On-Device AI & The Legal Industry:

Balancing Innovation with Confidentiality

Agenda

- The Opportunity and Challenges of AI in the Legal Industry
- Introduction to On-Device AI
- Under the Hood: Small LLMs, Distributed Inference, and Compound AI Systems
- Thinking in Layers: Cloud, 3rd Party Managed, Internal, and On-Device
- Building a Holistic AI Strategy
- The Future: A Hybrid Approach

Introductions & Background

Fred Bliss

Founder & Advisory @
Intersect Next

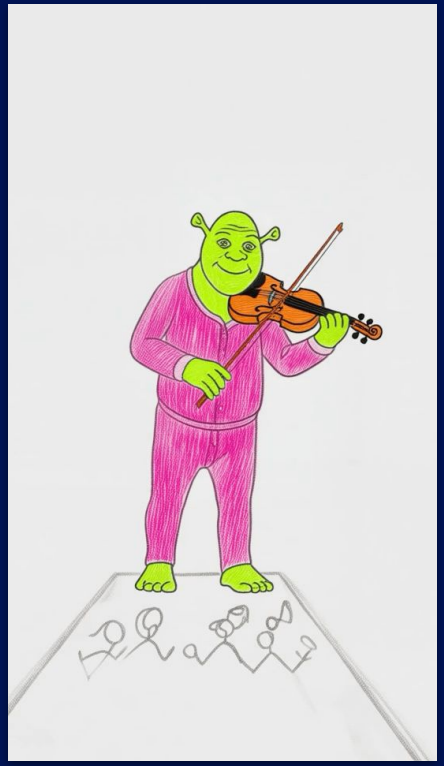
fred@intersectnext.ai

Eugene Goryunov

Partner & co-chair of AI & Deep Learning practice group
@ Haynes Boone

eugene.goryunov@haynesboone.com

Where we were in 2014 vs today



AI's Potential in Legal: Opportunities and Challenges

- Initial ‘ChatGPT paradigm’ of UI/UX mirror today’s most common use cases
- Toward a ‘Golden Era’ of custom-developed, industry-specific, specialized applications
- As deeper integration approaches, open security questions remain (ie: indirect prompt injection)

Organizational Data as a Differentiator

Data Privacy, Security, and Client Confidentiality Concerns

Cloud Parallels & The Need for a Layered Approach

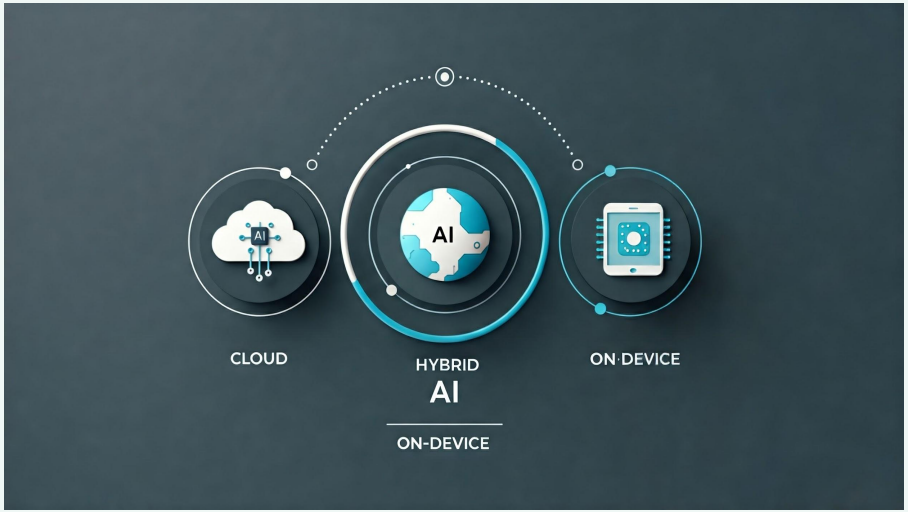
Common Adoption Paths

- Single Provider Approach
- Cloud-Based AI, Multiple Providers
- “On-Premise” Approach (3rd Party Managed)
- Internally Managed Approach
- Hybrid

Thinking in Layers

How can we encourage innovation and experimentation with a rapidly changing technology while maintaining privacy, security, and confidentiality?

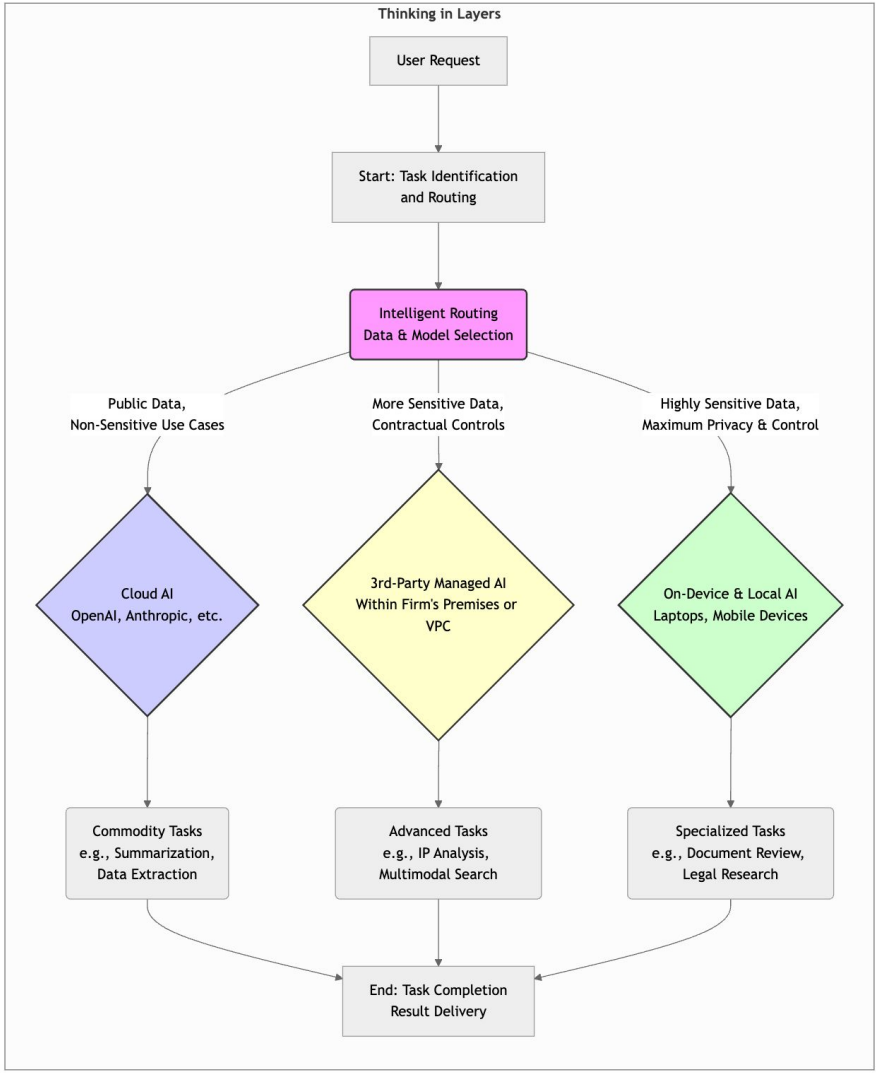
A Framework for Privacy-Centric AI: Tiered Solutions for a Hybrid World



Thinking in Layers

- AI Deployments in the Cloud
- Managed by a 3rd Party
- Internally Hosted
- On-Device Inference

... and everything in-between



Introducing On-Device AI

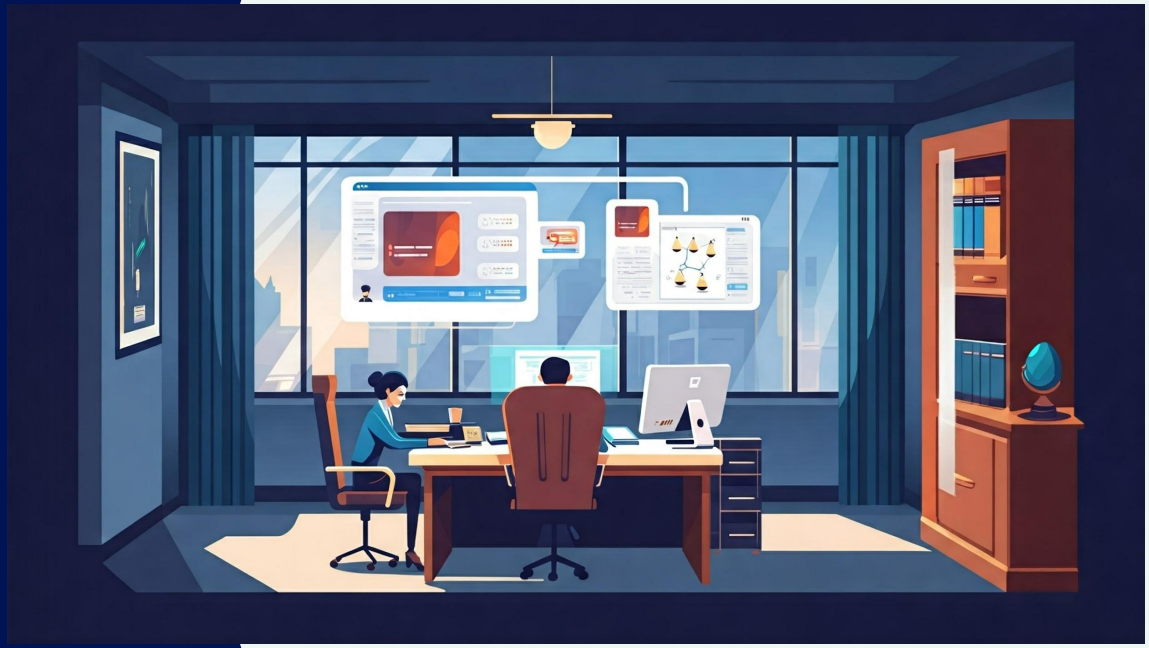
Large Foundation Models...

... distilled down to highly optimized small models

... that can run on laptops, mobile devices, and more

On-device AI Benefits:

- Privacy
- Latency
- Cost
- Experimentation
- Adaptability
- Ownership
- Custom Applications



How On-Device AI Works

SLMs and enabling techniques:

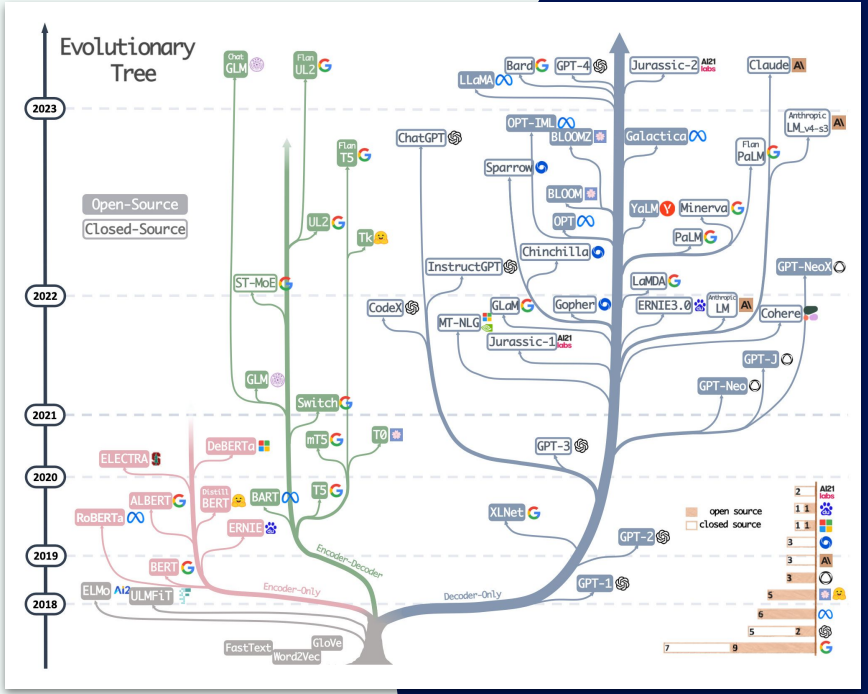
- Quantization
- Optimization
- Open source community
- “GPU poor”

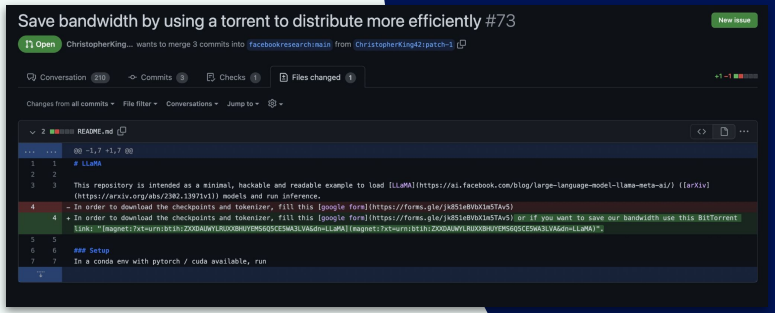
Examples of applications:

- Apple Intelligence
- Windows Copilot
- Our mobile devices

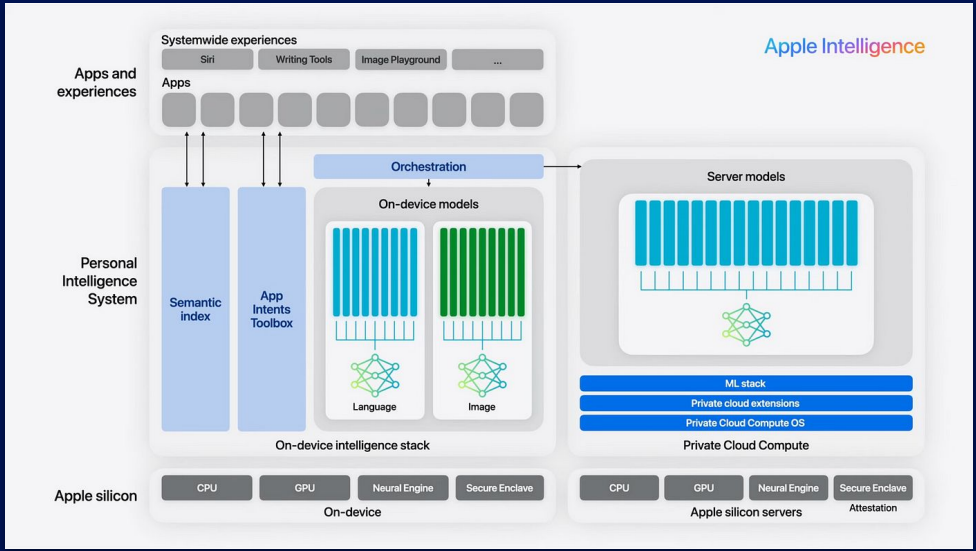
A Short History on Local LLMs

… and the trajectory of open source, local models, and the ecosystem





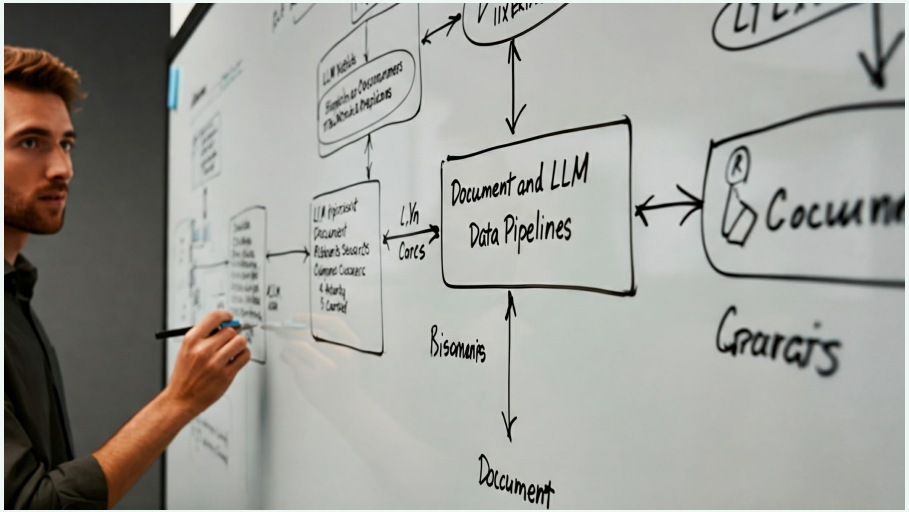
March 2023



Apple Intelligence as an example and enabler of both deep integration and distributed inference

On-device LLM and multimodal use cases

From summarization, audio transcription, and visual annotation



To AI-driven search, LLM data pipelines, and more powerful analytics

Practical Considerations

Key Points

** Stay informed on progress*
** Beware of privacy mishype (as always)*
** Understand nuances and limitations of Today vs. Near Term*

Considerations

- Current capabilities vs. 1-2 years ago
- Rise & Implication of Open Source Multimodal and Vision LLMs
- Integration with Applications
- Unique Use Cases for Low-Latency
- Hybrid Use Cases
- Distributed Use Cases

Beyond Single Models:

The Rise of Compound and Hybrid AI Systems

The Shift from Models to Compound AI Systems

Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, Ali Ghodsi

Feb 18, 2024

AI caught everyone's attention in 2023 with Large Language Models (LLMs) that can be instructed to perform general tasks, such as translation or coding, just by prompting. This naturally led to an intense focus on models as the primary ingredient in AI application development, with everyone wondering what capabilities new LLMs will bring. As more developers begin to build using LLMs, however, we believe that this focus is rapidly changing: **state-of-the-art AI results are increasingly obtained by compound systems with multiple components, not just monolithic models**.

For example, Google's AlphaCode 2 set state-of-the-art results in programming through a carefully engineered system that uses LLMs to generate up to 1 million possible solutions for a task and then filter down the set. AlphaGeometry, likewise, combines an LLM with a traditional symbolic solver to tackle olympiad problems. In enterprises, our colleagues at Databricks found that 60% of LLM applications use some form of retrieval-augmented generation (RAG), and 30% use multi-step chains. Even researchers working on traditional language model tasks, who used to report results from a single LLM call, are now reporting results from increasingly complex inference strategies: Microsoft wrote about a chaining strategy that exceeded GPT-4's accuracy on medical exams by 9%, and Google's Gemini launch post measured its MMLU benchmark results using a new CoT@32 inference strategy that calls the model 32 times, which raised questions about its comparison to just a single call to GPT-4. This shift to compound systems opens many interesting design questions, but it is also exciting, because it means leading AI results can be achieved through clever engineering, not just scaling up training.

In this post, we analyze the trend toward compound AI systems and what it means for AI developers. Why are developers building compound systems? Is this paradigm here to stay as models improve? And what are the emerging tools for developing and optimizing such systems —an area that has received far less research than model training? We argue that **compound AI systems will likely be the best way to maximize AI results in the future**, and might be one of the most impactful trends in AI in 2024.

Increasingly many new AI results are from compound systems.

Forward-Looking: Contextual UIs and Integration

- Seamless integration with enterprise applications
- Innovative (yet familiar) UI/UX paradigms powering new ways of working
- Automation of the Mundane, with more control over the Parts That Matter
- Control over the processing and routing when dealing with sensitive data

How Small, Large, and Everything In-Between will Play a Role in Compound AI Systems

- Small models for efficient on-device inference
- Large models for complex tasks
- Hybrid approach within custom applications and orchestration workflows

Building a Holistic Organizational AI & Data Strategy

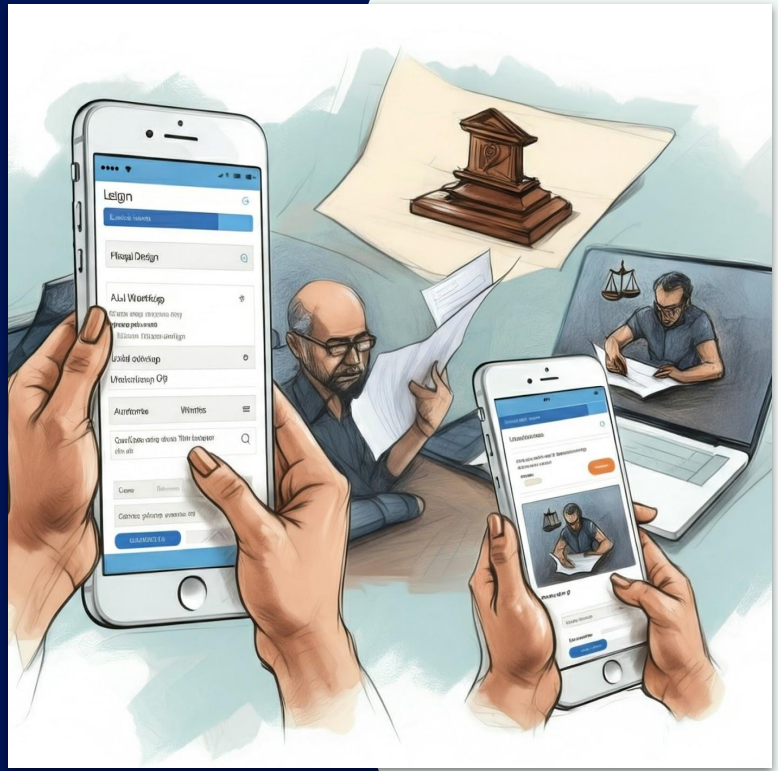
- Trusted, tech-agnostic partner
- Build a foundational vision
- Plan for change (people, process)
- Adoption, as always, is the hardest part

- Take a data-centric approach
- Bottoms-up approach - small, impactful use cases
- Your data is still your moat and competitive advantage
- “Automate the BS”

The Importance of AI Evaluation & Feedback Loops

The Future is Hybrid: Towards Smarter, Seamless AI Integration

- Increased on-device adoption
- Role of intelligent systems
- Multimodal capabilities
- Legal industry's leadership in responsible AI



Thank You!

Fred Bliss

Founder & Advisory @
Intersect Next

fred@intersectnext.ai

Eugene Goryunov

Partner & co-chair of AI & Deep Learning practice group
@ Haynes Boone

eugene.goryunov@haynesboone.com