

אופציה 1 לתרגיל בית מספר 5

תכנון ניסוי: חלונות הקשר בפרקטיקה

Lab: Context Windows in Practice

סטודנט

© Dr. Segal Yoram - כל הזכויות שמורות

November 2025

גרסה 1.0

תוכן העניינים

3	1	מבוא: Introduction
3	2	מטרות הניסוי: Lab Objectives
3	3	ניסוי 1: המחט בערמת השחת: Experiment 1: Needle in Haystack
3	3.1	פרטי הניסוי: Experiment Details
3	3.2	הנתונים: Data
4	3.3	קוד פסאודו: Pseudocode
4	4	ניסוי 2: השפעת גודל חלון ההקשר: Experiment 2: Context Window Size Impact
4	4.1	פרטי הניסוי: Experiment Details
4	4.2	הנתונים: Data
5	4.3	קוד פסאודו: Pseudocode
5	5	ניסוי 3: השפעת RAG: Experiment 3: RAG Impact
5	5.1	פרטי הניסוי: Experiment Details
5	5.2	הנתונים: Data
6	5.3	קוד פסאודו: Pseudocode
6	5.4	תוצאות צפויות: Expected Results
6	6	ניסוי 4: אסטרטגיות ניהול הקשר: Experiment 4: Context Engineering Strategies
6	6.1	פרטי הניסוי: Experiment Details
7	6.2	הנתונים: Data
7	6.3	קוד פסאודו: Pseudocode
7	7	טבלה ריכוזית: Summary Table
8	8	סיכום: Summary
8	9	הנחיות להגשה: Submission Instructions

1 מבוא: Introduction

מטלה זו עוסקת בניתוח ולימוד של חלון ההקשר Context Windows. שורת הניסויים המוצעת להלן מהווה מסגרת רעיונית כללית, ואתם מוזמנים לפרש, לפתח ולחקור את הנושאים בכל דרך שתמצאו לנכון. עבור כל ניסוי עליכם להגדיר שאלות מחקר, לבצע את הניסויים ולנתח את הממצאים, רצוי תוך הצגת ניתוח סטטיסטי ויזואלי (באמצעות גרפים או טבלאות). מומלץ לחזור על כל ניסוי מספר פעמים כדי להבטיח תוקף סטטיסטי לתוצאה.

שימו לב: המסקנות שלכם אינן חייבות לחפוף לחומר שהוצג בשיעור. אתם רשאים להגיע לתובנות עצמאיות, ובלבד שתנמקו אותן היטב; במקרים אלו מומלץ להיעזר בסימוכין חיצוניים ולהציע הסבר לפערים שגיליתם. קחו את הניסויים למקום שמעניין אתכם ולכיוון החקירה האישי שלכם – ההנחיות הללו נועדו לשמש כ'סיעור מוחות' ואינן בגדר הגדרות סגורות.

2 מטרת הניסוי: Lab Objectives

הסטודנט יבין ויחווה בחיי-קוד שתי תופעות מרכזיות:

1. **Lost in the Middle** - ירידה בדיוק כאשר מידע רלוונטי נמצא באמצע חלון ההקשר
2. **Context Accumulation Problem** - איך הנתונים מצטברים בסוכנים וגורמים לכישלון הניסוי מחולק לארבעה תת-ניסויים מודולריים, כל ניסוי הוא עצמאי אך בונה על הקודם.

3 ניסוי 1: המחט בערמת השחת: Experiment 1: Needle in Haystack

3.1 פרטי הניסוי: Experiment Details

- משך: כ-15 דקות
- רמת קושי: בסיסי
- מטרה: הדגמה של Lost in the Middle

3.2 הנתונים: Data

- טקסט סינתטי: 5 מסמכים, כל אחד עם 200 מילים
- כל מסמך מכיל עובדה אחת קריטית (למשל: "מנכ"ל החברה הוא דוד כהן")
- העובדה תופיע במיקומים שונים: התחלה / אמצע / סוף

Experiment 1: Lost in the Middle Simulation

```
# Generate synthetic documents with embedded facts
def create_documents(num_docs=5, words_per_doc=200):
    documents = []
    for i in range(num_docs):
        doc = generate_filler_text(words_per_doc)
        fact_position = random.choice(['start', 'middle', 'end'])
        doc = embed_critical_fact(doc, fact, fact_position)
        documents.append(doc)
    return documents

# Query LLM and measure accuracy by fact position
def measure_accuracy_by_position(documents, query):
    results = {'start': [], 'middle': [], 'end': []}
    for doc in documents:
        response = ollama_query(doc, query)
        accuracy = evaluate_response(response, expected_answer)
        results[doc.fact_position].append(accuracy)
    return calculate_averages(results)

# Expected: High accuracy at start/end, low at middle
```

תוצאה צפויה: דיוק גבוה בהתחלה/סוף, נמוך באמצע.

4 ניסוי 2: השפעת גודל חלון ההקשר: Experiment 2: Context Window Size Impact

4.1 פרטי הניסוי: Experiment Details

- משך: כ-20 דקות
- רמת קושי: בינוני
- מטרה: הדגמה של how context window size affects accuracy

4.2 הנתונים: Data

- הגדלה הדרגתית של מספר המסמכים: 2, 5, 10, 20, 50
- לכל גודל: מדידת זמן תגובה + דיוק + אורך הקשר בפועל

Experiment 2: Context Window Size Analysis

```

# Measure performance across different context sizes
def analyze_context_sizes(doc_counts=[2, 5, 10, 20, 50]):
    results = []
    for num_docs in doc_counts:
        documents = load_documents(num_docs)
        context = concatenate_documents(documents)

        start_time = time.time()
        response = langchain_query(context, query)
        latency = time.time() - start_time

        results.append({
            'num_docs': num_docs,
            'tokens_used': count_tokens(context),
            'latency': latency,
            'accuracy': evaluate_accuracy(response)
        })
    return results

# Plot: Accuracy degradation vs context size
# Expected: Accuracy decreases as window grows

```

תוצאות: גרף שמראה את ירידת הדיוק עם הגדלת החלון.

5 ניסוי 3: השפעת RAG: RAG Impact**5.1 פרטי הניסוי: Experiment Details**

- **משך:** כ-25 דקות
- **רמת קושי:** בינוני+
- **מטרה:** השוואה בין שתי אסטרטגיות:
- ללא RAG: כל המסמכים בחלון
- עם RAG: רק המסמכים הרלוונטיים (באמצעות similarity search)

5.2 הנתונים: Data

- מאגר של 20 מסמכים בעברית (נושאים: טכנולוגיה, משפט, רפואה)
- שאילתה: "מה הם התופעות הלוואי של התרופה X?"

5.3 קוד פסאודו: Pseudocode

Experiment 3: RAG vs Full Context Comparison

```
# Step 1: Chunking - split documents into chunks
chunks = split_documents(documents, chunk_size=500)

# Step 2: Embedding - convert to vectors
embeddings = nomic_embed_text(chunks)

# Step 3: Store in ChromaDB
vector_store = ChromaDB()
vector_store.add(chunks, embeddings)

# Step 4: Compare two retrieval modes
def compare_modes(query):
    # Mode A: Full context (all documents)
    full_response = query_with_full_context(all_documents, query)

    # Mode B: RAG (only similar documents)
    relevant_docs = vector_store.similarity_search(query, k=3)
    rag_response = query_with_context(relevant_docs, query)

    return {
        'full_accuracy': evaluate(full_response),
        'rag_accuracy': evaluate(rag_response),
        'full_latency': full_response.latency,
        'rag_latency': rag_response.latency
    }

# Expected: RAG = accurate & fast, Full = noisy & slow
```

5.4 תוצאות צפויות: Expected Results

- RAG: תשובות מדויקות ומהירות

- Full Context: רעש וסבל, תשובות פחות מדויקות

6 ניסוי 4: אסטרטגיות ניהול הקשר: Experiment 4: Context Engineering Strategies

6.1 פרטי הניסוי: Experiment Details

- משך: כ-30 דקות

- רמת קושי: מתקדם

- מטרה: בחינת אסטרטגיות ניהול הקשר (Write, Select, Compress, Isolate)

6.2 הנתונים: Data

- סימולציה של סוכן רב-צעדי שמבצע 10 פעולות עוקבות
- כל פעולה יוצרת output שמתווסף להקשר

6.3 קוד פסאודו: Pseudocode

Experiment 4: Context Engineering Strategies

```
# Strategy 1: SELECT - Use RAG for relevant retrieval only
def select_strategy(history, query):
    relevant = rag_search(history, query, k=5)
    return query_llm(relevant, query)

# Strategy 2: COMPRESS - Automatic history summarization
def compress_strategy(history, query):
    if len(history) > MAX_TOKENS:
        history = summarize(history)
    return query_llm(history, query)

# Strategy 3: WRITE - External memory (scratchpad)
def write_strategy(history, query, scratchpad):
    key_facts = extract_key_facts(history)
    scratchpad.store(key_facts)
    return query_llm(scratchpad.retrieve(query), query)

# Compare all strategies across 10 sequential actions
def benchmark_strategies(num_actions=10):
    results = {'select': [], 'compress': [], 'write': []}
    for action in range(num_actions):
        output = agent.execute(action)
        history.append(output)
        for strategy in ['select', 'compress', 'write']:
            result = evaluate_strategy(strategy, history)
            results[strategy].append(result)
    return results
```

7 טבלה ריכוזית: Summary Table

טבלה 1: סיכום הניסויים: Experiments Summary

ניסוי / Exp	נושא / Topic	כלים / Tools	זמן / Time	תפוקה / Output
1	Lost in Middle	Ollama + Python	תוקד 15	Accuracy by position graph
2	Context Size	Ollama + LangChain	תוקד 20	Latency vs size graph
3	RAG Impact	Ollama + Chroma	תוקד 25	Performance comparison
4	Engineering	LangChain + Memory	תוקד 30	Strategy performance table

8 סיכום: Summary

ניסויים אלו מדגימים את האתגרים המרכזיים בעבודה עם חלונות הקשר גדולים:

1. **בעיית ה-Lost in the Middle:** מידע באמצע החלון נוטה להיאבד
2. **גודל חלון ההקשר:** ככל שהחלון גדל, הדיוק יורד
3. **יעילות RAG:** אחזור ממוקד משפר דיוק ומהירות
4. **אסטרטגיות ניהול:** Select, Compress, Write מספקות פתרונות שונים

9 הנחיות להגשה: Submission Instructions

על הסטודנט לתכנן ולחשוב באיזה אופן משכנע להציג את תוצאות הניסוי החקר של הניסוי, המסקנות מהניסוי. מומלץ לתקף את התוצאות בגרפים לפי שיקול דעת הסטודנט.

הערה: האמור במסמך זה מיועד לנשים וגברים כאחד, והשימוש בלשון זכר הוא מטעמי נוחות בלבד.