Практичне завдання №10: **Експерименти з параметрами моделі** (Temperature, Frequency Penalty) **у** Playground

Мета: Навчитись практично використовувати Google AI Studio (або інший Playground) для керування поведінкою LLM через зміну ключових параметрів API; зрозуміти, як Temperature впливає на креативність, а Frequency Penalty — на повторюваність тексту.

Ключові концепції: API, Playground, Temperature, Frequency Penalty, Top-P, Top-K, Max Tokens, детермінізм.

Завдання 1: **Дослідження "креативності"** (Temperature)

Ідея: Емпірично з'ясувати, як параметр Temperature («температура») впливає на детермінізм (стабільність) та креативність (випадковість) відповідей моделі.

Інструменти: Google AI Studio (рекомендовано, оскільки є безкоштовним та легким для старту) або OpenAI Playground (чи інша зручна для вас «пісочниця»).

Завдання:

1. Зайдіть у Google AI Studio та створіть новий промпт ("Create new" -> "Freeform prompt").
2. Використовуйте один і той самий промпт:

   «*Напиши 3 унікальні ідеї для стартапу у сфері еко-технологій.*»

Крок А (**низька температура** – **режим** «**Точність**»):

1. На панелі налаштувань праворуч встановіть Temperature на 0.1 (або максимально близьке до 0 значення).
2. Натисніть «Run» 5 **разів поспіль**.
3. Скопіюйте та зафіксуйте всі 5 отриманих відповідей.

Крок Б (**висока температура** – **режим** «**Креативність**»):

1. На панелі налаштувань праворуч встановіть Temperature на 1.0 (або 0.9).
2. Натисніть «Run» 5 **разів поспіль** (з тим самим промптом).
3. Скопіюйте та зафіксуйте всі 5 отриманих відповідей.

Оцініть:

1. **Порівняйте** 5 **відповідей з кроку А** (T=0.1): Наскільки вони схожі між собою? Чи є серед них однакові ідеї?
2. **Порівняйте** 5 **відповідей з кроку Б** (T=1.0): Наскільки вони різноманітні? Чи є серед них більш «неочікувані» або «дивні» ідеї?
3. **Зробіть висновок**: Для якого типу завдань ви б використовували низьку температуру (напр., вилучення фактів, написання коду, класифікація), а для якого — високу (напр., мозковий штурм, написання вірша, генерація варіантів)?

Завдання 2: **Боротьба з повторами** (Frequency Penalty)

Ідея: Зрозуміти, як параметр Frequency Penalty («штраф за частоту») «штрафує» модель за повторення одних і тих самих слів (токенів) у відповіді.

Завдання:

Крок А (**провокація зациклення** (**без штрафу**)):

1. У Google AI Studio встановіть Temperature на 0.7 (для балансу).
2. Встановіть Frequency Penalty на 0.0.
3. Використайте промпт, що провокує повтори:
   «*Напиши довгий абзац про важливість інновацій. Повторюй слово "інновація" якомога частіше, щоб підкреслити, наскільки важлива інновація для сучасної компанії. Інновація — це ключ до успіху.*»
4. Запустіть промпт і зафіксуйте результат (ви маєте побачити багато повторів слова «інновація»).

Крок Б (**застосування штрафу**):

1. Не змінюючи промпт та температуру, встановіть Frequency Penalty на 2.0 (максимальне значення).
2. Запустіть промпт ще раз.
3. Зафіксуйте результат.

Оцініть:

1. Що сталося з текстом у Кроці Б?

2. Як саме модель уникла повторення слова «інновація»? (Чи замінила вона його синонімами? Чи просто почала писати про інше, уникаючи цього слова?)
3. В яких випадках цей параметр може бути корисним, окрім боротьби з очевидним зацикленням? (Напр., для написання більш «багатого» та стилістично різноманітного тексту).

Завдання для СРС: **Дослідження локальних** Open-Source LLM (**для просунутих**)

Ідея: Отримати досвід встановлення та взаємодії з LLM, яка працює виключно на вашому комп'ютері, та зрозуміти її ключові переваги (конфіденційність) та обмеження (база знань).

Інструменти: LMStudio (або Ollama, Jan).

Завдання:

1. **Встановлення**: Завантажте та встановіть LMStudio з офіційного сайту.
2. **Завантаження моделі**: Через інтерфейс програми (вкладка пошуку) знайдіть та завантажте одну з невеликих, але потужних open-source моделей. Рекомендовані варіанти (наприклад):
   - https://lmstudio.ai/models/microsoft/phi-4-mini-reasoning
   - https://lmstudio.ai/models/qwen/qwen3-4b-thinking-2507
   - https://lmstudio.ai/models/deepseek/deepseek-r1-0528-qwen3-8b
   - https://lmstudio.ai/models/mistralai/mistral-nemo-instruct-2407
3. **Офлайн-експеримент**: Перейдіть на вкладку чату (💬), виберіть завантажену модель. **Вимкніть Інтернет на вашому комп'ютері** (вимкніть Wi-Fi/Ethernet), щоб переконатися, що модель працює 100% локально.
4. **Проведіть серію тестів з моделлю**:
   - **Тест** 1 (**обмеження знань**): Задайте питання про нещодавню подію (напр., «*Який фільм виграв Оскар за найкращий фільм цього року?" або "Який курс гривні до долара сьогодні?*»).
   - **Тест** 2 (**конфіденційність**): Напишіть промпт: «*Я надаю тобі конфіденційний текст моєї компанії для аналізу. Ця інформація не повинна залишати мій комп'ютер. [Вигаданий конфіденційний текст]. Знайди в ньому три ключові ризики.*»
   - **Тест** 3 (**швидкість та якість**): Попросіть модель виконати креативне завдання (написати вірш) та логічне завдання (вирішити просту задачу). Зверніть увагу на швидкість генерації.

1. **Напишіть звіт** (1-2 **сторінки**):
4. Яке ПЗ та яку модель ви використовували?
1. Як модель відреагувала на **Тест** 1 (**Обмеження знань**)? Чому вона дала таку відповідь?
2. Які ваші враження від **Тесту** 2 (**Конфіденційність**)? Які фундаментальні переваги дає локальна модель у цьому контексті порівняно з хмарними сервісами?
3. Опишіть свої загальні враження: чим досвід взаємодії з локальною LLM відрізняється від використання ChatGPT або Gemini? Які ви бачите плюси (напр., конфіденційність, відсутність цензури, робота офлайн) та мінуси (напр., якість відповідей, вимоги до "заліза", обмеженість знань)?