# ASR Science - 2023 Q4 QBR

Page 0: Metrics

## 

* ASR Science Metrics - 2023 Q4 QBR ([ASR Science Metrics - 2023 Q4 QBR](https://quip-amazon.com/tplYAfTGDfuw))

Q4 & EOY Highlights

**en-US Large Model**: We successfully deployed to production on 11/20, our first all neural large model (1B transducer, 1B rescorer) ASR, achieving a milestone of unifying all profiles and supporting conversational (non-Alexa speech). This model was also used to support a public demo for Remarkable Alexa on 9/20 and was ready for their initial prod launch on 9/29, but later postponed to 12/15. Starting on 1/18, this model was deployed to EFD en-US traffic on a new runtime GPU fleet in a new datacenter (CMH), reaching 65% by 1/31, and will reach 100% later before the end of February. The 11/20 deployment supported 100% of mobile shopping’s Nile (a.k.a. AoAA) traffic - achieving a 31% relative tail slot-WER reduction (from 13.7% to 9.5%) compared to the small AoAA model. Overall we achieved a 7.2% WER on the LetsChat conversational beta evaluation dataset (goal <8% WER), a 19.1% relative improvement vs. the small production model (8.9% WER). We also achieved an early cutoff rate of 1.2% (goal <2%) as measured against a conversational speech evaluation dataset, vs. 3.6% for our small production model, completing the DSLT goal [#626558](https://kingpin.amazon.com/#/items/626558). We further improved our hierarchical attention neural biasing (NB) and gazetteer rescoring, on top of the accuracy gains in Q3 (5% to 42% relative slot WER compared to small prod en-US model across 8 slots). We added NB support for tail skills, achieving 14% relative WER reduction on skill\_targeted\_test testset (from 14.7% for the November release to 12.6% for the new model supporting skills). For contact names, we matched the accuracy of the smaller en-US which includes NB and FSTs (4.8% slot WER). We introduced context-literal based zero-shot NB for on-screen context, which, combined with rescoring, brings 61% rel. WER reduction (from 12.0% to 4.7%) on the Trending Entities test set and 61% rel. WER reduction (from 8.7% to 3.4%) on the Questions of the Day, in offline evaluations compared to the baseline model without on-screen hints, achieving the accuracy expectation for on-screen hints VP goal [#580803](https://kingpin.amazon.com/#/items/580803). We also introduced single-turn session context in the first production rollout on 12/5, which achieved 3% relative WER reduction (from 7.1% to 6.9%) on Let’s Chat traffic as well as 4% relative reduction (7.8% to 7.5%) on a verbal hints test set, and later increased that to 5 turns in a subsequent January rollout. The large model’s hot-fixing now uses [adapters (LoRA)](https://mcm.amazon.com/cms/MCM-90464209) in the second pass, achieving 15% relative WER-S reduction and surpassing the effectiveness of hot-fixing in the small models, which completes VP goal[#586020](https://kingpin.amazon.com/#/items/586020).

**Speech-to-speech (S2S) Demo/Proof of Concept:** On 2/1, we demonstrated capabilities of the proof-of-concept 7B S2S model, including 9 sub-tasks: automatic speech recognition (ASR), text-to-speech (TTS) generation, emotion recognition, speaker attributed ASR, speech-to-API, speech-to-speech via API call, text-chat bot and speech-to-speech chatbot, and speech-to-speech translation, all performed via a single LLM fusing speech and text modalities with our large acoustic foundation encoder. In 2024 we will leverage the capabilities in the Ermis LLM to deliver offline applications, such as translation and localization of Wondery podcasts (VP [#762878)](https://kingpin.amazon.com/#/items/762878), Audible books (VP #[763033](https://kingpin.amazon.com/)), dubbing (voice porting and semantic translation) of Prime Video content (VP [#763226)](https://kingpin.amazon.com/#/items/763226), audio capabilities for self-service ad generation (VP #[761029](https://kingpin.amazon.com/)), and online conversational AI for Alexa (VP [#761872](https://kingpin.amazon.com/#/items/761872)) and customer service call containment (VP [#762969](https://kingpin.amazon.com/#/items/762969)), and chat bot functionality through Lex (Proposed).

**EP Arbitrator rollout:** We launched the new end-pointing model (“EP arbitrator”) in eleven locales completing the AI goal [#580946](https://kingpin.amazon.com/#/items/580946): en-US/CA on 4/18 (yielding 20bps CPDR reduction), hi-IN on 8/22 (yielding 25bps CPDR reduction), en-GB on 9/22 (yielding 23bps CPDR reduction), ja-JP on 10/5 (yielding 26bps CPDR reduction), es-MX/ES/US on 11/16, en-AU on 11/21, en-IN on 10/31 (yielding 43bps CPDR reduction), de-DE on 12/7. For en-US, the achieved CPDR improvement corresponding to a 29.7% relative reduction in early EP rate on the four target domains (Home Automation, Entertainment, Info and Shopping) measured on partial utterance patterns (by the Macaw model). For other locales, we do not have a reliable mechanism to measure early EP rate, hence, are only reporting the above CPDR improvements for this final goal update.

**Auto OEM Deliverables:** We completed 2023 delivery of LVC S2I models for BMW and Stellantis programs and currently working on improving S2I model accuracy for 3P skills and resolving bugs.  For BMW, we delivered updated S2I models for 14 locales in December 2023 (en-AU, en-CA, en-GB, en-IN, en-US, es-ES, es-MX, es-US, fr-CA, fr-FR, fr-ROW, hi-IN, nl-NL, pt-BR), and 2 more on 1/9/2024 (de-DE, de-ROW). These models yielded less than 10% WER on 3P test sets. For OEM Stellantis (STLA), we delivered en-US model with Brandon wakeword (VP goals: BMW [#585170](https://kingpin.amazon.com/#/items/585170), Stellantis [#585174](https://kingpin.amazon.com/#/items/585174)).  And for OEM Mahindra and Mahindra, we delivered en-IN & hi-IN S2I and cloud models with Hey Inglo CWW.

**On-Device Intent Execution (ODIE):**We launched ODIE on 12 devices by 11/20, completing the DSLT joint goal [#573160](https://kingpin.amazon.com/#/items/573160). We achieved both of the goal success criteria: (a) TPR (true positive rate) of 70% - TPR represents the rate at which the model correctly recognizes a match of an utterance with a supported intent; (b) UPL below 750ms at TM95 for locally executed utterances (measured at 732ms). The model achieves a match accuracy of 95.7% (this is the ratio of correct matches on supported intents) and a 2.7% false acceptance rate.

**Reduced hallucinations in production:** In September, we initiated a dedicated program to systematically address ASR hallucination errors negatively impacting customer experience and Alexa domain metrics. In January 2024, we concluded the program with the successful completion of the hallucination reduction goal across de-DE, en-GB, and en-US, completing the VP goal [#735060](https://kingpin.amazon.com/#/items/735060). Compared to the peak hallucinations in 2023, we reduced hallucinations by 66% for en-US (v58), 22% for en-GB (v65), and 78% for de-DE (v74). We achieved these reductions by fine-tuning the core transducer on machine-transcribed unintelligible speech and negative training examples, and establishing hallucination metrics to prevent future regressions. Further, we developed a muting mechanism based on the device-directedness feature, which we launched with the en-US large v2.6 model on 1/25.

**Interactive self-learning:** We built an en-US teacher system that learns to improve itself from observing user and system interactions, completing the AI goal [#585254](https://kingpin.amazon.com/#/items/585254) by 12/5. We trained a 200M teacher transducer model using causal and non-causal context, for both audio and text modalities, improving WER by 10.1% relative compared to the baseline teacher trained with no context. In addition, we trained a context aware 1B en-US teacher model using causal and non-causal context, for both audio and text modalities, improving WER by 10.4% relative compared to the 1B large production model using causal text context only.

Q4 Lowlights

**ADS Gold transcription quality:** We did not complete seven VP goals because we have low confidence in the quality of the existing human transcriptions and metrics derived from these transcriptions, including WER-based goal measurements. On 12/1, ADS informed ASR that ADS gold transcription workflows in en-US, en-CA, en-GB, en-AU, en-IN, hi-IN and ar-SA were impacted by remote DAs not following the correct process from January through October of 2023. With no extra ADS budget nor time to re-transcribe test sets used to measure our goals, we cancelled three VP goals ([#580809](https://kingpin.amazon.com/#/items/580809), [#635862](https://kingpin.amazon.com/#/items/635862), [#660270](https://kingpin.amazon.com/#/items/660270)) because we reached the accuracy goal at 80%, or better, and marked four goals as *did not meet* ([#586518](https://kingpin.amazon.com/#/items/586518), [#586526](https://kingpin.amazon.com/#/items/586526), [#580808](https://kingpin.amazon.com/#/items/580808), [#585179](https://kingpin.amazon.com/#/items/585179)). ADS corrected the process by 1/31. Upcoming improvements include: roll out of TeacherGold transcription workflow reducing cost of transcription by an estimated 25% on dev/test queues (ETA: 2/26) and disabling DVR on 1/26.

**Poor run-time accuracy for custom wake word (CWW):** BMW winter testing was impacted due to poor CWW performance, root-caused to the on-device wake word model recognizing CWW without prefixes, such as “hey/hello/okay”. Our analysis indicated that CWW performance can be improved by changing the pre-roll to 300ms, however, it will impact latency of Alexa traffic. Based on this insight, we worked with the WW team and reduced the wakeword start offset as a stopgap. This enabled BMW to test several CWW and phrases across en-US, en-GB, de-DE (except in pt-BR, where issues persisted). Due to Data Processor (DP) compliance designed by Auto, we have challenges with accessing the cloud reject utterances which further delayed the analysis. As a long-term fix, we are working with [ASR Engine](https://issues.labcollab.net/browse/ASE-3314.) to enable WW-dependent pre-roll selection.

**Affinity-based music playback failure (COE**[**#285360**](https://www.coe.a2z.com/coe/285360/content)**):** On 10/22, an Alexa customer had poor CX when asking Alexa to “play the album Big Thank by King Youngblood”. They had to repeat the request three times before they got the desired response. It was particularly frustrating that Alexa did not understand this request, considering that this customer asks to play this music regularly. This incident showed we need continued investment in personalization of Music and other entities in ASR (with high quality customer affinity catalogs and ASR model’s responsiveness to personalized content) and in robustness of the error recovery systems downstream of ASR (with personalized query rewrites and catalog fuzzy matching).

**es-US ZIGGY wake word rejection degradation (COE**[**#288773**](https://www.coe.a2z.com/coe/288773/content)**):** The es-US ZIGGY wake word regressed after transitioning es-US traffic to the es-\* unified large model. This resulted in a reject rate of 54% for ZIGGY initiated wake word utterances compared to the previous es-US model reject rate of 4%. To mitigate the customer impact, we re-routed es-US Ziggy initiated traffic to the non-unified es-US model on 12/4 through 01/09. We root-caused the regression to the application of an existing BlueRec hot-fix (transform ‘ziggy’ => ‘si’) originally configured to only apply for es-ES and es-MX. Since the unified es-\* model is a repackaged es-ES model, any hot-fix previously applied to es-ES will now apply to all es-\* locales/profiles. We will inspect the BlueRec's hot-fix mechanism before launching the single multi-lingual large model package.

Hot Topics

**Impact due to P4 shortage:** AGI Sensory requested 200 P4s in H1 (with a plan to pivot to Trn1 in H2), to run experiments and conduct ablation studies in order to begin pre-training of the 30B S2S model by 3/1. This is needed for delivering key milestones for non-streaming S2T, T2S, and S2S translation by June 2024, and streaming speech processing in H2 2024 (Project Ermis [goal](https://kingpin.amazon.com/#/items/761079)). On 11/29, we were informed that AGI does not hold this much capacity in ARN. If we stay in ARN and resourcing continues to be limited to 48 P4s for pre-training and 12 for fine tuning, and by leveraging an additional 15 when not needed for large en-US or large multi-lingual models, we estimate an increase from 20 to 80 days to train each 30B candidate. This will have a cascading effect on all S2S milestones and push the 30B S2S delivery into 2025, turning above 2024 goal to RED. In H2 2024, when feature gaps between P4 GPUs and Trn1 are closed (support for conformer training), we will transition to use 810 Trn1’s (requested on 11/23) to train a 100B S2S model and eliminate the dependency on P4s. Without additional P4 capacity, the 100B or larger pre-trainings become effectively impossible.  When capacity becomes available, we have a ~4-week lead time to move 16.2 PB of data for S2S and large ASR model pretraining and finetuning.

Program updates

**Efficient bootstrapping of ASR for new languages (BOOM/Glossa):** We developed a framework to efficiently bootstrap ASR for new languages, where we: (a) distilled knowledge from a large 1B multilingual Teacher ASR (BOOM) model trained on Alexa and non-Alexa data (SayHi, AWS, public corpora); (b) augmented synthetic TTS data to improve coverage for common utterances (machine translated from high resource locales);  (c) fine-tuned with about 150 hours of in-domain 3P data. Using this framework, we developed ASR to support Auto and FireTV use-cases in 3 new languages, where we achieved less than 10% WER targets for Korean (Auto 3.0%), Polish (Auto 4.5%, FireTV 4.8%) and Turkish (Auto 9.4%, FireTV 4.7%) on 3P test sets. These experiments confirmed a 4x reduction in labeled training data (from 600hrs to 150hrs) and 6x reduction in science bandwidth (from 18HC months to 3HC months) for bootstrapping ASR for new languages, and paved way for rapid future language expansion.

**Faithful Speech Generation:** On 10/11 and 10/25 we measured faithfulness of synthetic audio data for ja-JP and nl-NL. This completed the VP goal [#579601](https://kingpin.amazon.com/#/items/579601) to deliver synthetic speech generation models that can deliver near production-quality speech in all Alexa locales by 12/15. Ten locales met the 2% faithfulness bar (WER difference <2%, RMSE <2%): en-GB, en-IN, de-DE, pt-BR, en-AU, en-CA, fr-CA, ja-JP, nl-NL and es-MX, with the remaining locales meeting the 5% bar (WER difference <5%, RMSE <5%): en-US, es-US, es-ES, ar-SA, it-IT, hi-IN and fr-FR. The adoption of the Freese TTS models in DOME is scheduled for Q1/2024 and will close the existing language gap for launched locales. Further, this will be enabled for nl-NL prior to the launch. Shopping and Automotive teams are exploring the Freese models for training and evaluation data generation.

**Everlearn:** In Q4, we achieved the goals [586020](https://kingpin.amazon.com/#/items/586020) (28% more hot-fixing success against the goal of 20%), [675088](https://kingpin.amazon.com/#/items/675088) (teacher model IL onboarding), [735060](https://kingpin.amazon.com/#/items/735060) (50% hallucination reduction), and [579601](https://kingpin.amazon.com/#/items/579601) (faithful speech generation, all locale). All planned Everlearn technologies such as data generation automation, data incorporation into model builds and updates, incremental update, and from-scratch build automation has been successfully rolled out in 2023. By including DOME generated data in model training, we completed late (by 1/31/24) the goals [579602](https://kingpin.amazon.com/#/items/579602) (Everlearn launch), [578722](https://kingpin.amazon.com/#/items/578722) (DOME effectiveness), and [586962](https://kingpin.amazon.com/#/items/586962) (Trending WERR). Due to ADS Gold transcription quality issues (confirmed on 12/6/2023), we canceled the goal [635862](https://kingpin.amazon.com/#/items/635862) (10% Tail WERR) as we measured only 8% relative WER reduction (which is an underestimate due to quality issues).

**FTP:** In Q4, we completed the necessary implementation and onboarded to PyRama for transducer training, leading to 2.5x throughput improvement over Phasa. We utilized this throughput improvement to train a multilingual 1B transducer leading to the completion of the VP goal [#732722](https://kingpin.amazon.com/#/items/732722) (3% - 30% relative WER reduction in Phasa evaluation on 13 out of 16 locales) despite compute and time constraints. We also doubled down on procuring, storing, managing and preparing non-Alexa 1P/3P datasets by working with legal and related 1P teams. Having resolved non-Alexa data storage issues with DPES managed s3 buckets, we prepared ~2M hours of non-Alexa (Voxpopuli, MLS, Common Voice, SayHi etc.) unsupervised data across [100+ languages](https://quip-amazon.com/8xwPAsEOKLop/Sahasra-2B-general-ASR-covering-50-languages). This dataset was used to train a 2B streaming foundation model (completed 02/08). We further focused on preparing supervised non-Alexa datasets for 50 languages for training a 2B 50-lingual transducer (planned for Q1 2024). We identified challenges with labeled transcripts for PrimeVideo, and we initiated 1B teacher-based SSL transcript generation for multiple 1P/3P data sources (SayHi, VoxPopuli, MLS, LibriVox etc). We are also researching data presentation, language grouping and architecture changes (prediction network, efficient vocabulary, adapters etc) with 200M transducers that will inform the strategy for training the 2B 50-lingual transducer in Q1.

**Context Modeling:** In Q4, we delivered updated contextualization technology to large en-US ASR models, with improvements in personalization and session context performance. We launched session context, consisting previous turn ASR hypothesis and TTS response. We launched an updated version of personalization support (neural biasing and gazetteer rescoring, the latter as part of Runtime program), which added support for skill biasing and maintained the slot gains we saw in Q3 and closed the gap from the small en-US ASR model regarding contact names. We launched biasing for on-screen hints and on-screen selections. Accuracy improvements from these technology features are provided in the highlights section. In Q4, in collaboration with ASE, we launched Context engine, a component within Pryon that produces text embeddings dynamically during runtime, a necessary component in support of screen context biasing and session context. In Q4 we [kicked-off a new phase for NB tech](https://quip-amazon.com/bfGNAaMoUQtJ/Neural-Biasing-Next-Phase-Kick-off-10-18-2023), for integrating NB and UCE within the Pyrama framework. The need for this technology resets stemmed from: (a) the need to urgently resource and prioritize Pyrama migration across programs; (b) experimental inefficiencies on training NB with embeddings from UCE, since UCE is in Pyrama while NB in Phasa, which were slowing down our team’s efforts; (c) the need to better align NB technology for large ASR models with S2S, which is better served by pivoting UCE from BERT to a speech-text LLM architecture. This new phase was communicated by an offsite among key stakeholders of the Context modeling program, to ensure alignment in the new direction and the corresponding technology roadmap. In Q4, we ran a 2-week hackathon to advance the Pyrama transition for Context Modeling; we are currently on track to meet the goal of being ready to build models from Pyrama by end of March 2024.

**HW Acceleration/Neural efficiency:** As part of the HW Acceleration program roadmap and aiming to reduce runtime costs, we concentrated on 8-bit and 4-bit weight-only quantization for the 1B and 2B Conformer models, aiming production for 4-bit quantization for 2B conformer in Q2/2024. Benchmarking with NemoRT custom inference engine, we observed 40-50% latency reduction for 4-bit quantization vs. 8-bit quantization at kernel level. The WERs are also promising: out of 30 testsets, we found degradation in only two of them, and that was below 2% relative. The 4-bit quantization will also support 7B and 30B S2S models. We also completed the N:M sparsification feature and validated the end-to-end accuracy for the 1B Conformer. The NeMoRT-bench results demonstrate a 6-10% runtime speed-up under the batch size of 1. When measured in Alexa\_prize\_onlygoldens and lets-chat-alpha, we observed less than 0.5% relative WER degradation when enabling bf16+sparsity. We were unable to deploy this feature to production since summer 2023 due to bugs in Nvidia TensorRT software which they acknowledged but were unable to resolve so far. Since then, we pivoted our focus to 8- and 4-bit quantization.

#### Runtime / Multi-Task:

* **Beat Whisper on SpeechStew:** In Dec’23, using a preliminary version of the v3 large model trained with more diverse non-Alexa data sets (Common Voice, Vox Populi, Wiki TTS data, Prime Video), we observed a 52% relative reduction in WER between v3 and v2 (from 16.3% to 7.9%) on CommonVoice subset of SpeechStew, beating Whisper by 12% relative (7.9% vs 9%).  Across all 7 SpeechStew subsets, we reduced WER by 27% relative (from 17.1% to 12.4%), compared to the v2 large model, narrowing the gap to Whisper from 53% to 11% relative.  In Jan’24, we further optimized the v3 large model and updated 1B RescoreBERT with SpeechStew training data to outperform Whisper (11.3% vs 11.4%) and completed the VP goal [732701](https://kingpin.amazon.com/#/items/732701).
* **Core Transducer Optimizations:** In 2023 we migrated all (cloud) ASR models to conformer architecture for all locales and unified the model across profiles for 11 out of 15 locales, reducing the number of different models from up to 38 to 18. Apart from this reduction, we also achieved WER reductions ranging from 2% to 11% rel. across locales/profiles. These unification investments into modelling serve as the foundation for our 2024 plans to further unify across languages and deploy a single multi-lingual package that serves all Alexa traffic. We are also feature complete for non-causal [amortized networks (VNAC) to enable VAD-based adaptive chunking](https://wiki.labcollab.net/confluence/display/Doppler/Non-causal+Amortized+Networks) expected to deploy on the Large en\_US v3.0. The latest experiments show rel. WER reductions of 11% on long-form, 11% on FTV, and 14% on AMI.
* **Large Rescorer:**We launched a unified 1B ContextBERT as part of the en\_US Large Model that yielded accuracy gains across profiles (**EFD**: 9.1% rel. WERR,  6.1% rel. WER-S reduction ; **FTV**: 6% rel. WERR ; **AoAA**: 7% rel. WERR), and are now feature complete for our 1B Multi-lingual ContextBERT(02/10). In support of these and future deployments, we also developed novel technologies such as: Gazetteer Neural Biasing, MWER-based Distillation, Scaling Laws of Rescorers, Parameter-efficient training for RescoreBERT, RescoreGPT, and LLM in Generative Mode for ASR.
* **Multi-Task**: We completed the EP arbitrator deployment for small ASR models (refer to highlight above). We are on track to complete the multi-task DD model by 2/15. We developed DD muting to reduce hallucinations and launched the feature with the en-US v2.6 large model on 1/25.
* **Package unification:**We completed the deployment of unified neural transducer models, unified neural re-scorer models, and a unified model package for es\* locales/profiles on 11/16, reducing the number of Spanish language model packages from six to one. We automated packaging, benchmarking, and deployment of the es\* unified model package. Completing this developmetn enables the 2024 goal to deploy a single multi-lingual large model package for all locales and device profiles.

**ODIE:** During Q4, we further developed the [Embedding Space Transformation Network (ESTN)](https://quip-amazon.com/sgpVAocJV1Vm), a method that transforms embeddings from one model’s latent space to another. In every new model deployment, the embedding space changes, and without ESTN this mismatch invalidates the (adapted) acoustic embeddings in the cache. These no longer fit to the new acoustic encoder and have to be removed, requiring each time a new embedding adaptation period (~3 weeks) until the cache is fully updated. With ESTN we addressed this issue, as we were able to match the performance of native embeddings with minimal (below 3% relative) TPR reduction. ESTN is scheduled to be in production with R63 (Q2 2024). In Q4, we also started the migration of the training code from Phasa to PyRama and expect to complete this work by the end of Q1 2024.

Tooling Updates

We surveyed 100+ scientists to identify productivity blockers for ASR model building process. We identified 67 issues, with the top two: 1) inability to copy-paste in the Sensai Studio, Notebook and Terminal; 2) inability to access model training logs instantly.

The Sensai November update started restricting copy-paste out of the browser, negatively impacting user productivity ([D104386702](https://t.corp.amazon.com/D104386702)). We partnered with the Sensai team to support science developer use cases but had to escalate the situation in January when further restrictions were released. The Sensai team filed an exception request approved on 1/29. Copy-paste functionality in Sensai was reenabled on 1/30 and was perceived very positively by scientists across AGI.

Regarding instant access to model training logs in the Sensai browser, the team reports limited read access and search functionality support. The team requests that leadership reviews a cost-benefit analysis of engineering design decisions before proceeding to implementation. In the example of removing the ability to copy-paste a particular utterance ID, the privacy risk reduction came at a cost of a large negative impact to the team productivity.

By 2/28, science and engineering will work together to identify high-impact low-effort issues, and identify plans to fix them using resourcing from specific programs.

### ---   You may stop reading further   ---

Appendix

## Product updates

#### en-US/CA

* We deployed a Conformer yielding 18bps CPDR reduction. We also improved the model performance for top gaming skills (e.g., Volley) and Disney, yielding up to 24% relative WER improvement.
* Completed the en-US and en-CA hybrid model deprecation working with Alexa Skills team and Disney to transition skill traffic to RNN-T contributing to $170k in 2023 and deprecated 50 legacy and hybrid models across all locales contributing to 1M+ USD cost saving in 2023.
* en-US large model transition
  + We conducted 10 A/B experiments on en-US EFD profile, reducing 50-90% hallucinations across domains and improved overall CPDR by 15.06% relative (from 13.34% to 11.33%) and friction by 8.9% (from 4.60% to 4.19%) between v0.1 and v2.3 models. But, the large model shows 17-23 bps regression when compared to the production small model mainly attributed to comms (153 bps) and shopping (40 bps) domains. We will finish the transition by 2/29.
  + We conducted 7 A/B experiments on FireTV profile, improving overall accuracy and hallucinations on the large model. This contributed to a 3.2% relative CPDR improvement (from 24.89% to 24.09%) between v1.2 and v2.3 models, outperforming the production small model by 112 bps and friction by 17 bps. We will transition FireTV traffic to the large model by 2/16.
  + We conducted 7 A/B experiments on skills traffic, improving overall accuracy on the large model. This contributed to 22.78% relative CPDR improvement (from 14.84% to 11.46%) and 72.64% relative friction improvement (from 2.96% to 0.81%) when comparing v2.3 with the production small model. We will transition skills traffic to large model by 2/29.

#### ja-JP

* We deployed a profile-unified Conformer yielding 13% rel. CERR and 110bps CPDR reduction. We achieved another 80bps CPDR reduction with EP operating point optimization.

#### es-MX/US/ES

* We deployed a unified single package to serve all es-\* traffic with no accuracy degradation on Doppler and  40-60bps improvements in DD-NWWER for FTV.  We also integrated DOME TTS data yielding >50% rel. WER reduction on DOME test sets.

#### star-IN

* We on-boarded to incremental learning (IL) for en-IN and hi-IN enabling KTLO for small models in 2024. We added DOME-TTS ingestion into transducer-IL for en-IN. We also improved transducer training obtaining 2% average WER reduction for en-IN and hi-IN (YoY 20.4% for en-IN and 13.6% for hi-IN).
* We deployed [EP arbitrator](https://wiki.labcollab.net/confluence/display/Doppler/EP+Arbitrator+for+Streaming+Conformer+models), that allows natural pauses and fillers in speech, for en-IN EFD and achieved 43bps CPDR reduction ([Frida AB test](https://frida.alexa.amazon.dev/experiment/40827001-4690-4065-b645-8c24fbe61dc9)).

#### en-GB/AU

* en-GB was onboarded to e2e IL pipeline, to help enable KTLO for small models in 2024

#### ODIE

* We launched ODIE on 12 devices as of 11/20 for en-US. This model with local cache has a TPR (true positive rate) of 70%. For these locally executed utterances, we observed a CPDR of 2.3% (vs ~6% in cloud) and friction of 1.5% (vs 3.3% in cloud) for the supported intents.

## Goals

Below table gives the list of goals we met/completed, completed late, cancelled and DNM.  For more details, sharing the kingpin dashboard link - <https://kingpin.amazon.com/#/businessReview/reports/38bdfc9a-1bf1-4a6f-a5ac-55dff42257c7>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No | **First Level Category** | **ID** | **Status** | **Title** |
| 1 | VP Goals | 585358 | Completed | Reduce ASR service cost from 1.82 cents to 1.27 cents per thousand utterances (30% improvements). |
| 2 | DSLT | 585383 | Completed | Achieve best-in-class voice search speech recognition accuracy for popular and trending requests on FireTV in en-US, en-IN, and de-DE. |
| 3 | VP Goals | 579601 | Completed | Launch synthetic speech generation models that can deliver near production-quality speech in all Alexa locales |
| 4 | DSLT | 585289 | Completed | Achieve best in class voice search speech recognition accuracy for Alexa on Amazon app (AoAA) for en-IN and en-US. |
| 5 | DSLT | 626558 | Completed | Deliver a new streaming ASR system for LLM-powered free-form conversational experiences. |
| 6 | DSLT | 573160 | Completed | Make Alexa the fastest smart home assistant by delivering <750ms response times through edge-based innovation on 12 devices in en-US. |
| 7 | VP Goals | 585170 | Completed | Enable BMW contractual deliverables by launching 1) Local Voice Control signal-to-intent models for \*\*\*Car control, Music control, Communications, Local Information, and Navigation domains and 2) data processing compliant model training pipeline for BMW |
| 8 | VP Goals | 585483 | Completed | Reduce tail WER-slot for Shopping ItemName by 20% relative based on unweighted average for en-US, en\_IN, and en\_GB. |
| 9 | VP Goals | 586020 | Completed | Improve the percentage of misrecognized entities corrected by ASR hot-fix pipeline by 20% for Information, Entertainment and Shopping domains within 24h for en\_US large model |
| 10 | Alexa AI | 580946 | Completed | Reduce early cutoffs by 20% relative across 4 domains in at least 10 locales |
| 11 | DSLT | 585288 | Completed | Deliver 3 new languages for Auto and FireTV by developing massively multilingual Spoken Language Understanding (SLU) Teacher models. |
| 12 | VP Goals | 585399 | Completed | Reduce Sentence Error Rate (SER) by 10% on Info Q&A traffic WW |
| 13 | VP Goals | 585174 | Completed | Enable Stellantis contractual deliverables by launching Local Voice Control on the Vega OS System |
| 14 | Alexa AI | 585254 | Completed | Interactive Self-Learning AI: Build an ASR model which learns to continuously improve itself by observing the interactions with its environment. |
| 15 | VP Goals | 712179 | Completed | Reduce Info slot-SER in Local Info for PlaceName by 10% relatively in en-GB, de-DE and 5% relatively in en-IN, es-US and it-IT locales |
| 16 | Alexa AI | 574584 | Completed | Improve the accuracy (IRER) for Navigation and Local Search domains, and decrease the footprint of the Signal-to-Interpretation (S2I) model for Edge Auto in enUS, enCA, frFR and frCA |
| 17 | VP Goals | 641649 | Completed | Improve the accuracy (IRER) for Navigation and Local Search intents of the Signal-to-Interpretation (S2I) model for Edge Auto in en-US, en-CA, fr-FR and fr-CA |
| 18 | VP Goals | 580810 | Completed | Reduce AWS cost for personalization model builder service by 57% relative from projected $15.6M in to $6.7M by EOY 2023 |
| 19 | VP Goals | 616943 | Completed | Hold the line on ASR Delivery Independent latency for all intentful utterances to not exceed Week 50'22 value of 1025 ms at TM [95%:99]. |
| 20 | VP Goals | 615354 | Completed | Hold the line on ASR Delivery Independent latency for all intentful utterances at Trimmed Mean 95 to not exceed Week 50'22 value |
| 21 | VP Goals | 584067 | Completed | Deploy unified model packages across profiles (EFD, FireTV, AoAA) for all locales |
| 22 | VP Goals | 615494 | Completed | [COE Cascaded VP-Goal for shehzad@] Publish HSE and Sev1 COEs within 14 days (P50) from incident date and complete High Priority COE Action Items within 28 days (P90) from incident date for Alexa Org under shehzad@ |
| 23 | VP Goals | 675088 | Completed | Keep teacher ahead of runtime-model in incremental updates for en-US, en-GB & de-DE. |
| 24 | VP Goals | 711154 | Completed | Deliver a new streaming ASR system for LLM-powered free-form conversational and voice search experiences on Alexa on Amazon App in en\_US |
| 25 | VP Goals | 732722 | Completed | Build a large 1B multilingual ASR transducer model that achieves WER on par or better than the current “small” production models in TK locales |
| 26 | VP Goals | 579602 | Completed | Launch of Everlearn Infrastructure WW (ex-Continuous Learning): Continuous and automated incremental training and weekly release of teacher, student, cloud and on-device models for all locales |
| 27 | DSLT | 735060 | Completed | Reduce ASR hallucination errors by 50% for en-US, en-GB, and de-DE |
| 28 | VP Goals | 586962 | Completed Late | Reduce trending WER by 15% relative for en-US, en-GB, de-DE |
| 29 | VP Goals | 578722 | Completed Late | Increase DOME effectiveness: Reduce word error rate to below 10% for at least 60% of must-win domain ingestions across locales |
| 30 | VP Goals | 732701 | Completed Late | Achieve state-of-the-art WER on “SpeechStew” public test set using a large en-US model |
| 31 | VP Goals | 732775 | Completed Late | Demonstrate the NextGen ASR architecture, showcasing state-of-the-art transcription accuracy, multi-task functionality (e.g. emotion tagging, speaker diarization), and speech-to-speech generation capability. |
| 32 | VP Goals | 584636 | Cancelled | Reduce world-wide tm95 ASR Delivery Independent Latency by 10% relative from 369ms to 332ms |
| 33 | VP Goals | 588476 | Cancelled | Reduce ASR contribution to UPL for Top Customer NVA intents by 12% as compared against Week 50 of 2022 (w/s 12/11/22) |
| 34 | VP Goals | 588207 | Cancelled | Reduce ESER estimation error by 15% on device-directed tail utterances and ensure customer facing confidence score stability between model releases |
| 35 | VP Goals | 702177 | Cancelled | Achieve 15% relative slot-SER reduction for table reservation, sports ticketing and concert ticketing slots to support Aloco monetization experience launch in en\_US by 12/30. |
| 36 | VP Goals | 586299 | Cancelled | Achieve 10% relative better performance (e.g. WER, slot-WER) for deployed teacher compared to the runtime model for trending, long tail, multi-turn, and personalized slots for en-US,  en-GB, and de-DE. |
| 37 | VP Goals | 586290 | Cancelled | Build 1B+ parameter multilingual acoustic foundation and teacher models, achieving a reduction in WER by 20% relative in underrepresented scenarios and teacher device-directedness EER by 20% |
| 38 | VP Goals | 635862 | Cancelled | Reduce Tail WER by 10% world-wide |
| 39 | VP Goals | 586518 | Cancelled | Reduce the FireTV sentence error rate (SER) by 10% relative for voice search in en\_US/IN/GB, es\_US, pt\_BR and de\_DE. |
| 40 | VP Goals | 660270 | Cancelled | Reduce WER in India by 20% |
| 41 | VP Goals | 580809 | Cancelled | Achieve 5% WER absolute for verbal hints in en-US |
| 42 | Alexa AI | 585988 | DNM | Standardize string representation in multi-script locales (ja-JP, en-IN) via mandatory migration to consistent Alexa-wide text normalization and reduce CPDR in ja-JP and en-IN by 10% relative |
| 43 | VP Goals | 585088 | DNM | Launch Pryon-2 M1 engine to ASR LaserService |
| 44 | VP Goals | 586505 | DNM | Deploy 200ms post-roll for FireTV for all locales and reduce average SER by 10% relative |
| 45 | VP Goals | 585179 | DNM | Reduce Points of Interests (POI) and street address slot WER by 12% YoY in en-US and en-GB. |
| 46 | Alexa AI | 584045 | DNM | Deploy a 170M parameter unified second-pass rescoring model for en-US on accelerated hardware with at least a 5% relative reduction in tail WER across Shopping, Entertainment, Info and Local Search domains |
| 47 | VP Goals | 580808 | DNM | Multi-turn user repeats and rephrases: Reduce WER  by 15% relative for user repeats and rephrases in en\_US, DE and GB. |
| 48 | VP Goals | 586526 | DNM | Reduce word error rate by 10% relative traffic-weighted averaged over all endpoints for Alexa Mobile Assistant |
| 49 | VP Goals | 715910 | DNM | Reduce Customer Perceived Defect Rate (CPDR) for all en\_US traffic contributed by ASR by 60 bps YoY |
| 50 | VP Goals | 579587 | DNM | Achieve 5% absolute slot WER for personal content and 30% slot WER relative improvement for rare personal content measured on a list of 14 personal catalogs (smart home device names, contact names, etc.). |
| 51 | VP Goals | 583352 | DNM | Deliver 20% relative WER improvement for on-screen selections (dialogs) on EFD devices in en\_US, DE and en\_GB locales |
| 52 | VP Goals | 580803 | DNM | Achieve 5% WER (absolute) at p99 threshold (worst performing 1% utterances) for on-screen hints. |
| 53 | VP Goals | 579596 | DNM | Reduce mis-recognitions by 50% relative for affinity based personal catalogs for Song Name, VideoName and ItemName slots in en-US, en-GB, de-DE. |

## Publications and Academic collaboration:

Refer to [Speech/ASR Publications & Patents](https://quip-amazon.com/2Cj8AasBAibd) and [ASR Publications & Patents 2023](https://quip-amazon.com/ciZ8AOwCDa08) for full list and details on patents, publications etc.

* **Publications** 20 accepted papers in 2023 (9 ICASSP, 1 ICML, 6 InterSpeech, 3 ASRU), 5 accept to ICASSP 2024.
  + **2 top-3% paper award at ICASSP 2023** (“[Cross-utterance ASR Rescoring with Graph-based Label Propagation](https://arxiv.org/pdf/2303.15132.pdf)”, “[On-the-fly text retrieval for end-to-end ASR adaptation](https://www.amazon.science/publications/on-the-fly-text-retrieval-for-end-to-end-asr-adaptation)“)
* **Amazon Research Liaisons.**We managed three university relationships (John Hopkins, University of Illinois Urbana Champaign and Virginia Tech) on atypical speech and data selection, and with the University of California Santa Barbara (jointly with AGI Foundations org) on compressed tensor-training of ASR and NLU transformer models. We supervised one PhD student at TU-Braunschweig. We mentored three PhD students at JHU on foundation modeling, multi-party ASR and contextual ASR. We developed datasets for contextual ASR and for atypical speech through these university collaborations. These collaborations led to 2 NeurIPS papers, 1 ASRU paper in 2023, and 3 papers submitted to INTERSPEECH, COLING and ACL in 2024.
* **Patents & Invention disclosures:** 3 issued patents and 14 IDF

## Salient updates

**New Data Foundation program.** Data preparation is key to our success and includes several major steps (procurement, synthetic data generation, label preparation for all learning tasks, ranging from ASR, end-pointing, backchanneling, to emotion and TTS, and more). Our current data preparation tools only cover a subset of these steps. To build competitive speech2speech models, we started a new program to consolidate and extend our data preparation pipeline in collaboration with other orgs (ASR, TTS, APT, NU).

**New team publication guidelines** - We want to change our publications to fewer but more impactful publications, we want to use publications to establish us as a technology leader in our field. When we started Alexa we were busy the first years to build an innovative product and to establish us as one of the leading labs in our field. In the first years we had to simply catch up and to build a state-of-the-art technology stack. Once we achieved this we started publishing at scale. We have now a similar presentation at conferences in our field as other leading industrial research labs, our papers get read and frequently cited. Now it is time to take the next step, to become a technology leader in our field - and our field is not only Speech anymore, but we have shifted to building general ML solutions focused on interactive and self-learning AI which we (so far) apply mainly to Speech problems. Concretely, this means that we want fewer publications but each publication being of high impact. This also means that each publication is expected to be bigger in ambition, more exhaustive experimental work, and larger in data than our today’s average publication. This means also we target the prestigious ML/AI conferences (ICML, ICLR, NeurIPS, etc.) in addition to the traditional Speech conferences (ICASSP, Interspeech, etc.). We are defining new publications guidelines are to achieve technology leadership in our field.