Chat Completion API Fundamentals

Understanding the Foundation of LLM Interactions

REQUEST → LLM SERVICE → RESPONSE

What Are Chat Completion APIs?

- HTTP-based interfaces for interacting with LLMs.
- Most LLMs follow OpenAI's API format as a de-facto standard.
- Use RESTful principles with JSON for requests and responses.
- Enable programmatic access to powerful language model capabilities.

Authentication & Headers

Requests must be authenticated using an API key sent in the headers.

```
Authorization: Bearer YOUR_API_KEY
Content-Type: application/json
```

🔑 Essential Concepts

- **Stateless:**Each API request is independent and self-contained.
- **Token Counting:**Costs are based on input + output tokens.
- **Rate Limits:**APIs have usage restrictions to ensure fair use.

Typical Request Components

```
{
  "model": "gpt-3.5-turbo",
  "messages": [
    {"role": "user", "content": "Hello!"}
  ],
  "temperature": 0.7,
  "max_tokens": 100
}
```

Typical Response Structure

```
{
  "id": "chatcmpl-abc123",
  "choices": [{
    "message": {
      "role": "assistant",
      "content": "Hello! How can I help?"
    }
  }],
  "usage": {
    "prompt_tokens": 10,
    "completion_tokens": 6,
    "total_tokens": 16
  }
}
```