**Tech Promotion Work Summary**

# Employee Instructions

1. **Partner with your manager to define a next-level goal** (e.g., project, research, presentation, etc.) that allows you to demonstrate one or more next-level capabilities, development areas, and/or Leadership Principles. Next-level criteria can be found in your [Role Guideline](https://ivy-help-center.talent.a2z.com/article/article-1568200618517-Gtmgg1nif) in the “Moving to…” a level section. If your guideline is missing this section, your manager will need to clarify which expectations and technical skills at the next level they want you to demonstrate.
2. **Document *Planned Work* in Part I**. The purpose of documenting a planned goal is to get clear on what is expected and how you and your manager foresee that it gives you the opportunity to demonstrate the next level. If you get a new manager, this can inform them of previous agreements and inform them about what you are working on (relative to the next level).
3. **When you complete the work, fill out Part II*.*** This is a best practice. If you change teams or get a new manager, this gives the new manager a documented record of the work you have done to demonstrate the next level. It is also a record of the impact you had and how your work demonstrates our [Leadership Principles](https://inside.amazon.com/en/About/corevalues_EN/Pages/LeadershipPrinciples.aspx).
4. **Provide at least two work samples that best demonstrate the work** **delivered** (e.g., scope, impact, expertise, technical skill, etc.). Work samples vary by job role. Examples:
   1. Technical work samples - Demonstrate technical and/or scientific expertise. These include scientific research papers, models, algorithm design documents, technical specifications, working mockups/prototypes, code samples, code review commentary (both given and received), architecture/design documents, patent submissions, etc.
   2. Standard work samples - Demonstrate other skills. These include PR/FAQs, narratives, Product Roadmaps, OP1 documents, MBR/QBR materials, functional specifications, UX designs, presentation decks, broadcast videos, project proposals/ROI case studies, wikis, resource planning documents, etc.
5. **After your manager has documented their feedback in Part III, discuss whether it would be helpful to gather additional peer and/or customer feedback**. This is highly recommended, but not required. Your manager gathers the feedback and decides whether some comments are appropriate to add to **Part III**. Work Summary feedback providers can be any level, however higher levels offer better next-level perspectives.

# Manager Instructions

1. **Partner with your employee on a one or more goals that give them the opportunity to demonstrate the next level**. Review the next-level criteria defined in the employee’s [Role Guideline](https://ivy-help-center.talent.a2z.com/article/article-1568200618517-Gtmgg1nif). Design projects or other goals that deliver value for the team’s customer, technology, and business domain. Goals can also give them the opportunity to demonstrate Leadership Principles or resolve a gap identified when hiring or in a previous attempt to promote**.**
2. **If the employee is getting a Planned or Completed Work Review** – Make sure the employee’s summary is clear and well written.
3. **When the work is completed:**
   1. Evaluate the quality of the employee’s work and the results they delivered against the next-level capabilities (and Tech bar) expressed in the employee’s Role Guideline
   2. Document your feedback in **Part III**
   3. Make sure their work samples demonstrate key skills required at the next level
   4. Discuss whether there are peers or customers the employee thinks would be good to get feedback from about the specific effort in the work summary.

# PART I: PLANNED WORK

(Part I required for a [Planned Work Review](https://ivy-help-center.talent.a2z.com/article/article-1588633382719-aPKCqTQob).)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Employee Information** | | | | |
| **Employee Name:** | Kai Zhen | **Current Job Title:** | Applied Scientist II |
| **Manager Name:** | Hieu Duy Nguyen | **Proposed Job Title:** | Applied Scientist III |
| **Steam Member:** | Dave Limp | **Current Business Title:** | Applied Scientist |  |
| **Steam Direct:** | Tom Taylor | **Proposed Business Title:** | Sr. Applied Scientist |  |  |
|  | | | | |
| **Goal (e.g., Project, narrative, presentation, DEMONSTRATE A LEADERSHIP PRINCIPLE)** | | | | |
| *Use your Role Guideline to define a single goal that allows you to demonstrate the type of work expected at the next level, our Leadership Principles, or other area to be considered for a promotion. Summarize in one or two paragraphs.* | | | | |
| The landscape of Automatic Speech Recognition (ASR) has undergone a significant shift towards Generative AI, redefining its computational paradigm. Core ASR neural components now boast over 1 billion parameters (as seen in Let’s Chat) and are projected to reach 7 billion and beyond in the next generation of speech-to-speech architectures. This evolution promises Alexa users a more natural and expressive conversational experience through enhanced speech responses. However, it also entails substantial costs in terms of both training and inference time. Ensuring cost parity, alongside maintaining acceptable user-perceived latency, remains crucial to sustain the business viability of these ASR services (Kingpin Goal xxxxx).  This project aims to develop a versatile neural efficiency tool with an emphasis on sparsification. The primary objective is to achieve runtime efficiency by accelerating core Alexa ASR components on Nvidia’s Ampere-based GPU instances, which serves as a pivotal factor in attaining “latency- and cost-neutral operating points”. The specific goals encompass (a) adapting the existing compression-aware training methodology to the finetuning stage to reduce training costs; (b) developing and merging highly reusable, production-ready code in Phasa for Conformer, PyRama for Transformer, and future Speech2Text LLMs.; (c) collaborating with both internal and external hardware teams facilitate the end-to-end accuracy and latency testing regarding sparsification using TensorRT-9 SDK; (d) establishing alignment with Language Modeling team to showcase or productize sparsification in their BERT-based Rescorer for en-US locale. Measuring the goals: (a) achieving an affordable training cost with a target duration of <2 days to attain model sparsity; (b) implementing the sparsity pattern that matches what the hardware actually supports to realize >10% latency reduction observed from the p50 production data; (c) minimizing the accuracy impact, ideally 2% or less relative degradation from most test sets. To execute these goals effectively: (a) initiating mini model releases with the release owner to showcase consistent accuracy within desired ranges when sparsity is enabled; (b) hosting bi-weekly sync meetings with key figures from hardware and ASR science teams to address potential blockers and facilitate collaboration; (c) providing regular updates to L6+ stakeholders via email or LR highlights to gather guidance and support for next steps. To keep the goal realistic, I should (a) make the sparsification algorithm kernel-wise or layer-wise configurable, such that we could retain certain weights' density to ensure an acceptable accuracy level; (b) prioritize the productization for one ASR component, such as the language model, first, while showcasing the effectiveness of the other component. To exhibit the qualities of the next level, I need to (a) demonstrate strong ownership of designing the sparsification algorithm and leading technology implementation while coordinating cross-team efforts; (b) deliver a robust, high-quality, and end-to-end neural efficiency solution on time with tangible production impact; (c) earn trust and bring influence to multiple teams regarding their current and future production roadmap. | | | | |
|  | | | | |
| **Associated Next-Level Criteria and/or Leadership Principles** | | | | |
| *Cut and paste the next-level criteria or Leadership Principles that this work intends to demonstrate. A single goal can demonstrate readiness in one or more areas. If the work is to demonstrate a different requirement for promotion, document it here.* | | | | |
| **Associated Next Level Criteria:**   * You lead the design, implementation, and successful delivery of large, critical, or difficult scientific solutions involving a significant amount of work. These efforts can be new solutions or a refactor of existing solutions. You heavily influence the design and write a significant portion of the “critical-path” code. * You demonstrate influence over 1-2 teams, either via a collaborative efforts or by increasing their scientific knowledge, usage of advanced or specialized techniques, or engineering best practices. * You influence your team’s science and business strategy by making insightful contributions to team priorities and approach. You take the lead in identifying and solving ambiguous problems, science deficiencies, or areas where your team’s solutions bottlenecks the innovation of other teams. You make solutions simpler. * Your solutions, code, designs, and scientific artifacts (e.g., papers) set a great example to others. You work very efficiently and routinely deliver the right things.   **Leadership Principles:**   * **Deliver Results:** Leaders focus on the key inputs for their business and deliver them with the right quality and in a timely fashion. Despite setbacks, they rise to the occasion and never settle. * **Have Backbone; Disagree and Commit:** Leaders are obligated to respectfully challenge decisions when they disagree, even when doing so is uncomfortable or exhausting. Leaders have conviction and are tenacious. They do not compromise for the sake of social cohesion. Once a decision is determined, they commit wholly. * **Think Big**: Thinking small is a self-fulfilling prophecy. Leaders create and communicate a bold direction that inspires results. They think differently and look around corners for ways to serve customers. | | | | |
|  | | | | |

# PART II: Results

(Complete this section after you finish the work. Parts I, II, and III required for [Completed Work Reviews](https://ivy-help-center.talent.a2z.com/article/article-1588633382719-aPKCqTQob).)

|  |
| --- |
| **Quality of Work/Challenge** |
| *Summarize how your work demonstrated the* ***challenge*** *expected at the next level. Where possible, give examples.* |
| **Ambiguity:**  The definition of sparsification is ambiguous with multiple factors. What’s the criterion to prune weights? Is the magnitude-based pruning optimal for Alexa production stack? Should the sparsity be structured or unstructured? More importantly, what sparsity structure should we pursue and how it can be accelerated on the hardware to realize the actual benefits?  In terms of sparsification policy, there has also been debate on when to invoke sparsification on which layers. For example, making sparsifiaction a built-in component from scratch when training a model is considered of minimal accuracy impact. However, the cost and model build overhead can be formidable. While feasible for previous small model releases, sparsity-aware training may not be the perfect fit for large-ASR and next-gen speech-to-speech models. In contrast, post-training sparsification compresses the model in one-shot without requiring any data and training cost. Yet, the accuracy degradation, even small, might be a potential blocker when releasing the feature.  **Scope of Influence:**  **Scientific and Technical Complexity:**  **Impact:**  In this project, we developed sparsity-aware finetuning (SAF), a groundbreaking paradigm to compress Alexa's acoustic and language models by up to 50% without sacrificing performance, featuring both training time and runtime efficiency. SAF is crucial to achieve cost parity and improve accuracy-latency tradeoffs. Implemented successfully in Lets-Chat v2.0, SAF reduced latency by 17.8% on p50 data setting and ~40.0% on p99 in the BERT-based Rescorer on TensorRT-9 while maintaining high accuracy across various test sets. SAF's application extended to compressing the 1B param Conformer, showcasing <0.5% relative degradation on both Alexa\_prize\_onlygoldens and lets-chat-alpha with up to 10% latency reduction even with batch size of 1. The combination with INT8 quantization led to 61.9% throughput improvement and 38.3% latency reduction, making it viable to host 7B-30B param Llama on a single GPU. Merged in Phasa/PyRama Core through rigorous validation and successful production intake, SAF solidified its role as a versatile and efficient tool within Alexa's neural efficiency methods, poised for present and future hardware advancements.  **Execution:**  **Knowledge:**  **General qualifications:**  I began my journey with AHS-ASR in May 2020 as an applied scientist intern, already having published a paper on ASR model sparsification during this internship. Subsequently, I transitioned to a full-time role with the AHS ASR group at Amazon PIT, where I led several impactful neural efficiency projects for both on-device and cloud ASR, significantly influencing production and research initiatives. I earned a Ph.D. in computer science and cognitive science from Indiana University Bloomington, where I spearheaded multiple speech and audio neural waveform coding projects, pioneering a new research area. The results of my work were published in prestigious signal processing and speech processing conferences and journals, including Interspeech, ICASSP, IEEE Signal Processing Letters (SPL), and IEEE T-ASLP. In recognition of my research contributions, the Cognitive Science Program at IU honored me with the Outstanding Research Award in 2021. Additionally, I hold six US patents as an inventor, showcasing my contributions to innovative technological solutions. Before joining Amazon, I contributed to BERT-based multi-modal recommendation systems as a machine learning relevance intern at LinkedIn from 2018 to 2019. My educational background includes an M.S. in Computer Science from Tsinghua University in 2015 and a B.S. in Software Engineering from Xidian University in 2012. During my time at Xidian University, I received the Honorable Mention award for the International Mathematical Contest in Modeling (MCM) and was a three-time recipient of the National Scholarship. |

|  |
| --- |
| **Impact/Measure of Success** |
| *Summarize in one or two paragraphs the impact this effort had on customers, your team/organization/business area, partners, and others. What data (quantitative1 metrics or qualitative measures) can you provide?* |
| In this project, we proposed, implemented, and productized the Alexa-specific policy-based sparsity-aware finetuning (SAF) paradigm. This approach can compress the Alexa acoustic model (1B+ param Conformer) and language model (1B+ param BERT) by up to 50% without compromising their predictive performance. Achieving this holds significant value in attaining cost parity and enhancing accuracy-latency tradeoffs. SAF also features reducing training costs without necessitating training-from-scratch. It efficiently delivers the expected sparsified/pruned ASR components from their pre-trained versions in just a few shots or less than 1 day's effort. This design of SAF aligns perfectly with the development process of Let’s Chat Alexa ASR, minimizing production intake risks.  In 2023 Q4, we've successfully productized SAF by incorporating N:M sparsity in BERT-based 1B param Rescorer for the Lets-Chat v2.0 release. Among the 20 layers, we pruned 50% of the weights to zero across 14 layers using N:M SAF. This resulted in a 17.8% relative reduction in latency on TensorRT-9, utilizing <7x16> inputs as the p50 production setting. In terms of accuracy, the sparse Rescorer maintained consistent predictive performance (<1.5% relative degradation) from over 40 test sets in end-to-end benchmarking. As a results, the sparse ContextBERT retained a Word Error Rate Reduction (WERR) of >20% on contact-name test set, achieved from Gazetteer modeling and first-pass Neural Biasing, from the previous mainline production model, bringing down the accuracy gap against the small cloud RNN-T model on contact-name to <1%, while simultaneously achieving significant reduction for the memory footprint and latency. In addition to BERT-based language model, SAF's application in compressing the 1B param Conformer, the acoustic model, was showcased: with over 90% kernels in the reduced-frame-rate encoder of the Conformer pruned by 50%, the Let’s Chat GPU package only exhibits <0.5% relative degradation on both Alexa\_prize\_onlygoldens and lets-chat-alpha, yet with up to 10% latency reduction measured via NeMoRT-Bench.  SAF is directly compatible with INT8 quantization on TensorRT-9 for hardware acceleration. I coordinated with Nvidia and requested the hardware benchmark: Compared to non-sparse BF16 baseline, the sparse INT8 Transformer yields 61.9% throughput improvement and 38.3% latency reduction, both measured on the p50 data with the batch size of 7 and sequence length of 16. With NeMoRT’s custom acceleration kernel as an ONNX plug-in, 2:4 sparsity yields ~40% latency reduction. Along with INT8 quantization, one can host Llama up to 30B parameters on a single GPU.  The implementation of SAF underwent rigorous scrutiny by SDEs for both PyRama and Phasa before merging into their respective Core folders in the mainline. Additionally, as the technology owner, I oversaw multiple rounds of mini model releases in tandem with the release owner, ensuring the correctness of SAF at each step toward final deployment. These efforts bolster the usability and robustness of our sparsification scheme for stakeholders, enabling timely delivery without exerting much burden to the release owner. Moreover, SAF stands out as a versatile neural efficiency paradigm, supporting not only N:M fine-grained sparsity but also various other arbitrary sparsity patterns, such as block sparsity, with minimal effort. This adaptability ensures SAF remains one of Alexa’s effective neural efficiency tools, irrespective of hardware types, both presently and in the future. |
|  |

|  |
| --- |
| **Leadership Skill & InFluence (Section required for promotions into L6+ only)** |
| Please describe how this work gave you the opportunity to demonstrate any other leadership skills (e.g., drive a best practice, influence a needed change, build consensus on an approach)? How did you apply judgment (e.g., design choices, trade-offs, priorities) or make decisions with long-term effects? |
| In this project, we pioneered Sparsity-Aware Finetuning (SAF) to compress Alexa's models by up to 50% while maintaining performance and boosting runtime efficiency, crucial for cost parity. Implemented in Lets-Chat v2.0, it cut latency by 17.8% (p50) and ~40.0% (p99) in BERT-based Rescorer on TensorRT-9, preserving accuracy. It compressed the 1B param Conformer with <0.5% degradation on test sets and up to 10% latency reduction even at batch size 1. Combined with INT8 quantization, SAF achieved 61.9% throughput improvement and 38.3% latency reduction, enabling hosting of large models like Llama on a single GPU. Merged into Phasa/PyRama Core after validation, SAF solidified its role in optimizing neural efficiency for present and future Alexa's systems.  **Deliver Results:**  While sparsification stands as a well-established model compression technique in the literature, delivering it on the right hardware to bring actual benefits to our customers’ experience remains a complex endeavor. To begin with, implementing Conformer and BERT in two separate code bases, namely Phasa and PyRama respectively, necessitates the development of a sparsification-aware fine-tuning (SAF) mechanism for both platforms, demanding a substantial amount of effort. Additionally, both Conformer and BERT are with multiple stages of training in production setting: Enabling SAF on the right stage with right version necessitates tremendous efforts on coordination.  To deliver results, I consistently engaged with the release owners for both Conformer and BERT to build consensus early on that SAF does not impact accuracy. Take BERT as an example, we first proved that SAF yielded little-to-no accuracy degradation on 170M BERT back in Q2/2023. Because of that, the Seattle-LM team (Yile Gu, Senior Applied Scientist and Jari Kolehmainen, Senior Applied Scientist) was inclined to support us on applying SAF to 1B BERT in Q3/2023, and then ContextBERT with Gazetteer finetuning in Q4/2024 with end-to-end Dory-BlueShift decoding on over 40 test sets for accuracy verification. Moreover, I proactively collaborated with Buddha N (Senior SDE from NeMoRT) and Sonal Pareek (Senior SDE from NeMoRT) on hardware testing for latency metrics for several rounds, meticulously documenting quantitative evidence of latency reduction and throughput improvement through sparsification.  From initiating early consensus-building to conducting rounds of model-level testing and culminating in comprehensive end-to-end testing, we successfully delivered results: Alexa’s first sparse Transformer that maintains in-range accuracy across over 40 test sets and achieves approximately 20% latency reduction. Furthermore, we established a robust and reusable SAF paradigm that naturally applies to most ASR/NU neural components in the PyRama mainline.  **Have Backbone; Disagree and Commit:**  We encountered significant challenges during the rollout of sparsification for the large ASR model release. One major issue stemmed from the instability of the TensorRT-9 software development kit (SDK), causing a delay in the general access (GA) SDK's release by over a quarter, from August 28th to December 7th. Consequently, our sparse model, initially scheduled for release in November, wouldn't yield any latency benefits until after the SDK's deployment. This led to setbacks in obtaining p4 instances and necessitated a potential delay to our release plan. As the technology owner, I authored documentation detailing the current state of sparsification [xxx], emphasizing its technological readiness. All necessary codes have been merged into the mainline, and enabling sparsity for BERT requires minimal effort, just half a day of fine-tuning. Additionally, I highlighted previously demonstrated evidence regarding accuracy and latency improvements. I communicated extensively via email and discussions with various stakeholders involved in hardware acceleration and large ASR modeling within the LR, including Denis Filimonov (Principal Applied Scientist at Alexa-ASR), Ariya Rastrow (Senior Principal Applied Scientist at Alexa-ASR), Ehry MacRostie (Senior Manager at Alexa Speech), and Rolando Jimenez (Principal TPM at Alexa Speech), among others.  Despite the challenges, I remained steadfast and committed to the project. With the support of these stakeholders, we collectively decided to proceed with the original release plan. Eventually, we successfully delivered sparse BERT in v2.0 on time.  **Think Big:**  **I make bold moves w. caution.**  **Think corners. Never be conveniently naïve. Work with hot-fixing, stuff.** |

|  |  |  |
| --- | --- | --- |
| **WORK SAMPLES** | | |
|  | **Work Samples Name and Link** |
|  | [First 8-bit cloud Conformer model release (de-DE)](https://wiki.labcollab.net/confluence/pages/viewpage.action?pageId=2054818930) | Wiki for general quantization (GQ) and its intake for cloud ASR (de-DE v59) | |
|  | [First 5-bit en-US Bluebottle model release](https://wiki.labcollab.net/confluence/display/SHELBY/BlueBottle+R15+en-US+RNN-T+Release#Project-388125852) | Wiki for sub-8-bit trained en-US Bluebottle R15 release as the first 5-bit on-device ASR model | |
|  | [First 5-bit non en-US Crosstown model release](https://wiki.labcollab.net/confluence/display/SHELBY/Training+and+Delivery+of+Crosstown+es-ES+v3+ASR+model) | Wiki for sub-8-bit trained es-ES Crosstown R3 release as the first 5-bit trained non English ASR model | |
|  | [Launch announcement](https://quip-amazon.com/bTBQAjhiM2KM/Launch-Announcement-8-bit-Cloud-Conformer-Training-via-General-Quantization) | Launch Announcement for 8-bit Cloud Conformer Training via General Quantization | |
|  | [Runtime Modeling LR Highlights](https://quip-amazon.com/nykpAZKcR9kV/Runtime-Modeling-Bi-Weekly-Update-07112023#temp:C:TAZ529cda86fbcf4f208965c3b3f) | Highlights on 8-bit Cloud Conformer Training via General Quantization after its deployment | |
|  | [Live Latency Monitor](https://monitorportal.amazon.com/igraph?SchemaName1=Service&DataSet1=Prod&Marketplace1=USAmazon%3Abrownie&HostGroup1=ALL&Host1=ALL&ServiceName1=AlexaHybridEngine&MethodName1=FirstPassRecognition&Client1=ALL&MetricClass1=NONE&Instance1=NONE&Metric1=stage1_pryon_latency_msec.es-ES&Period1=OneDay&Stat1=p50&ValueUnit1=microsecond&LiveData1=true&Label1=p50%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color1=1600ff&Visible1=false&UserLabel1=p50%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName2=Service&Stat2=p90&Label2=p90%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color2=0a0a0a&UserLabel2=p90%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName3=Service&Stat3=p99&Label3=p99%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color3=ff00ff&Visible3=true&UserLabel3=p99%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&SchemaName4=Service&Period4=OneHour&Stat4=n&ValueUnit4=millisecond&LiveData4=false&YAxisPreference4=right&Label4=n%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&Color4=02e0e0&Visible4=false&UserLabel4=n%20-%20OneHour%20%5Bmin%3A%20%7Bmin%7D%2C%20avg%3A%20%7Bavg%7D%2C%20max%3A%20%7Bmax%7D%5D&HeightInPixels=720&WidthInPixels=1560&GraphTitle=stage1_pryon_latency_msec&DecoratePoints=true&GraphType=zoomer&HorizontalLineLeft1=%23color%3Dblue%20Sev2%20%28p50%29%20-%20@%201920.58&HorizontalLineLeft2=%23color%3Dgreen%20Sev2%20%28p90%29%20-%20@%204206.02&HorizontalLineLeft3=%23color%3Dmagenta%20Sev2%20%28p99%29%20-%20@%209432.25&HorizontalLineLeft4=%23color%3Dblue%20Sev3%20%28p50%29%20-%20@%201139.68&HorizontalLineLeft5=%23color%3Dgreen%20Sev3%20%28p90%29%20-%20@%202616.33&HorizontalLineLeft6=%23color%3Dmagenta%20Sev3%20%28p99%29%20-%20@%206151.32&StartTime1=-P470D&EndTime1=-PT0H) | See how the pryon latency noticeably dropped after the sub-8-bit model’s deployment at the end of Aug, 2022. | |
|  | [Sub-8-bit quantization-aware training](https://assets.amazon.science/fe/84/ad0cdd7c4967b17aaf670fe0194b/sub-8-bit-quantization-aware-training-for-8-bit-neural-network-accelerator-with-on-device-speech-recognition.pdf) | Conference: Interspeech, 2022  Title: Sub-8-Bit Quantization Aware Training for 8-Bit Neural Network Accelerator with On-Device Speech Recognition  Authors: **Kai Zhen**, Hieu Duy Nguyen, Raviteja Chinta, Nathan Susanj, Athanasios Mouchtaris, Tariq Afzal, and Ariya Rastrow  **Deployed to on-device ASR.** | |
|  | [General quantization](https://assets.amazon.science/0c/03/41fc077547799c2350ccb3a4ac15/sub-8-bit-quantization-for-on-device-speech-recognition-a-regularization-free-approach.pdf) | Conference: Spoken Language Technology Workshop (IEEE SLT), 2023  Title: Sub-8-Bit Quantization for On-Device Speech Recognition: A Regularization-Free Approach  Authors: **Kai Zhen**, Martin Radfar, Hieu Nguyen, Grant Strimel, Nathan Susanj, Athanasios Mouchtaris  **Deployed to both on-device and cloud ASR.** | |
|  | Patent on ASR model quantization | File Number: P77898-US01  Title: COMPRESSION OF MACHINE LEARNED MODELS  Inventors: **Kai Zhen**, Hieu Nguyen, Raviteja Chinta, Tariq Afzal, Anastasios Alexandridis, Athanasios Mouchtaris, Ariya Rastrow | |

**--SUMMARY SHOULD NOT EXCEED 3 PAGES--**

# PART III: Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| **Manager FeedbACK** | | | |
| *Date:* | | | |
|  | | | |
|  | | | |
| **Additional FeedbACK** | | | |
| **feedback provider Name:** |  | **Feedback provider business title:** |  |
|  |  |  |  |
| **Feedback date:** |  | **Feedback provider level:** |  |
| **Feedback type:** | Select | **Relationship to candidate:** | *(e.g., customer, peer, stakeholder, teammate)* |
|  | | | |
| **Additional FeedbACK** | | | |
| **feedback provider Name:** |  | **Feedback provider business title:** |  |
| **Feedback date:** |  | **Feedback provider level:** |  |
| **Feedback type:** | Select | **Relationship to candidate:** | *(e.g., customer, peer, stakeholder, teammate)* |
|  | | | |
|  | | | |