In-Class Assignment AIF1: Hallucination Risk Scorecard

Digital Finance – BSc Course Prof. Dr. Joerg Osterrieder

Context. A mid-sized investment bank wants to deploy an LLM-based agent across six workflows: (1) **trade execution** (submit orders to exchange), (2) **KYC summary** (1-paragraph client risk profile from docs), (3) **code-gen** (write unit tests for pricing library), (4) **email drafting** (reply to client queries), (5) **portfolio rebalance** (propose buy/sell list to meet target weights), (6) **market research** (weekly sector report). The bank needs a hallucination-tolerance design for each.

Q1. Classify each of the 6 tasks as **LOW / MEDIUM / HIGH** hallucination tolerance. Justify in one line.

Solution. (1) Trade exec **LOW** — a wrong ticker = immediate $-loss and exchange fines. (2) KYC summary **LOW** — mis-stating a sanctions flag = regulator enforcement. (3) Code-gen **HIGH** — CI/tests catch errors before prod. (4) Email drafting **MEDIUM** — a human reviewer reads before send. (5) Portfolio rebalance **LOW** — wrong weights = tracking error + fiduciary breach. (6) Market research **MEDIUM** — hallucinated statistics reach analysts who should cross-check. Rule of thumb: **LOW** if output acts directly (money or compliance), **HIGH** if output is verified by a deterministic downstream system.

Q2. For the 3 **LOW-tolerance** tasks, design a **human-in-the-loop (HITL)** gate and one **deterministic guardrail**. One sentence per task.

Solution. Trade exec: HITL = trader confirms any order > $500k on a 4-eyes screen; guardrail = hard $ + notional limits per ticker + circuit breaker on >2% NAV. **KYC summary**: HITL = compliance officer signs off summary before filing; guardrail = RAG-only retrieval forces citations to source docs, and a regex check blocks unsupported claims (no ticker / name outside retrieved corpus). **Portfolio rebalance**: HITL = PM approves the order ticket; guardrail = weight-sum-to-1 assertion + pre-trade compliance check + 10% turnover cap per day. Acceptable alternatives: dual-agent cross-check (executor + critic), deterministic calculation module separate from LLM.

Q3. The bank's CISO argues that any LOW-tolerance task should be *deterministic only* (no LLM, ever). Give **one argument for, one against**.

Solution. For: a deterministic pipeline has a provable error rate — LLM hallucination is open-ended and not reducible below $\approx 0.5\%$ even with 2026-era models; for fiduciary/regulated actions, "best-effort" is not acceptable. **Against**: LLMs catch edge cases deterministic rules miss (unusual KYC phrasing, ambiguous client instructions); a hybrid with HITL + guardrail has a lower *combined* error rate than either alone. Reasonable balanced answer: use LLM for *proposal* + deterministic for *execution* — the agent drafts, the human and the rule-engine dispatch.