In-Class Assignment AIF2: Agent Cost Model

Digital Finance – BSc Course Prof. Dr. Joerg Osterrieder

Context. A retail bank runs **10,000 customer chats per day**. Each chat averages **8 turns** with mean **1,500 input tokens + 400 output tokens** per turn. Human agents are paid \$25/hour fully-loaded and handle **4 chats/hour**. The bank wants to deploy a GPT-class agent at **80% automation** (80% of chats handled end-to-end by the agent; 20% escalated to a human). Assume 2026 frontier-model pricing: \$3 per 1M input tokens, \$12 per 1M output tokens. One escalation consumes a full human chat.

Q1. Compute the **daily token cost** of running the agent on 10k chats. (All 10k chats go through the agent; 2k are then handed off, but the agent still processed them first.)

Solution. Per chat: $8 \times 1{,}500 = 12{,}000$ input tokens, $8 \times 400 = 3{,}200$ output tokens. Per-chat cost = $(12{,}000/10^6) \times \$3 + (3{,}200/10^6) \times \$12 = \$0.036 + \$0.0384 = \$0.0744$. **Daily token cost** = $10{,}000 \times \$0.0744 = \$744$**/day**, i.e. $\approx \$272k$/year. Key insight: frontier LLM inference for 10k chats/day is <\$1k/day – the bottleneck is not cost but accuracy.

Q2. Compute the **daily human-cost saving**. Baseline (no agent): 10k chats / 4 per hour = 2,500 agent-hours $\times$ \$25 = \$62,500/day. With 80% automation, humans handle 2k chats. Compare fully-loaded costs and report the **break-even automation rate** at which the agent stops saving money.

Solution. With agent: humans handle $2{,}000/4 = 500$ hours = \$12,500 + \$744 tokens = \$13,244/day. **Saving** = \$62,500 − \$13,244 = \$49,256**/day**, i.e. \$12.3M/year. Break-even: the agent costs \$744/day regardless. Let a = automation rate; human cost = $(1-a) \cdot 10{,}000/4 \cdot \$25 = \$62{,}500(1-a)$. Agent breaks even when $\$62{,}500(1-a) + \$744 = \$62{,}500$, so $a \approx 1.2\%$. **The agent pays for itself at just 1.2% automation** – virtually any deflection rate is profitable, which is why cost is rarely the binding constraint.

Q3. Name **2 hidden costs** not in this model that materially change the TCO. One sentence each.

Solution. (i) **Evaluation + red-team staffing**: a regulated-industry agent needs ≈2–5 FTE ML-ops + prompt engineers to maintain eval harnesses, run adversarial tests, and gate model upgrades – easily \$400–800k/year, 2× the token bill. (ii) **Compliance + legal**: bank-grade chatbots in the EU need MiCA/DORA + GDPR Article 22 review, audit logging for every LLM call (typically \$50k–200k/year in SIEM + storage), and human appeal rights on any automated adverse decision. Acceptable alternatives: (iii) fine-tuning + vector-store infra, (iv) SLA premiums for low-latency hosted inference, (v) brand-risk insurance covering hallucination-caused fines.