Designing an Agent

Today you decide to get a new working dog on your digital farm, here are the decisions you need to make.

1) Defining the Job = Use Case

High level jobs might be...

- The Herding Dog (Data Processing Agent)
- The Retrieval Dog (Information Agent)
- The Guard Dog (Monitoring Agent)
- The Service Dog (Workflow Agent)

2) Commands = Tools/MCP

Some tricks people have templates for...

- Sequential Thinking for complex tasks
- Filesystem for local file system management
- Legacy Analysis for debugging COBOL
- Consilium for consensus driven generation

3) Choosing the Breed = Model Selection

A Labrador can be a working dog or family pet, breeding matters

- Llama 3 70B (Open Source) - Advanced reasoning for data transformation. Handles complex information & classification.
- NileChat-3B (Open Source) - Specialised for Egyptian & Moroccan Arabic, culturally aware.
- Qwen-2-1.5B (Open Source) High quality for its size; ideal for CPU inference and edge devices.
- Gemini 2.5 Pro (API) - Massive 2M token context for multi-source synthesis. Superior web search & document analysis integration.
- DeciLM 7B (Open Source) - Fast, efficient inference for real-time log analysis. Optimised for event-driven systems .
- Mixtral 8x7B (Open Source) - Expert at parsing instructions for multi-step processes. Strong logic and state management.

Note: easiest to think of "B" as "bigness" since more "B" means bigger memory, bigger compute power, bigger money to use and a bigger range of capabilities and applications within a single model (usually). "B" doesn't always mean "better" as they are expensive for narrow or repetitive tasks.

4) Memory & Recall = State Management

Dogs learn through repetition and experience, however with agents you can 'install' some memories.

Short-term Memory
(Working Memory)
- Like: Remembering the current session
- Technical: Session storage, caching
- Use Case: Context for an interaction

Long-term Memory
(Knowledge Retention)
- Like: Remembering last session
- Technical: Vector databases, RAG
- Use Case: Context for a user profile

Scent Memory
(Similarity Search)
- Like: A hunting dog tracking by scent
- Technical: vector search, anomalies
- Use Case: Finding related information

Hard Memory
(Structured data)
- Like: Knowing where treats are kept
- Technical: Rules, tables, transformers
- Use Case: Retrieval of unaltered data

5) Health monitoring = Observability

A neglected dog will eventually bite you or run away, so take care of your dogs.

Trials
(Evaluations)
- Simulations
- Accurate/Complete
- Divergence

Vital Signs
(Core Metrics)
- Response Time
- Success Rate
- Cost Efficiency

Behavior Monitoring
(Quality Metrics)
- Hallucination Check
- Tool Patterns
- Error Analysis

Performance Tracking
(Business Impact)
- Work Completed
- User Satisfaction
- ROI Measurement

Agent Definition: An AI agent is an autonomous system that perceives its environment through tools and data sources, reasons about tasks using an LLM as its "brain" and acts upon the environment using tools without constant human direction.

Model Context Protocol (MCP): servers are standardised connectors that provide AI assistants with access to external tools and data sources. They act as bridges between AI models and external systems. If you understand APIs, you know how these work.

Training/tuning: Models come in all shapes and sizes, often for specialised tasks or contexts they will perform better with fine tuning and in some cases more extensive machine learning. Pre-tuned models are often available for most scenarios.