-construction model, to avoid self-preference bias). We run inference in FP8 precision using vLLM [33], with the temperature set to 0 and maximum generation lengths of 16,000 and 12,000 tokens for claim extraction and final judgment, respectively. The LLM-as-a-judge evaluation required approximately 23 H100 GPU-seconds per dialogue. [2] This pipeline first performs claim extraction from the SOAP note. Then, in the second stage, the judge model evaluates the *Faithfulness, Coverage, Structure*, and *Conciseness* of the claim list relative to the dialogue transcript. Faithfulness evaluates claim support and contradictions. Coverage evaluates the completeness of documentation relative to transcript evidence. Structure evaluates SOAP formatting and section placement. Conciseness evaluates redundancy and low value content. We consider faithfulness and coverage to be the most important properties of SOAP notes, with structure and conciseness included for their roles in readability.

IV. RESULTS

A. Submitted System Descriptions

Nine teams participated (eight in LW, five in HW, four in both) and could also submit off-leaderboard contrast systems. Table III summarizes their designs; we highlight the trends here and analyze their effect in Section V.

Every system adapted a pretrained open audio-language model rather than training one from scratch: Qwen2.5-Omni-3B [1] was the most common backbone (six teams), with Voxtral [6] and (HW-only) Qwen3-Omni-30B-A3B [5] as alternatives. Teams differed mainly in how they adapted the backbone rather than which one they chose. Most used full fine-tuning or LoRA/QLoRA [41], [42], while JYVA instead trained a QFormer [43] to align diarization-aware audio features. Four teams then added metric-aligned RLVR rewarding

[2] In experiments on notes produced by the baseline model, we found that the LLM gets slightly more generous as it is quantized (within 0.1) on each rating dimension. It is additionally substantially more generous (+0.4) when run with vLLM compared to Ollama, for unknown reasons.

TABLE III

DESIGN SPACE OF THE FINAL BETRAC SUBMISSIONS. A SPLIT BACKBONE SWATCH DENOTES DIFFERENT LW / HW MODELS. ✓ IN EXT./AUG. MARKS USE OF EXTERNAL CLINICAL CORPORA; (TTS) MARKS SYNTHETIC-SPEECH AUGMENTATION. IN STRUCT./TOOLS, (ENCODER) MARKS DIARIZATION INTEGRATED AT THE AUDIO ENCODER AND (TOOL) AN EXTERNAL DIARIZATION STEP. A DASH DENOTES A TECHNIQUE NOT USED. BACKBONE COLORS: ■ QWEN2.5-OMNI, ■ VOXTRAL, ■ QWEN3-OMNI-30B. † LW ONLY; ‡ SLATE LAB'S CONTRAST SYSTEM ONLY.

System	Track	Backbone	Params	Adapt.	RL/Pref	Ext./Aug.	Long-ctx	Struct./Tools
TalTech [34]	L,H	■ Voxtral 3B / 24B	3/24B	full	GSPO	–	–	–
KUSLP [35], [36]	L,H	■ Qwen2.5-Omni / ■ Qwen3-Omni-30B	3B / 30B	LoRA	GRPO	✓ (TTS)	time-compress	–
NTT-HI-CS [37]	L,H	■ Qwen2.5-Omni-3B	4.0B	full	DAPO	✓†	avg-pool	–
ASLP [38]	L,H	■ Qwen2.5-Omni / ■ Qwen3-Omni-30B	4B / 30B	full	GRPO	✓	–	–
JYVA [39]	L	■ Qwen2.5-Omni + DiariZen	6.0B	QFormer	–	✓	–	diar (encoder)
KIT-ISL-AI4LT [40]	L	■ Qwen2.5-Omni-3B	4.0B	LoRA	–	✓	–	–
Slate Lab	H	■ Qwen3-Omni-30B	35B	LoRA	–	–	–	diar (tool)‡
DirectSense	L	■ Voxtral-Mini-3B	4.7B	prompt-only	–	–	–	–
IASP	L	■ Qwen2.5-Omni-3B	4.1B	QLoRA	–	–	–	–

C-F1 (and sometimes ROUGE or SOAP-format terms): TalTech (GSPO [44]), NTT-HI-CS (DAPO [45]), and KUSLP and ASLP (GRPO [46]). DirectSense did no fine-tuning at all, prompt-tuning a frozen Voxtral-Mini-3B [6]. All teams trained on Synth-DoPaCo, four supplementing it with external corpora [18], [19], [47] and KUSLP additionally augmenting the audio via TTS re-rendering. With conversations averaging nine minutes, three teams added explicit long-context mechanisms in order to shorten the input sequence and thus reduce computational burden: feature average-pooling (NTT-HI-CS) and waveform time-compression (KUSLP).

Notably, the Heavyweight track's headline affordances went almost unused: no system used retrieval, ontologies, or tool-calling agents—all teams generated the note in a single pass. The only exception was speaker diarization, used by two systems: JYVA, which integrates DiariZen [48] at the audio encoder, and Slate Lab's contrast system, which runs a separate Pyannote step. All post-processing was deterministic header/format cleanup, with no learned model editing the note.

B. *Metric Scores*

Table II reports the results for each team, baseline, and reference on each metric and Figure 1 shows the relationships across metrics and datasets. The reference notes are judged as the strongest by the LLM judge. They are particularly strong in faithfulness (4.85 ID, OOD) and structure (4.89 ID, 4.78 OOD). They are judged to be weaker in coverage (4.66 ID, 4.27 OOD) and conciseness (4.39 ID, 4.64 OOD).

TalTech won both the LW and HW tracks of the challenge. Their systems were ranked first in C-F1 in all settings except LW OOD, where they ranked second. Their LW system would be the likely LW pick of the LLM judge, receiving the highest faithfulness scores both ID and OOD and having generally high other scores. Their HW system was possibly the strongest in the competition overall. It is highly faithful (4.80 ID, 4.63 OOD), had the lowest hallucination rates (0.08% ID, 0.30% OOD), and scored well in structure and conciseness, often exceeding the reference notes. However, it suffers from weaker coverage, especially OOD (4.14), which we will discuss in more detail.

NTT-HI-CS received second place in LW and third place in HW. The LW track was extremely close—NTT-HI-CS's LW system had C-F1 of just 0.003 lower than TalTech, and, in LW OOD, they had higher C-F1 than TalTech. Their ROUGE and Concept Recall scores were the highest in LW, both ID and OOD. Their LLM judge scores were weaker than TalTech in all settings except conciseness in LW ID.

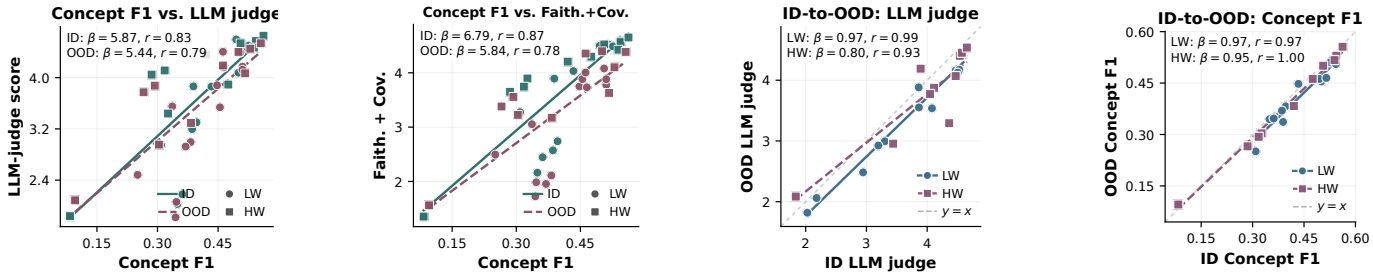
KUSLP received third place in LW and second in HW. Their HW system was highly faithful, receiving the highest score ID and second highest OOD, and produced well structured and concise notes, but suffered from weak coverage, especially OOD (3.67).

The HW system of ASLP deserves special mention. Despite having lower Concept and ROUGE scores ID and OOD, it was highly regarded by the LLM judge, where it received the highest coverage score ID and second highest OOD, and the highest structure score OOD. This system produces longer notes (15–30% longer than TalTech's) that the LLM judged as more comprehensive but less faithful. We discuss this case further in the next section.

V. DISCUSSION

TalTech, NTT-HI-CS, KUSLP, and ASLP (HW) built the strongest systems, and all four used the same recipe: full or LoRA fine-tuning, then some form of RLVR, on a Qwen or Voxtral base model. This was the winning approach in the challenge. What set the top teams apart was mostly how they adapted the model, not which base model they picked: the two best Lightweight systems use different base models yet score within 0.003 C-F1 of each other.

In Fig. 1, we examine the relationships between C-F1 metric vs. the average LLM as a judge score, as well as metric scores transfer from ID to OOD, using all systems including contrasts. We find a clear pattern. C-F1 transfers near perfectly from ID to OOD (Fig. 1d), with $r \geq 0.97$. LLM judge scores transfer a bit less well, with $r \geq 0.93$ (Fig. 1c). In each case, we see no clear signs of overfitting to ID, with outlying points on both sides of the line (e.g., Slate Lab received an average LLM judge score of 3.9 ID and 4.2 OOD). These findings are notable given that the ID audio are entirely synthetic, and OOD are real recorded conversations. This suggests that synthetic data generation, and particularly RLVR in synthetic environments, is a viable path forward to improving the general long context summarization capabilities of LALMs. This near-linear ID $\rightarrow$ OOD relationship echoes the "accuracy on the line" phenomenon, where OOD performance is strongly



(a) **(b)** **(c)** **(d)**

Fig. 1. Relationships between evaluation performance on the ID and OOD subsets across primary submissions, contrast systems, and baselines. (a) C-F1 vs. the aggregate LLM-as-a-judge score on ID and OOD data. (b) C-F1 vs. the average of faithfulness and coverage on ID and OOD data. (c) ID vs. OOD LLM-as-a-judge performance. (d) ID vs. OOD C-F1 performance. Marker shape distinguishes LW and HW systems. Regression slopes and Pearson correlations are shown within each panel.

linearly predicted by ID performance across a wide range of models [49]. Our results extend the observation to generative clinical summarization and, notably, to a synthetic-to-real distribution shift.

The clearest signs of weakness in transferability are between C-F1 and the LLM judge scores (Fig. 1a-b), which we find to be attributable to at least four factors: C-F1 does not give credit for readability, it does not penalize repetition, it handles negated concepts incorrectly, and errors in the reference note propagate to C-F1 but not the LLM judge. We take these in turn. In Fig. 1a, we see two clusters of points with LLM judge score of around 2 that are separated by ~ 0.2 points of C-F1. The group on the left, with low C-F1, are notes that have no relationship to the dialogue but still receive high LLM judge readability scores. The group on the right are highly repetitive notes, which are judged poorly for conciseness but still receive high C-F1 (which is invariant to repetition). Removing the structure and conciseness judge dimensions tightens the fit on ID data, but not on OOD data (Fig. 1b).

A salient disagreement between the automated metrics and the judge is HW TalTech vs. ASLP. ASLP scores higher on judged coverage, both ID and OOD, yet lower on every automated metric, including Concept Recall, which is supposed to reward capturing more of the reference's concepts. The main cause is negation. ASLP documents extensive pertinent negatives ("Denies cough, chest pain, ..."), but the reference notes contain few negatives. In addition, the C-F1 matcher does not handle negation: it extracts "cough" and "chest pain" as *present* concepts. The result is that including negatives tends to reduce Concept Precision and F1, even if they are supported by the transcript. Negations account for roughly 57% of ASLP's precision gap and 35% of its F1 gap against TalTech. The LLM judge does not see the reference note and rewards supported negative statements with higher coverage scores. Thus, ASLP HW is a stronger system than C-F1 implies, due to peculiarities of the automated metrics.

VI. Conclusion

These results suggest that end-to-end LALMs can produce faithful SOAP notes from long clinical audio, and, more broadly, that training and reward-optimizing on synthetic audio can transfer to real recordings—offering a scalable route to long-context audio understanding that sidesteps the data-scarcity and privacy barriers of real clinical speech.

Several caveats temper this. The reference notes are a strong comparator but not a perfect ceiling. The best systems do not quite match their faithfulness (the one dimension on which the references remain unbeaten), yet several systems equal or exceed them on coverage, structure, or conciseness. The references are themselves imperfect and could be improved; for instance, they appear to under-document pertinent negatives relative to some submissions (§ V), so exceeding them on a given dimension does not certify a note as clinically ideal. More fundamentally, our synthetic audio does not yet reproduce the non-lexical cues (e.g., prosody, hesitation, audible symptoms) that motivate end-to-end modeling in the first place. These are challenging to simulate with current open-weight technology. On the metric side, no single automated score is fully trustworthy: C-F1 is cheap and reproducible but blind to negation, repetition, and readability, while the LLM judge captures these but brings its own biases and cost (and, as we observe, sensitivity to quantization and serving stack).

Finally, the Heavyweight track's headline affordances (retrieval, tools, and agents) went almost entirely unused. We read this less as a lack of interest than as a sign that audio-space processing tools remain far less mature than their text counterparts, and are hard to employ without crossing the no-transcription rule.

Acknowledgment

This work was supported by the 2025 Jelinek Memorial Summer Workshop on Speech and Language Technologies (JSALT 2025), hosted at the Brno University of Technology and organized by the Center for Language and Speech Processing (CLSP) at Johns Hopkins University. We acknowledge the resources of the Ohio Supercomputer Center [50].

AI-Generated Content Disclosure: AI tools were used to assist with language editing and grammar enhancement across the manuscript. The authors reviewed the final text and take full responsibility for it.

REFERENCES

[1] J. Xu, Z. Guo, J. He *et al.*, "Qwen2.5-Omni technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[2] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S. gil Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=FjByDpDVIO

[3] Amazon Artificial General Intelligence, "Amazon nova sonic: Technical report and model card," *Amazon Technical Reports*, 2025. [Online]. Available: https://www.amazon.science/publications/amazon-nova-sonic-technical-report-and-model-card

[4] Y. Li, J. Liu, T. Zhang *et al.*, "Baichuan-Omni-1.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2501.15368

[5] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[6] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy, S. Gandhi, S. Ghosh, S. Mishra, T. Foubert, A. Rastogi, A. Yang, A. Q. Jiang, A. Sablayrolles, A. Héliou, A. Martin, A. Agarwal, A. Roux, A. Darcet, A. Mensch, B. Bout, B. Rozière, B. D. Monicault, C. Bamford, C. Wallenwein, C. Renaudin, C. Lanfranchi, D. Dabert, D. S. Chaplot, D. Mizelle, D. de las Casas, E. Chane-Sane, E. Fugier, E. B. Hanna, G. Berrada, G. Delerce, G. Guinet, G. Novikov, G. Martin, H. Jaju, J. Ludziejewski, J. Rute, J.-H. Chabran, J. Chudnovsky, J. Studnia, J. Barmentlo, J. Amar, J. S. Roberts, J. Denize, K. Saxena, K. Yadav, K. Khandelwal, K. Jain, L. R. Lavaud, L. Blier, L. Zhao, L. Martin, L. Saulnier, L. Gao, M. Pellat, M. Guillaumin, M. Felardos, M. Dinot, M. Darrin, M. Augustin, M. Seznec, N. Gupta, N. Raghuraman, O. Duchenne, P. Wang, P. Saffer, P. Jacob, P. Wambergue, P. Kurylowicz, P. Chagniot, P. Stock, P. Agrawal, R. Delacourt, R. Sauvestre, R. Soletskyi, S. Vaze, S. Subramanian, S. Garg, S. Dalal, S. Gandhi, S. Aithal, S. Antoniak, T. L. Scao, T. Schueller, T. Lavril, T. Robert, T. Wang, T. Lacroix, T. Bewley, V. Nemychnikova, V. Paltz, V. Richard, W.-D. Li, W. Marshall, X. Zhang, Y. Wan, and Y. Tang, "Voxtral," 2025. [Online]. Available: https://arxiv.org/abs/2507.13264

[7] S. Arora, K.-W. Chang, C.-M. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H.-Y. Lee, K. Livescu, and S. Watanabe, "On the landscape of spoken language models: A comprehensive survey," *Transactions on Machine Learning Research*, 2025.

[8] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=xWu5qpDK6U

[9] O. Ahia, M. Bartelds, K. Ahuja *et al.*, "BLAB: Brutally long audio bench," 2025. [Online]. Available: https://arxiv.org/abs/2505.03054

[10] J. Chen, Z. Guo, J. Chun, P. Wang, A. Perrault, and M. Elsner, "Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance," in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

[11] S. Kumar, v. Sedláček, V. Lokegaonkar, F. López *et al.*, "MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 27, pp. 22 688–22 697, 2026. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/39430

[12] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[13] V. Rennard, G. Shang, J. Hunter, and M. Vazirgiannis, "Abstractive meeting summarization: A survey," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 861–884, 2023.

[14] S. Cornell, C. Boeddeker, T. Park, H. Huang, D. Raj, M. Wiesner, Y. Masuyama, X. Chang, Z.-Q. Wang, S. Squartini *et al.*, "Recent trends in distant conversational speech recognition: A review of chime-7 and 8 dasr challenges," *Computer Speech & Language*, vol. 97, p. 101901, 2025.

[15] A. Ben Abacha, W.-w. Yim, G. Adams, N. Snider, and M. Yetisgen, "Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 503–513. [Online]. Available: https://aclanthology.org/2023.clinicalnlp-1.52/

[16] S. A. Haider, S. Prabha, C. A. Gomez-Cabello, S. Borna, A. Genovese, M. Trabilsy, B. G. Collaco, N. G. Wood, S. Bagaria, C. Tao, and A. J. Forte, "Synthetic patient–physician conversations simulated by large language models: A multi-dimensional evaluation," *Sensors*, vol. 25, no. 14, 2025.

[17] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu, and H. Yu, "NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes," in *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024.

[18] A. Papadopoulos Korfiatis, F. Moramarco, R. Sarac, and A. Savkov, "PriMock57: A dataset of primary care mock consultations," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022.

[19] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. Yetisgen, "ACI-BENCH: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation," *Scientific Data*, vol. 10, no. 1, p. 586, 2023.

[20] F. Fareez, T. Parikh, C. Wavell, S. Shahab, M. Chevalier, S. Good, I. De Blasi, R. Rhouma, C. McMahon, J.-P. Lam, T. Lo, and C. W. Smith, "A dataset of simulated patient-physician medical interviews with a focus on respiratory cases," *Scientific Data*, vol. 9, no. 1, p. 313, jun 2022. [Online]. Available: https://doi.org/10.1038/s41597-022-01423-1

[21] Y. Labrak, D. Grünert, S. Baroudi, J. Chun, P. Cyrta, S. Burdisso, A. Hassoon, D. Liu, A. Rothschild, R. Van Deusen, P. Motlicek, A. Perrault, R. Marxer, and T. Schaaf, "Generating synthetic doctor-patient conversations for long-form audio summarization," in *Interspeech*, 2026.

[22] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. [Online]. Available: https://aclanthology.org/W19-5034/

[23] Google DeepMind, "Gemma 3 model card," 2025. [Online]. Available: https://ai.google.dev/gemma/docs/core/model_card_3

[24] K. Team, Y. Bai, Y. Bao *et al.*, "Kimi K2: Open agentic intelligence," 2025. [Online]. Available: https://arxiv.org/abs/2507.20534v1

[25] Google DeepMind, "Gemma 4 model card," 2026. [Online]. Available: https://ai.google.dev/gemma/docs/core/model_card_4

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[27] A. Yang, A. Li, B. Yang *et al.*, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[29] X. Shi, X. Wang, Z. Guo *et al.*, "Qwen3-ASR technical report," 2026. [Online]. Available: https://arxiv.org/abs/2601.21337

[30] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, Jul. 2000.

[31] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[32] Google Research, "`rouge-score`: A python implementation of rouge," 2019. [Online]. Available: https://github.com/google-research/google-research/tree/master/rouge

[33] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the*

29th Symposium on Operating Systems Principles, ser. SOSP '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 611–626. [Online]. Available: https://doi.org/10.1145/3600006.3613165

[34] A. Olev and T. Alumäe, "Robust summarization of doctor-patient conversations: Taltech systems for the beyond transcription challenge," 2026. [Online]. Available: https://arxiv.org/abs/2607.17230

[35] J. Ahn, G. Kim, and C. Kim, "Faithfulness has a price for end-to-end audio-grounded clinical note generation," 2026.

[36] ——, "Dissecting metric-aligned reinforcement learning for audiollm clinical note generation," 2026.

[37] K. Matsuura, T. Kano, T. Ashihara, and M. Delcroix, "A lightweight end-to-end doctor-patient conversation summarization system: The ntt submission to betrac 2026," 2026.

[38] Z. Zhang, M. Shao, W. Tian, T. Zuo, L. Li, and L. Xie, "Bridging the modality gap in long-form clinical audio: A comparative study of lightweight and heavyweight end-to-end soap generation," 2026. [Online]. Available: https://arxiv.org/abs/2609.14467

[39] Y. DKHISSI, V. VIELZEUF, A. CAUBRIERE, and J.-E. DARTOIS, "Jyva at betrac: Audio-based medical note generation with soap section-specific queries," 2026.

[40] E. Y. Ugan, F. Retkowski, Y. Ko, T.-B. Nguyen, M. Züfle, J. Niehues, and A. Waibel, "Speech-to-soap: End-to-end summarization of medical dialogues: Kit@betrac 2026," 2026. [Online]. Available: https://arxiv.org/abs/2608.24327

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, vol. abs/2106.09685, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:235458009

[42] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized llms," in *Neural Information Processing Systems (NeurIPS)*, vol. abs/2305.14314, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258841328

[43] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[44] C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, J. Zhou, and J. Lin, "Group sequence policy optimization," *ArXiv*, vol. abs/2507.18071, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280017753

[45] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, W. Dai, Y. Song, X. Wei, H. Zhou, J. Liu, W. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y.-X. Wu, and M. Wang, "DAPO: An open-source llm reinforcement learning system at scale," in *Neural Information Processing Systems (NeurIPS)*, vol. abs/2503.14476, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277104124

[46] Z. Shao, P. Wang, Q. Zhu, R. Xu, J.-M. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," *ArXiv*, vol. abs/2402.03300, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267412607

[47] A. Ben Abacha, W.-w. Yim, Y. Fan, and T. Lin, "An empirical study of clinical note generation from doctor-patient encounters," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2291–2302. [Online]. Available: https://aclanthology.org/2023.eacl-main.168/

[48] J. Han, F. Landini, J. Rohdin, A. Silnova, M. Diez, and L. Burget, "Leveraging self-supervised learning for speaker diarization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[49] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 7721–7735. [Online]. Available: https://proceedings.mlr.press/v139/miller21b.html

[50] Ohio Supercomputer Center, "Ohio supercomputer center," 1987. [Online]. Available: http://osc.edu/ark:/19495/f5s1ph73