Foundation Models & AI Engineering – Exam Ready Notes

1. Foundation Models

- Large-scale models trained on massive datasets using self-supervision.
- Includes Large Language Models (LLMs) and large multimodal models.
- Designed to be general-purpose rather than task-specific.
- Can be adapted to many downstream tasks with minimal additional data.

2. Self-Supervision in Multimodal Models

- Learning paradigm that does not require manually labeled data.
- Uses natural correlations in data (e.g., image and accompanying text).
- Scales efficiently because labeled data is expensive and limited.
- Enables training on internet-scale datasets.

3. CLIP (Contrastive Language–Image Pretraining)

- Trained on 400 million image–text pairs collected from the internet.
- Uses natural language supervision instead of human-annotated labels.
- Learns a joint embedding space for images and text.
- Capable of zero-shot image classification across many tasks.
- Not a generative model; it does not produce open-ended outputs.

4. Embedding Models

- Transform raw data (text, image, audio) into numerical vectors.
- Embeddings capture semantic meaning and relationships.
- Similarity in embedding space reflects similarity in meaning.
- Multimodal embeddings align different modalities in the same vector space.

5. Generative Multimodal Models

- Built on top of multimodal embedding backbones.
- Can understand and generate across modalities.
- Examples include vision-language assistants and multimodal chat systems.
- Enable tasks like image understanding, captioning, and visual reasoning.

6. Shift from Task-Specific to General-Purpose Models

- Earlier ML models were trained for a single task only.
- Task transfer was difficult or impossible.
- Foundation models perform many tasks out of the box.
- Performance can be further optimized for specific tasks.

7. Model Adaptation Techniques

Prompt Engineering

- Carefully designed instructions and examples.
- No model training required.
- Fast and low cost but limited control.

Retrieval-Augmented Generation (RAG)

- Model retrieves relevant information from external sources.
- Improves factual accuracy and reduces hallucination.
- Keeps responses grounded in up-to-date data.

Fine-Tuning

- Further training on domain-specific datasets.
- Highest control and task performance.
- Higher cost compared to prompting and RAG.

8. AI Engineering

- Discipline focused on building applications using foundation models.
- Differs from traditional ML engineering which focuses on model training.
- Emphasizes prompt design, RAG pipelines, and system integration.

9. Drivers of the Rise of AI Engineering

- General-purpose AI capable of solving many tasks.
- Massive increase in enterprise and venture capital investment.
- Low entry barrier due to model-as-a-service APIs.
- AI enables rapid prototyping and faster time to market.

10. Foundation Model Use Cases

- Coding and software development assistance.
- Image and video generation and editing.
- Writing, summarization, and documentation.
- Education and tutoring systems.
- Conversational agents and copilots.
- Workflow automation and data processing.

11. Enterprise Adoption Patterns

- Preference for internal-facing applications.
- Lower risk compared to customer-facing systems.
- Common tasks include summarization, classification, and knowledge management.
- Closed-ended tasks are easier to evaluate and control.

12. Coding as the Most Popular Use Case

- AI significantly improves productivity for simple and repetitive tasks.
- High gains in documentation and code generation.
- Lower gains for complex system design and architecture.
- More effective for frontend than backend tasks.

13. Creative and Writing Applications

- AI excels at probabilistic and creative tasks.
- Widely used in marketing, design, and content creation.
- Writing tasks benefit due to high volume and error tolerance.
- Improves speed, clarity, and consistency of output.

14. Key Exam Takeaways

- Foundation models are scalable, adaptable, and general-purpose.
- CLIP proved large-scale multimodal self-supervision is effective.
- Embedding models underpin modern generative AI systems.
- Prompting, RAG, and fine-tuning are core AI engineering techniques.
- AI engineering is one of the fastest-growing engineering disciplines.