Created by: Lamhot Siagian

AI Terminology for Software Testing — Beginner → Advanced

🌱 Foundations (Beginner)

Term	Definition
AI	Systems performing tasks that usually require human intelligence.
ML	Algorithms that learn patterns from data to make predictions.
Model	Mathematical function trained to map inputs to outputs.
Training / Inference	Learning parameters from data / using the model to predict.
Label	Ground-truth target used during supervised learning.
Bias / Variance	Underfitting vs. overfitting tendencies in a model.
Overfitting	Great on training data, poor on unseen data.
Generalization	Model's ability to perform well on new data.

🔤 NLP & LLM Basics

Term	Definition
NLP	Techniques for processing and understanding human language.
Tokenizer	Maps text to tokens (subwords/words) consumed by models.
Embedding	Vector representation of text/code capturing semantics.
Transformer	Neural architecture relying on self-attention.
LLM	Large Language Model trained to predict next tokens.
Context Window	Max tokens the model can consider in one request.
System/User/Tool Messages	Roles in chat prompting that guide behavior.
Temperature/Top-p	Sampling controls that trade off creativity vs. determinism.

🧪 AI for Testing — Core Ideas

Term	Definition
AI-Augmented Testing	Using AI to generate, execute, or evaluate tests.
Self-Healing Tests	Locators/steps auto-adjust when UI changes.
Test Case Generation	Model-driven synthesis of test steps/data.
Test Oracle	Mechanism to decide pass/fail (rule-based or LLM-based).
Spec-to-Test	Deriving tests from requirements/design/API schemas.
Exploratory AI	Agent navigates UI/API to discover defects.
Code Review w/ LLM	LLM suggests fixes, smells, and security issues.

📦 Data & Datasets

Term	Definition
Prompt Dataset	Collection of (input → expected) pairs for evaluation.
Gold Set	Curated, trusted labels for benchmarking.
Synthetic Data	Model-generated examples for coverage/edge cases.
Data Drift	Input distribution shifts over time.
Concept Drift	Target meaning/relationship changes over time.
Privacy Filters	Techniques to remove/obfuscate sensitive info.
PII	Personally Identifiable Information; requires protection.

📝 Prompting & Guardrails

Term	Definition
Prompt Engineering	Designing inputs to get reliable model outputs.
Chain-of-Thought (CoT)	Encouraging stepwise reasoning (often summarized in prod).
Few-Shot	Supplying exemplars to guide the model.
System Prompt	Top-level instructions that set role and constraints.
Guardrails	Policies/filters to block unsafe/irrelevant outputs.
Function/Tool Calling	LLM chooses structured tools (e.g., "click", "GET /users").
JSON Schema Output	Constrain responses to parseable formats.

📚 Retrieval & RAG

Term	Definition
RAG	Retrieval-Augmented Generation; model uses fetched context.
Vector Store	DB for embeddings enabling semantic search.
Chunking	Splitting docs into retrieval-friendly pieces.
Hybrid Search	Combine dense (vectors) + sparse (BM25) retrieval.
Hallucination	Confident but incorrect output; mitigated via RAG/grounding.
Attribution	Citations linking outputs to sources.
Context Relevance	How well retrieved chunks answer the query.

📈 Evaluation Metrics (Classical & LLM)

Metric	What it indicates
Accuracy / F1	Correctness balance for classification tasks.
BLEU / ROUGE	N-gram overlap with references (text gen).
METEOR / BERTScore	Semantic similarity beyond surface tokens.
Hallucination Rate	Frequency of unsupported claims.
Groundedness	Output supported by retrieved sources.
Task Success	End-to-end completion of user goal.
Latency / Throughput	Response time and requests/second.
Cost per Test	Spend per evaluation run or per pass/fail.

🛠️ Agents, Tools & Orchestration

Term	Definition
AI Agent	LLM + memory + tools + policy executing tasks autonomously.
Planner/Executor	Decompose tasks / perform steps via tools.
Memory	Short/long-term state used across steps/sessions.
Toolbox	APIs the agent can call (e.g., Playwright, Git, Jira).
ReAct	Reasoning + acting loop (think → act → observe).
MCP	Model Context Protocol; standardizes tool/server access.
Human-in-the-Loop	Manual approvals for risky steps.

⚙️ MLOps & CI/CD for AI Testing

Term	Definition
Experiment Tracking	Log prompts, seeds, metrics, artifacts.
Model Registry	Versioned storage for models & promotion stages.
Canary / A/B	Safely compare models/policies in prod.
Data Versioning	Track datasets & labels alongside code.
Prompt Versioning	Manage prompt variants as code.
Evaluation Pipeline	Automated offline/online tests gating release.
Observability	Telemetry: quality, drift, cost, latency, failures.

🧯 Risks, Safety & Compliance

Term	Definition
Safety Policy	Rules for harmful/unsafe content handling.
Red Teaming	Adversarial prompts to probe failures.
Jailbreak	Prompt designed to bypass safeguards.
Toxicity	Offensive/biased content requiring filtering.
Copyright/Attribution	Respect licensing, cite sources.
Explainability	Understanding model decisions (where possible).
Auditability	Reproducible logs of prompts, outputs, decisions.

🧬 LLM-Specific Testing Techniques

Technique	Purpose
Golden-Answer Checks	Exact/semantic match vs. expected outputs.
Reference-Free Eval	Judge quality without gold (LLM-as-Judge + heuristics).
Mutation Testing	Introduce small changes to test robustness.
Adversarial Prompts	Stress safety & instruction-following limits.
Determinism Harness	Fix seeds/temperature to compare runs.
Tool-Use Assertions	Verify correct tool selection & parameters.
Coverage Matrices	Map scenarios → prompts → metrics.

🛰️ Performance, Cost & Ops

Term	Definition
Tokens	Billing/latency unit; roughly word pieces.
Rate Limits	Throughput caps per minute/second.
Caching	Reuse responses for identical prompts.
Batching	Process multiple requests together when supported.
Distillation	Compress a large model into a smaller one.
Quantization	Lower-precision weights to speed & cut memory.
Latency Budgets	Allocations per step in an agent pipeline.

🧵 Vision/Code/Multimodal for QA

Term	Definition
Multimodal LLM	Accepts text+images (and possibly audio/video).
OCR	Extract text from screenshots/PDFs for assertions.
Vision Diffing	Compare UI renders pixel/feature-wise.
Code LLM	Model optimized for code generation/explanation.
AST Analysis	Parse code to trees for robust checks/refactors.
Screenshot Assertions	Image-based pass/fail for visual regressions.

📚 Acronyms & Shorthands

Acronym	Meaning
SLM/LLM	Small/Large Language Model
RLHF/RLAIF	Reinforcement Learning from Human/AI Feedback
LoRA/QLoRA	Parameter-efficient fine-tuning techniques
PEFT	Parameter-Efficient Fine-Tuning
BM25	Classic sparse text retrieval algorithm
ETL	Extract-Transform-Load (data pipeline)
SFT	Supervised Fine-Tuning
P95/P99	95th/99th percentile latency

🧭 Practical Mapping to QA Tools

Need	AI Term to Apply
Generate UI tests	Prompt Engineering, Few-Shot, Tool Calling (Playwright)
Stabilize flaky locators	Self-Healing, Embeddings for element similarity
Spec coverage	Spec-to-Test, Coverage Matrices, RAG over PRDs
Security hints	LLM Code Review, Policies, Red Teaming
Defect triage	RAG over tickets, Embeddings clustering
Release gating	Evaluation Pipeline, A/B, Canary
Cost control	Caching, Batching, Distillation, Quantization