澳門大學

協同創新研究院

協同

人文學院


澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

# 基於語言學奧賽的大語言模型推理性能評估與基準開發

# Evaluation and Benchmark Development of LLM's Reasoning Performance by Linguistic Olympiad

**仇瀟琳 MC365037**

理學碩士學位（數據科學—計算語言學）課程
專案報告

專案導師

袁毓林教授

2025 年 3 月 27 日

# 聲明

本人所提交的論文， 除了經清楚列明來源出處的資料外，其他內容均為原創；本論文的全部或部分未曾在同一學位或其他學位中提交過。

本人聲明已知悉及明白《澳門大學學生學術誠信處理規條》及《澳門大學學生紀律規章》。

簽名 : 仇瀟琳

學生姓名 : 仇瀟琳

學生編號 : MC365037

日期 : 2025 年 3 月 28 日

## 摘要

各種規模的大語言模型的頻繁發佈，為大語言模型的評測研究帶來了機遇與挑戰，目前，針對通用領域大語言模型的評測體系日趨成熟，而面向垂直領域的大語言模型評測仍在起步階段。世界上許多語言缺乏大規模語料庫，來訓練高質量的大型語言模型（LLMs），現有的 LLMs 基本無法理解甚至處理瀕危語言。因此，大型語言模型如何處理瀕危語言是一個重要且亟待解決的問題。缺乏針對瀕危語言的相關評測數據集，這種局面限制了對 LLMs 在極低資源語言上能力的系統性評估。儘管前人研究關注了低資源語言，但尚未深入探討極低資源語言，尤其是那些缺乏完整的語法描寫體系的瀕危語言。為彌補這一空白，我們聚焦於極低資源語言，旨在探索大模型在此類語言上的應用潛力。通過語言學奧林匹克競賽題目，我們評估了 GPT-4、deepseek、通義千問和智譜清言等大模型在處理、推理和翻譯多種極低資源語言方面的能力，並研究了提升模型表現的方法。研究結果表明，少樣本學習是 LLMs 處理和推理極低資源語言的關鍵。在瀕危語言甚至缺乏完整的語法描寫體系的情況下，少樣本對是激發語言價值的重要方法，為 LLMs 在極低資源語言上的應用提供了新的可能性。



**關鍵字：** 大模型 ；評測基準 ；極低資源語言；邏輯推理



## Abstract

The frequent release of large language models (LLMs) of varying scales has brought both opportunities and challenges to LLM evaluation research. Currently, the evaluation framework for general-domain LLMs is becoming increasingly mature, while assessments of vertical-domain LLMs remain in their early stages. Many languages worldwide lack large-scale corpora for training high-quality LLMs, and existing models are generally unable to understand or even process endangered languages. As a result, how LLMs handle endangered languages is a crucial and pressing issue. The lack of dedicated evaluation datasets for endangered languages limits the systematic assessment of LLM capabilities in extremely low-resource languages. Although prior research has examined low-resource languages, there has been little in-depth exploration of extremely low-resource languages — particularly endangered ones that lack comprehensive grammatical documentation. To address this gap, we focus on extremely low-resource languages,

aiming to explore the potential of large models in processing such languages. Using problems from the Linguistics Olympiad, we evaluate the capabilities of GPT-4, DeepSeek, Qwen (Tongyi Qianwen), and GLM (Zhipu Qingyan) in handling, reasoning, and translating multiple extremely low-resource languages, while also investigating methods to improve model performance. Our findings demonstrate that few-shot learning is key to enabling LLMs to process and reason in these languages. For endangered languages that lack complete grammatical systems, few-shot pairs serve as a vital method for unlocking linguistic value, offering new possibilities for applying LLMs to extremely low-resource languages.

**Key words:** Large Model; Evaluation Benchmark; Extremely Low-resource Language; Logical Reasoning

**目錄**

## 1 引言:語言大模型的發展與評估方面的挑戰

2022 年 11 月，OpenAI 發佈的大型語言模型 ChatGPT 的出現，為整個人工智慧技術領域帶來了前所未有的變革，推動著 AI 技術正式進入了一個嶄新的歷史階段。近年來，生成式語言大模型（LLMs）在自然語言處理領域取得了突破性進展，展現出驚人的文本生成、翻譯、問答等能力。然而，大模型的輸出具有不確定性，並且伴隨著模型規模和能力的不斷提升，當前大語言模型性能評估也面臨著前所未有的挑戰。大模型的實際效能能否達到預期目標，還需要更加系統且精准的評估體系以及持續的優化迭代。

與構建通用領域大模型相比，垂直領域的大語言模型構建更具有性價比。語言學奧林匹克競賽（IOL）是一項旨在考察參與者邏輯思維和語言分析能力的國際賽事，其題目通常涉及對陌生語言現象的推理和歸納，能夠有效評估個體的邏輯推理能力。為更好地將極低資源語言與大語言模型技術相結合，提高極低資源語言在大模型中的應用價值，亟需針對極低資源語言的大語言模型評測標準，為極低資源語言領域大語言模型評測提供參考。有鑒於此，本文研究怎樣利用語言學奧賽試題就大語言模型對語言結構的邏輯推理能力，對當前主流的語言大模型進行評測。通過測評得到當前對瀕危語言處理最為優異的大語言模型，為針對瀕危語言垂直領域大語言模型訓練的基線模型選取提供依據。

### 1.1 大語言模型推理能力現狀

Brown 等人（2020）提出，通過在大型文本語料庫上進行預訓練，再針對特定任務進行微調，可以在許多自然語言處理（NLP）任務和基準測試中取得顯著提升。研究指出 GPT-3 在眾多自然語言處理（NLP）任務中表現出色，包括翻譯、問答和完形填空等任務；同時也能夠出色完成需要即時推理或領域適應的複雜任務，例如單詞重組、在句子中使用新詞，甚至進行三位數的算術運算。近年來，大型語言模型（LLMs）和大型多模態模型（LMMs）取得了顯著進展，在多項任務中超越了普通人類的能力，並在多個領域達到了接近人類專家的熟練水準。2024 年，He 等人（2024）對 GPT-4 與其前身 GPT-3 進行了比較；結果表明，GPT-4 在模型規模上實現了顯著提升（參數量超過一萬億），並在多語言處理能力、上下文理解以及推理能力等方面全面超越了 GPT-3。

大模型在結構化問題推理方面，針對數學領域，無論是僅限於少量計算的純文本形式的推理問題，還是融合數學邏輯和背景知識的競爭水準問題，這些具有挑戰性的數據集正越來越多地被克服 (Cobbe et al., 2021; Lu et al., 2023; Zhou et al., 2023 )。針對物理領域，無論是物理知識的多步驟推理，還是基於定理的問題回答的數據集也逐漸被構建 (Arora et al., 2023; Hendrycks et al., 2021; )。

Bowen 等人(2024)對當前大語言模型的歸納推理能力進行了全面評估。他們認為，僅考慮規則的歸納過於狹隘且不現實，因為歸納推理通常與其他能力（如規則應用、結果/規則驗證和更新資訊整合）混合在一起。研究發現即便是最先進的大語言模型在執行簡單的推理任務時也會失敗，這表明大語言模型在執行這些任務時存在局限性。這些局限性在需要高度結構化推理的學術競賽中尤為突出，而國際奧林匹克競賽題目恰好為驗證大模型的複雜推理能力提供了測評數據集來源。

## 1.2 大模型在奧賽題上的測評研究

國際奧林匹克競賽（如數學、物理、化學、生物學、資訊學及語言學等）是全球範圍內旨在發掘和培養青少年學術潛力的高層次學術賽事。其核心目標是通過競技形式激發學生對科學及人文領域的興趣，促進跨學科思維與創新能力的培養。數學奧賽題通常強調抽象推理與邏輯嚴謹性，要求參賽者在有限時間內構建巧妙的證明或解決複雜的組合問題；物理奧賽題則注重理論與實驗的結合，考察學生對物理原理的深刻理解及其在現實情境中的應用能力；語言學奧賽題則通過分析陌生語言的結構與規則，考察參賽者的邏輯推理、模式識別及跨文化理解能力。

Trinh 等人(2024)提出了一個神經符號系統 AlphaGeometry，使大模型的表現接近國際數學奧林匹克競賽（IMO）金牌得主的平均水準。更值得注意的是，AlphaGeometry 能生成人類可讀的證明，經由人類專家評估，它能解決 2000 年和 2015 年 IMO 中的所有幾何問題。2024 年，He 等人(2024)推出了奧林匹克級別雙語多模態的科學基準測試 OlympiadBench，包含來自奧林匹克級別數學和物理競賽（包括中國的高考）的 8476 道題目。實驗表明，開源語言模型在數學和物理學領域發展迅猛，構建的物理奧賽的題目與先前的測試集相比，難度不斷加大，題型不斷豐富，題量不斷增加，為該領域在複雜性和多樣性方面樹立了新的標杆。但語言學奧賽題目因其對陌生語言規則的歸納與演繹特性，對大模型提出了截然不同的挑戰——這類任務不僅需要符號推理能力，更依賴對極低資源語言結構的跨文化解讀能力，而當前研究尚未涉足這一空白領域。

## 1.3 大模型在極低資源語言方面的評測現狀

Moseley (2010) 指出雖然像英語或西班牙語這樣的語言擁有大量可供獲取的數據，但世界上 7000 多種語言中的大多數卻缺乏豐富的語料庫，包括聯合國教科文組織所認定的大多數瀕危語言。Zhang 等人(2024)針對瀕危語言提出了一種無需訓練的方法 LINGOLLM，在 GPT-4 和 Mixtral 這兩個模型之上實現了 LINGOLLM，並在 8 種瀕危或資源匱乏的語言上對其在 5 個任務中的表現進行了評估，旨在讓語言大模型（LLM）

能夠處理其預訓練數據中罕見出現的語言。實驗說明，在語言大模型時代，語言知識對於瀕危語言具有巨大的價值。

這些前期探索為大模型邏輯推理能力的研究奠定了基礎，為本文對大型語言模型針對極低資源語言的評測提供了極其重要的理論參考。在此基礎上，我們利用語言學奧賽題目構建有關大模型對極低資源語言推理能力的評測基準，以填補目前大模型在極低資源語言評測領域的空白，我們希望這項工作將有利於提升大語言模型在極低資源語言推理方面的能力。

## 2 研究意義及主要內容

綜前§1所述，面對當今大模型技術日新月異的發展態勢，系統全面地構建多元且創新的評測基準對於衡量大型語言模型性能具有重要價值。在這一背景下，本文聚焦語言大模型在極低資源語言方面的評測空白，旨在全面且系統地探索如何有效測評大型語言模型在極低資源語言邏輯推理能力上的優劣。

本文的主要研究內容圍繞以下幾個方面展開：

(1)構建大模型對極低資源語言評測基準框架：基於既有大型語言模型各方面推理能力的研究成果，構建一項針對極低資源語言的評測基準，檢驗大型語言模型對極低資源語言理解及推理能力；

(2)進行大型語言模型性能評測：對當前具有代表性大型語言模型進行系統測評，通過測試全面地評估其在極低資源語言上的語言結構的理解推理能力；

(3)基於評測數據的深度分析與歸因研究：根據實驗所得出的測評數據與結果，分析大型語言模型在極低資源語言邏輯推理方面的表現，並探究造成這些問題的潛在原因。

本研究對大語言模型針對極低資源語言的邏輯推理任務中的性能優化具有重要意義。因為我們利用語言學奧賽試題構建了首個針對極低資源語言的評測基準，同時也系統地評估了大型語言模型在極低資源語言邏輯推理方面的表現，這對大型語言模型在極低資源語言推理能力上的優化提升提供重要的參考。本文的評測代碼將上傳至 Github 平臺供 AI 社區共用。

## 3 測試集概覽

本節主要闡述說明了測試集數據來源，詳細展示了測評的具體格式。

### 3.1 數據集來源

本研究採用的極低資源語言素材，來自國際語言學奧林匹克競賽（IOL）官方發佈

的 2013-2024 年真題中的典型題目。IOL 要求參賽者分析不同語言的語法、結構、文化和歷史，並通過語言分析和解決問題來展示他們的語言能力。該數據集集成了廣泛且多樣化的語言樣本，涵蓋等多種瀕危語言，並且設計了英文、中文、俄語三種高資源語言版本和羅馬尼亞語、愛沙尼亞語、斯洛文尼亞語三種低資源版本。

必須說明的是，本研究未對 IOL 題目庫題目內容進行任何形式的篡改或修飾，僅對各類語言試題進行了必要的整理。得益於此，IOL 題目為本研究構建一個高質量的極低資源語言測評數據集提供了堅實的測評數據支撐。

### 3.2 任務與問題格式

鑒於語言學奧賽題目設置和大語言模型現狀，本研究選擇了其中的源語言與目標語言的翻譯題目，專注於構建一個集中針對極低資源語言邏輯推理的評測數據集。

該任務核心在於測試大模型對極低資源語言理解及推理能力，要求模型能夠推理陌生的極低資源語言的句子結構和邏輯，從而推斷出翻譯結果。

本測評數據集設計了一系列基於推理的翻譯題例，提問格式如下（以 2014 年中文題目第一題為例）：


{
"source": ["第十二屆國際語言學奧林匹克競賽"],
"location": "中國北京",
"date": "2014 年 7 月 21-25 日",
"name": "個人賽題目",
"problem": {
"problem_1": {
"title": "第一題",
"score": 20,
"instructions": " 以下為貝納貝納語的一些動詞形式及其漢語翻譯:\n
nohobe = 我在打他 \n
kahalune = 我們將打你 \n
nokoho'ibe = 我們倆在打你 \n
nolenufu'inagihe = 因為我們倆在刺你們 \n
nolifi'ibe = 你們倆在刺我們 \n
nofunagihe = 因為我在刺他 \n
nofine = 你在刺他 \n

nifila'ibe = 你們倆將刺我 \n
nonahatagihe = 因為你在打我 \n
lenahalube = 我將打你們 \n
nahalanagihe = 因為你們將打我 \n
lahala'ibe = 你們倆將打我們 \n
nofutagihe = 因為我們在刺他 \n
lenifilu'ibe = 我們倆將刺你們 \n
noho'inagihe = 因為我們倆在打他",
"note": "△！貝納貝納語屬於跨新幾內亞語系. 在巴布亞新幾內亞, 大約四萬五千人使用該語 \n 言. ——戴誼凡 (伊萬・德爾然斯基)",
"question":[
{
"question": "翻譯成漢語:",
"options": ["nonifibe", "halu'ibe", "lifilatagihe", "nokufune", "nolahanagihe"],
"answer": ["你們在刺我", "我們倆將打他", "因為你將刺我們", "我們在刺你", "因為你們在打我們"]
},
{
"question": "翻譯成貝納貝納語:",
"options": ["你們倆在打他", "我們將刺你", "因為我們在打你們", "因為你們將刺他"],
"answer": ["noha'ibe", "kifilune", "nolenohotagihe", "filanagihe"]

}
]
}
}

# 4 實驗

## 4.1 評估指標

Kishore 等人（2002）提出 BLEU 值作為評估機器翻譯品質的自動指標，並指出 BLEU 的優勢在於，它通過在整個測試語料庫上平均化單個句子的判斷誤差（而非試圖精確預測每個句子的主觀人工評判），從而與人類評判保持高度相關性。本研究構建的評測數據集構建的題目均為翻譯題，所以採用的評估指標主要為 BLEU 值。BLEU（Bilingual Evaluation Understudy）值是一種用於評估機器翻譯品質的自動化指標，通過比較機器翻譯結果與人工參考翻譯的相似度來衡量其準確性。

計算方法如下：

n-gram 匹配：電腦器翻譯與參考翻譯在 1-gram、2-gram、3-gram 和 4-gram 上的匹配程度。

精度計算：統計機器翻譯中 n-gram 與參考翻譯匹配的比例。

brevity penalty（簡短懲罰）：如果機器翻譯比參考翻譯短，則施加懲罰，避免過短翻譯得分過高。

綜合得分：將各 n-gram 的精度加權平均，再乘以簡短懲罰，得到最終的 BLEU 值。

計算公式如下：

$$\mathrm{BLEU} = \mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

## 4.2 評估對象

在本研究中，我們聚焦於對 GPT-4o(即 ChatGPT)、deepseek、qwen 以及 glm 翻極低資源語言翻譯推理性能的深度評估。這四個大模型具備強大的通用語言理解和生成能力，且支持多種語言，儘管低資源語言的訓練數據有限，但其強大的遷移學習能力使其在低資源語言任務中仍具潛力。這四個模型有各自的特點和優勢，能夠全面覆蓋從通用到專用、從大規模到輕量化的不同需求。通過對比它們在低資源語言翻譯中的表現，可以更好地理解不同模型在該任務中的潛力與局限，為未來優化提供依據。

有鑒於此，我們認為 GPT-4o、deepseek、qwen 以及 glm 在該評測任務上的表現具有很高的參考價值。對 GPT-4o、deepseek、qwen 以及 glm 四個大型語言模型進行評測，可以反映當前大型語言模型在極低資源語言翻譯推理任務上的處理能力水準。綜上所述，在本項研究中，本研究選取 GPT-4o、deepseek、qwen 以及 glm 四個大模型作為評測對象，通過系統的對比分析，旨在為大型語言模型的發展方向提供寶貴的建設性參考意見。

## 4.3 情境學習

情境學習（In-Context Learning）（Brown et al., 2020）是一種基於大規模預訓練語言模型的任務適應策略。該方法是通過在輸入序列中嵌入相關的上下文提示資訊，使模型能夠在無需顯式參數更新的情況下，依據任務描述或少量輸入-輸出示例完成特定任務。與傳統的微調方法不同，情境學習不依賴於反向傳播或參數優化，而是通過設計有效的提示詞（prompt）來引導模型理解並執行任務。這一特性使其特別適用於大規模預訓練模型，因為這類模型通常難以直接微調，尤其是在標注數據稀缺的場景下。情境學習的核心優勢在於，它能夠在不依賴大量標注數據的前提下，通過調整提示詞的表達方式，使模型快速適應新任務，從而顯著降低了對計算資源和數據標注的依賴。這種能力不僅提升了模型的泛化性能，也為低資源場景下的任務適應提供了可行的解決方案。

本研究重點評估了模型在零樣本學習（Zero-Shot Learning, 0-shot）和少樣本學習（Few-Shot Learning, 3-shots）兩種模式下的性能表現。零樣本學習是指模型在沒有任何任務相關示例的情況下，僅依靠任務描述或指令來執行任務。這種設置能夠有效檢驗模型在面對全新任務時的泛化能力及其對未知場景的適應性。相比之下，3-shots 則為模型提供了少量任務示例，旨在通過有限的樣本資訊提升模型的表現。實驗結果表明，經過優化後，模型在兩種學習模式下的回應品質和任務完成度均得到了顯著提升，具體數據詳見第 5 部分的實驗結果分析。這一發現不僅驗證了模型改進的有效性，也為後續研究提供了重要的參考依據。

## 4.4 測試流程

該測評流程首先通過初始化配置，設置 API 密鑰和語言映射，並定義數據集路徑。接著，腳本按語言對數據集檔進行分組，確保每種語言的數據能夠被獨立處理。隨後，讀取 JSON 檔並調用 3-shots 情境學習模型進行翻譯評測，利用少量示例進行推理，生成翻譯結果。評測完成後，腳本將結果保存為 JSON 檔，並記錄已處理的檔，避免重複操作。最後，啟動測評流程，依次處理所有數據集檔，確保每個檔都能被高效評測。通過這一自動化流程，實現了對多語言數據集的翻譯評測，並支持 3-shots 情境學習，能夠高效生成評測結果並保存。

具體步驟見下：

（1）初始化與配置

API 配置：在 ModelApiEvaluator 類中，設置 API 密鑰（API_KEY）和語言映射選項（mapping_option），用於調用翻譯模型的 API。

檔路徑配置：通過 CURRENT_FILE_PATH 獲取當前腳本的路徑，並定義數據集的存儲路徑（data_set 檔夾）。

（2）檔分組

遍歷數據集檔夾：通過 group_files 方法遍曆 data_set 檔夾中的所有 JSON 檔，並根據父檔夾名稱對檔進行分組。

生成檔組：使用 get_group_files 方法將檔按語言分組，生成一個字典，鍵為語言名稱，值為對應語言的 JSON 檔列表。

（3）讀取與處理 JSON 檔

加載已處理檔記錄：通過 load_process 方法讀取 process.txt 檔，記錄已經處理過的檔，避免重複處理。

讀取 JSON 檔：使用 load_json_file 方法讀取每個 JSON 檔的內容，獲取數據集。

3-shots 情境學習：調用 model_three_shot 類的 evaluate_translation 方法，對數據集進行 3-shots 情境下的翻譯評測。

（4）保存結果

生成結果檔案名：將原始 JSON 檔案名替換為 3shot-results.json，作為結果檔的名稱。

保存結果檔：通過 save_results_to_json 方法將評測結果保存到指定的路徑中，路徑為 dataset_result/3shots/語言名稱/。

記錄已處理檔：使用 save_process 方法將已處理的檔路徑記錄到 process.txt 中，確保後續不會重複處理。

（5）運行測評

啟動測評流程：在 run 方法中，依次調用 group_files、get_group_files 和 read_json_files 方法，完成整個測評流程。

（6）執行腳本

主程序入口：在 if __name__ == "__main__":中，實例化 ModelApiEvaluator 類並調用 run 方法，啟動測評流程。

# 5 結果與分析

經過測試，得到以下結果。表 1 以可視化形式呈現了模 chatGPT-4o，deepseek，智譜清言和通義千問四個語言大模型在面對不同語言時，在兩種不同的情境學習條件(0S 和 3S)下的表現。

| 模型 | 語言 | 情境學習 | BLEU |
|---|---|---|---|
| chatgpt | 中文 | 0-shot | 2.53 |
| | | 3-shots | 2.53 |
| | 英語 | 0-shot | 6.51 |
| | | 3-shots | 6.10 |
| | 俄語 | 0-shot | 3.92 |
| | | 3-shots | 3.95 |
| | 斯洛文尼亞語 | 0-shot | 4.20 |
| | | 3-shots | 3.66 |
| | 愛沙尼亞語 | 0-shot | 4.89 |
| | | 3-shots | 4.05 |
| | 羅馬尼亞語 | 0-shot | 4.35 |
| | | 3-shots | 4.28 |
| deepseek | 中文 | 0-shot | 0.62 |
| | | 3-shots | 0.72 |
| | 英語 | 0-shot | 1.05 |
| | | 3-shots | 1.15 |
| | 俄語 | 0-shot | 0.66 |
| | | 3-shots | 0.80 |
| | 斯洛文尼亞語 | 0-shot | 0.71 |
| | | 3-shots | 0.85 |
| | 愛沙尼亞語 | 0-shot | 0.90 |
| | | 3-shots | 0.99 |
| | 羅馬尼亞語 | 0-shot | 0.69 |
| | | 3-shots | 0.68 |
| 智譜清言 | 中文 | 0-shot | 0.68 |
| | | 3-shots | 0.81 |
| | 英語 | 0-shot | 1.11 |
| | | 3-shots | 1.24 |
| | 俄語 | 0-shot | 0.71 |
| | | 3-shots | 0.94 |
| | 斯洛文尼亞語 | 0-shot | 0.91 |
| | | 3-shots | 1.00 |
| | 愛沙尼亞語 | 0-shot | 0.67 |
| | | 3-shots | 0.85 |
| | 羅馬尼亞語 | 0-shot | 1.18 |
| | | 3-shots | 1.11 |
| 通義千問 | 中文 | 0-shot | 2.19 |

| | | |
|---|---|---|
| 英語 | 3-shots | 3.17 |
| | 0-shot | 1.92 |
| | 3-shots | 3.60 |
| 俄語 | 0-shot | 1.23 |
| | 3-shots | 2.06 |
| 斯洛文尼亞語 | 0-shot | 1.77 |
| | 3-shots | 2.78 |
| 愛沙尼亞語 | 0-shot | 1.51 |
| | 3-shots | 2.07 |
| 羅馬尼亞語 | 0-shot | 1.60 |
| | 3-shots | 2.55 |

表 1 四個語言大模型在語言學奧賽上的表現

表 2 為四個不同語言大模型的綜合表現，在測試的四個模型中，ChatGPT 整體表現最優，平均 BLEU 得分為 3.93，顯著高於其他模型。其在英語任務中的 0-shot 表現尤為突出，達到了 6.51 的 BLEU 值，但中文任務的表現相對較弱（0-shot 和 3-shots 均為 2.53 左右）。在英語(6.51)、愛沙尼亞語(4.89)和斯洛文尼亞語(4.20)的 0-shot 任務中表現優越。通義千問的平均 BLEU 為 2.22，排名第二，其特點是 3-shots 學習效果顯著，尤其在英語任務中實現了最大增益（+1.68）。智譜清言在羅馬尼亞語 0-shot(1.18)和斯洛文尼亞語 3-shots(1.00)等低資源語言上有相對優勢。智譜清言和 DeepSeek 的平均 BLEU 分別為 0.89 和 0.76，表現相對較弱，但智譜清言在英語 3-shots 任務中達到了 1.24 的峰值，而 DeepSeek 的穩定性較高（標準差 0.17）。DeepSeek 表現相對均衡，但整體水準較低，無明顯優勢領域。總體來看，ChatGPT 在多語言任務中表現全面領先，而通義千問在情境學習增益上更具優勢。

| 模型 | 平均 BLEU | 最高 BLEU(語言/情境) | 最低 BLEU(語言/情境) | 表現穩定性(標準差) |
|---|---|---|---|---|
| ChatGPT | 3.93 | 6.51(英語 0-shot) | 2.53(中文 0/3-shot) | 1.32 |
| 通義千問 | 2.22 | 3.60(英語 3-shots) | 1.23(俄語 0-shot) | 0.76 |
| 智譜清言 | 0.89 | 1.24(英語 3-shots) | 0.67(愛沙尼亞語 0-shot) | 0.18 |
| DeepSeek | 0.76 | 1.15(英語 3-shots) | 0.62(中文 0-shot) | 0.17 |

表 2 不同語言大模型綜合表現對比

表 3 為語言大模型在不同語言上的表現，從語言維度看，所有模型在英語上的表現都顯著優於其他語言，ChatGPT 的英語 0-shot(6.51)甚至高於其他模型最佳表現的 2-6 倍。英語任務的平均 BLEU 最高（3.14），其次是愛沙尼亞語（1.83）和斯洛文尼亞語（1.81）。ChatGPT 在英語、愛沙尼亞語和俄語的 0-shot 任務中均取得了最高分，展現了強大的跨語言能力。中文任務的表現相對分化，通義千問在 3-shots 條件下以 3.17 的 BLEU 領先，是唯一接近 ChatGPT 英語水準的非英語表現，而 DeepSeek 的 0-shot 得分最低（0.62）。羅馬尼亞語和俄語的任務中，模型間差異較大（標準差分別為 1.41 和 1.31）。中文 0-shot 表現普遍弱於 3-shots，可能中文任務需要更多上下文。在高資源語言中，可能因為其語言結構的複雜性，或者訓練數據品質/數量問題，或者評測指標對瀕危語言的適應性俄語表現較弱。總之，不同模型對不同語言的處理能力存在顯著差距。

| 語言 | 平均 BLEU | 最高 BLEU(模型/情境) | 最低 BLEU(模型/情境) | 模型間差異(標準差) |
|---|---|---|---|---|
| 英語 | 3.14 | 6.51(ChatGPT 0-shot) | 1.05(DeepSeek 0-shot) | 2.12 |
| 愛沙尼亞語 | 1.83 | 4.89(ChatGPT 0-shot) | 0.67(智譜清言 0-shot) | 1.58 |
| 斯洛文尼亞語 | 1.81 | 4.20(ChatGPT 0-shot) | 0.71(DeepSeek 0-shot) | 1.39 |
| 羅馬尼亞語 | 1.76 | 4.35(ChatGPT 0-shot) | 0.68(DeepSeek 3-shots) | 1.41 |
| 俄語 | 1.66 | 3.92(ChatGPT 0-shot) | 0.66(DeepSeek 0-shot) | 1.31 |
| 中文 | 1.76 | 3.17(通義千問 3-shots) | 0.62(DeepSeek 0-shot) | 1.01 |

表 3 語言大模型針對不同語言的綜合表現

表 4 為不同語言大模型情境學習的增益，結果顯示情境學習的增益效果因語言大模型而異。通義千問學習能力最強，從 0-shot 到 3-shots 的平均增益最高（+0.75），尤其在英語任務中提升顯著（+1.68），且所有語言在通義千問的情境學習上都呈現正增長，說明其穩定的 few-shot 學習能力。智譜清言和 DeepSeek 的增益較小（分別為+0.15 和+0.07），且存在個別語言的負面效果（如羅馬尼亞語任務中，智譜清言的 3-shots 表現反而下降）。ChatGPT 整體呈現輕微負增益（-0.07），僅在俄語任務中有微弱提升（+0.03），而在愛沙尼亞語任務中 3-shots 表現大幅下降（-0.84）。經過情境學習，對於高資源語言，中文、英語和俄語均有正向增益顯著。而對於低資源語言，羅馬尼亞語、愛沙尼亞語和斯洛文

尼亞語增益不顯著或呈負面。這表明情境學習的效果高度依賴模型和語言組合，並非所有場景下都能帶來改進。

| 模型 | 平均增益(3-shots vs 0-shot) | 最大增益(語言) | 負面效果(語言) |
|---|---|---|---|
| 通義千問 | 0.75 | +1.68(英語) | 無 |
| 智譜清言 | 0.15 | +0.23(俄語) | -0.08(羅馬尼亞語) |
| DeepSeek | 0.07 | +0.15(愛沙尼亞語) | -0.01(羅馬尼亞語) |
| ChatGPT | -0.07 | +0.03(俄語) | -0.84(愛沙尼亞語) |

表 4 不同語言大模型情境學習增益

綜合來看，ChatGPT 在多語言任務中表現最優，但情境學習對其幫助有限；通義千問雖整體得分次之，卻能通過 3-shot 學習顯著提升性能；智譜清言和 DeepSeek 表現較弱，但穩定性較高。英語任務整體表現最佳，而模型對低資源語言（如愛沙尼亞語、斯洛文尼亞語）的處理能力差異較大。情境學習的有效性因模型和語言而異。

## 6 討論

基於 § 5 的分析，我們得出了一系列重要結論:

1. ChatGPT 整體表現最優，平均 BLEU 得分顯著高於其他模型，其次是通義千問，其 3-shot 學習效果顯著，智譜清言在低資源語言上有相對優勢，而 DeepSeek 表現相對均衡，但整體水準較低，無明顯優勢領域。

2. 情境學習顯著提升了模型在大部分高資源語言翻譯推理任務中的性能，在英語和俄語任務中的表現有力證實了少樣本學習對增強大模型在面對極低資源語言翻譯推理任務中的關鍵價值。但情境學習在大多數低資源語言（如愛沙尼亞語、斯洛文尼亞語）的翻譯推理任務上，語言大模型仍面臨著嚴峻考驗。

3.經過少樣本學習，部分語言的 bleu 值降低，我們推測這可能是由於 3-shots 提供的樣本數量非常有限，可能無法充分捕捉目標語言的複雜性和多樣性，模型可能無法從中學習到有效的模式，導致性能下降。並且某些語言（如中文）具有複雜的語法結構、豐富的辭彙和多義性，3-shots 樣本可能無法覆蓋這些複雜性，導致模型在少樣本學習後表現不佳。

4.模型在極低資源語言的推理上有一定的局限性，凸顯了其在應對陌生的瀕危語言推理時面臨的挑戰。

# 7 總結與展望

本文旨在填補對大型語言模型對極低資源語言邏輯推理能力評測方面的研究空白，探究大型語言模型在極低資源語言上的表現，我們依據語言學奧賽題目，構建了一個專門針對極低資源語言的測評集來系統地評估大模型對極低資源語言邏輯推理性能。

通過對評測數據細緻且深度的解析，我們詳細地展示了當前大模型在處理極低資源語言時的性能表現，揭示了其在面對極低資源的陌生語言時所面臨的的挑戰；同時也驗證了少樣本學習對大型語言模型的在翻譯推理任務上的表現有顯著提升；除此之外，我們還發現相對於高資源語言，在低資源語言的推理方面，少樣本輸入反而導致模型得分下降，並針對此現象可能的原因作出瞭解釋。

針對以上問題，後續工作中可以從通過增加樣本數量、優化樣本品質以及結合多任務學習，提升模型在少樣本情境下的泛化能力。通過引入更多的低資源語言數據或跨語言遷移學習，提升模型在這些語言上的表現，增強模型對低資源語言的適應性。並且可以結合多種評估指標（如 ROUGE、METEOR 或人工評估），更全面地衡量模型生成文本的品質。同時也要研究更適合少樣本學習和多語言任務的模型架構，進一步提升模型的性能和效率。

同時，由於語言學奧賽試題難度極大，導致本研究的調查問卷完成度和完成品質較差，導致本研究缺乏人類基線，後續可以設計更完善的問卷形式，構建人類基線。

通過持續優化和改進，我們有望在未來的多語言任務中實現更高的準確性和泛化能力，為自然語言處理領域的發展提供有力支持。

## 參考文獻


Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., et al. 2020. *Language Models are Few-Shot Learners. ArXiv, abs/2005.14165*.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z.L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., & Sun, M. (2024). *OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems.* Annual Meeting of the Association for Computational Linguistics.

Baktash, J.A., & Dawodi, M. (2023). *Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing.* ArXiv, abs/2305.03195.

Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. *A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models.* In Findings of the Association for Computational Linguistics: EACL 2024, pages 323–339, St. Julian's, Malta. Association for Computational Linguistics.

Trinh, T.H., Wu, Y., Le, Q.V. et al. *Solving olympiad geometry without human demonstrations.* Nature 625, 476–482 (2024).

Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. *A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Language Models.* In Findings of the Association for Computational Linguistics: EACL 2024, pages 323–339, St. Julian's, Malta. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. *Scieval: A multi-level large language model evaluation benchmark for scientific research.*

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *In Proceedings of the 59th Annual Meeting of the Association for Computational*

*Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774– 6786, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. *Measuring mathematical problem solving with the me math dataset.* arXiv preprint arXiv:2103.03874.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

**附錄**

附錄 1 六種語言原始試題形式及結構化後數據形式（以 2014 年語言學奧賽題第一題為例）

中文試題

**zh**

# 第十二届国际语言学奥林匹克竞赛

## 中国　北京　2014年7月21日 – 25日

### 个人赛题目

毋需抄题. 将不同问题的解答分述于不同的答题纸上. 每张纸上注明题号, 座位号和姓名. 否则答题纸可能被误放或张冠李戴.

解答需详细论证. 无解释之答案, 即便完全正确, 也会被处以低分.

---

**第一题 (20分).** 以下为贝纳贝纳语的一些动词形式及其汉语翻译:

| | |
|---|---|
| *nohobe* | 我在打他 |
| *kahalune* | 我们将打你 |
| *nokoho'ibe* | 我们俩在打你 |
| *nolenufu'inagihe* | 因为我们俩在刺你们 |
| *nolifi'ibe* | 你们俩在刺我们 |
| *nofunagihe* | 因为我在刺他 |
| *nofine* | 你在刺他 |
| *nifila'ibe* | 你们俩将刺我 |
| *nonahatagihe* | 因为你在打我 |
| *lenahalube* | 我将打你们 |
| *nahalanagihe* | 因为你们将打我 |
| *lahala'ibe* | 你们俩将打我们 |
| *nofutagihe* | 因为我们在刺他 |
| *lenifilu'ibe* | 我们俩将刺你们 |
| *noho'inagihe* | 因为我们俩在打他 |

(a) 翻译成汉语:

*nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.*

(b) 翻译成贝纳贝纳语:

- 你们俩在打他;
- 我们将刺你;
- 因为我们在打你们;
- 因为你们将刺他.

⚠ 贝纳贝纳语属于跨新几内亚语系. 在巴布亚新几内亚, 大约四万五千人使用该语言.

——戴谊凡 *(伊万·德尔然斯基)*

英文試題

# en(B)

## Twelfth International Olympiad in Linguistics

Beijing (China), 21–25 July 2014

Individual Contest Problems

Do not copy the statements of the problems. Write down your solution to each problem on a separate sheet or sheets. On each sheet indicate the number of the problem, the number of your seat and your surname. Otherwise your work may be mislaid or misattributed.

Your answers must be well-supported by argument. Even a perfectly correct answer will be given a low score unless accompanied by an explanation.

---

**Problem #1 (20 points).** Here are some verb forms of Benabena and their English translations:

| | |
|---|---|
| ***nohobe*** | I am striking him |
| ***kahalune*** | we will strike $you_{sg}$ |
| ***nokoho'ibe*** | we both are striking $you_{sg}$ |
| ***nolenufu'inagihe*** | because we both are piercing $you_{pl}$ |
| ***nolifi'ibe*** | you both are piercing us |
| ***nofunagihe*** | because I am piercing him |
| ***nofine*** | $you_{sg}$ are piercing him |
| ***nifila'ibe*** | you both will pierce me |
| ***nonahatagihe*** | because $you_{sg}$ are striking me |
| ***lenahalube*** | I will strike $you_{pl}$ |
| ***nahalanagihe*** | because $you_{pl}$ will strike me |
| ***lahala'ibe*** | you both will strike us |
| ***nofutagihe*** | because we are piercing him |
| ***lenifilu'ibe*** | we both will pierce $you_{pl}$ |
| ***noho'inagihe*** | because we both are striking him |

**(a)** Translate into English:

***nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.***

**(b)** Translate into Benabena:

- you both are striking him;
- we will pierce $you_{sg}$;
- because we are striking $you_{pl}$;
- because $you_{pl}$ will pierce him.

⚠ The Benabena language belongs to the Trans-New Guinea family. It is spoken by approx. 45,000 people in Papua New Guinea. —*Ivan Derzhanski*

俄語試題

**ru**

# Двенадцатая Международная олимпиада по лингвистике

## Пекин (Китай), 21–25 июля 2014 г.

### Задачи индивидуального соревнования

Не переписывайте условий. Решайте каждую задачу на отдельном листе (или листах). На каждом листе проставьте номер решаемой задачи, номер Вашего места и Вашу фамилию. Только в этом случае гарантируется точная оценка Вашей работы.

Полученные Вами ответы нужно обосновывать. Даже абсолютно верный ответ оценивается низко, если он приведён безо всякого обоснования.

---

**Задача №1 (20 баллов).** Даны формы глаголов на языке бенабена и их русские переводы:

| | |
|---|---|
| ***nohobe*** | я бью его |
| ***kahalune*** | мы побьём тебя |
| ***nokoho'ibe*** | мы двое бьём тебя |
| ***nolenufu'inagihe*** | потому что мы двое колем вас |
| ***nolifi'ibe*** | вы двое колете нас |
| ***nofunagihe*** | потому что я колю его |
| ***nofine*** | ты колешь его |
| ***nifila'ibe*** | вы двое уколете меня |
| ***nonahatagihe*** | потому что ты бьёшь меня |
| ***lenahalube*** | я побью вас |
| ***nahalanagihe*** | потому что вы побьёте меня |
| ***lahala'ibe*** | вы двое побьёте нас |
| ***nofutagihe*** | потому что мы колем его |
| ***lenifilu'ibe*** | мы двое уколем вас |
| ***noho'inagihe*** | потому что мы двое бьём его |

**(a)** Переведите на русский язык:

***nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.***

**(b)** Переведите на язык бенабена:

- вы двое бьёте его;
- мы уколем тебя;
- потому что мы бьём вас;
- потому что вы уколете его.

⚠ Язык бенабена относится к трансновогвинейской семье. На нём говорят около 45 000 человек в Папуа — Новой Гвинее. *—Иван Держанский*

羅馬尼亞語試題

**RO**

# A douăsprezecea Olimpiadă internaţională de lingvistică

## Beijing (China), 21–25 iulie 2014

### Probleme pentru competiţia individuală

Nu copiaţi datele oferite de problemă. Rezolvaţi fiecare problemă pe o foaie (sau foi) separată (separate). Scrieţi pe fiecare foaie numărul problemei, numărul locului dvs. şi numele dvs. de familie. Altfel, rezultatele dvs. ar putea fi atribuite în mod greşit unui alt participant.

Răspunsurile trebuie să fie bine argumentate. Un răspuns dat fără explicaţie, chiar dacă este absolut corect, va primi un punctaj scăzut.

---

**Problema nr. 1 (20 de puncte)** Sunt date unele forme ale verbelor în limba benabena şi traducerile lor în limba română:

| | |
|---|---|
| ***nohobe*** | eu îl lovesc |
| ***kahalune*** | noi te vom lovi |
| ***nokoho'ibe*** | noi ambii te lovim |
| ***nolenufu'inagihe*** | pentru că noi ambii vă străpungem |
| ***nolifi'ibe*** | voi ambii ne străpungeţi |
| ***nofunagihe*** | pentru că eu îl străpung |
| ***nofine*** | tu îl străpungi |
| ***nifila'ibe*** | voi ambii mă veţi străpunge |
| ***nonahatagihe*** | pentru că tu mă loveşti |
| ***lenahalube*** | eu vă voi lovi |
| ***nahalanagihe*** | pentru că voi mă veţi lovi |
| ***lahala'ibe*** | voi ambii ne veţi lovi |
| ***nofutagihe*** | pentru că noi îl străpungem |
| ***lenifilu'ibe*** | noi ambii vă vom străpunge |
| ***noho'inagihe*** | pentru că noi ambii îl lovim |

**(a)** Traduceţi în limba română:

***nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.***

**(b)** Traduceţi în limba benabena:

- voi ambii îl loviţi;
- noi te vom străpunge;
- pentru că noi vă lovim;
- pentru că voi îl veţi străpunge.

⚠ Limba benabena face parte din familia trans-neoguineeană. Este vorbită de aproximativ 45 000 de persoane în Papua Noua Guinee.

*—Ivan Derjanski*

愛沙尼亞語

# et

## Kaheteistkümnes rahvusvaheline lingvistikaolümpiaad

Peking (Hiina), 21.–25. juuli 2014

Individuaalvõistluse ülesanded

Ärge kirjutage ülesandeid ümber. Lahendage iga ülesanne eraldi lehel (lehtedel). Kirjutage lahendatava ülesande number, oma koha number ja nimi igale ülesande lahenduse lehele eraldi. Ainult sel juhul on Teie tulemuste täpne arvestus garanteeritud.

Põhjendage iga vastust. Täiesti õigeid, kuid põhjenduseta vastuseid hinnatakse madalalt.

---

**Ülesanne nr 1 (20 punkti).** On antud mõned verbivormid benabena keeles ning nende eestikeelsed tõlked:

| | |
|---|---|
| ***nohobe*** | ma löön teda |
| ***kahalune*** | me hakkame lööma sind |
| ***nokoho'ibe*** | me mõlemad lööme sind |
| ***nolenufu'inagihe*** | sest me mõlemad torkame teid |
| ***nolifi'ibe*** | te mõlemad torkate meid |
| ***nofunagihe*** | sest ma torkan teda |
| ***nofine*** | sa torkad teda |
| ***nifila'ibe*** | te mõlemad hakkate torkama mind |
| ***nonahatagihe*** | sest sa lööd mind |
| ***lenahalube*** | ma hakkan lööma teid |
| ***nahalanagihe*** | sest te hakkate lööma mind |
| ***lahala'ibe*** | te mõlemad hakkate lööma meid |
| ***nofutagihe*** | sest me torkame teda |
| ***lenifilu'ibe*** | me mõlemad hakkame torkama teid |
| ***noho'inagihe*** | sest me mõlemad lööme teda |

**(a)** Tõlkige eesti keelde:

***nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.***

**(b)** Tõlkige benabena keelde:

- te mõlemad lööte teda;
- me hakkame torkama sind;
- sest me lööme teid;
- sest te hakkate torkama teda.

⚠ Benabena keel kuulub trans-uus-guinea keelte hulka. Seda räägib umbes 45 000 inimest Paapua Uus-Guineas.

—*Ivan Deržanski*

斯洛文尼亞語

# sl

## Dvanajsta mednarodna olimpijada iz jezikoslovja

Peking (Kitajska), 21.–25. julij 2014

Naloge individualnega tekmovanja

Ne prepisuj opisov nalog. Rešitve posameznih nalog napiši vsako na svoj list papirja. Na vsakem listu jasno označi številko naloge, številko svojega sedeža in svoj priimek. Del tvojega dela bo sicer lahko izgubljen ali pripisan komu drugemu.

Odgovori morajo biti dobro utemeljeni. Tudi popolnoma pravilen odgovor bo dobil slabo oceno, če ob njem ne bo razlage.

---

**Naloga št. 1 (20 točk).** Podanih je nekaj glagolskih oblik v jeziku benabena ter njihovi prevodi v slovenščino:

| | |
|---|---|
| ***nohobe*** | jaz ga udarim |
| ***kahalune*** | mi te bomo udarili |
| ***nokoho'ibe*** | midva te udariva |
| ***nolenufu'inagihe*** | ker vas midva zabodeva |
| ***nolifi'ibe*** | vidva nas zabodeta |
| ***nofunagihe*** | ker ga jaz zabodem |
| ***nofine*** | ti ga zabodeš |
| ***nifila'ibe*** | vidva me bosta zabodla |
| ***nonahatagihe*** | ker me ti udariš |
| ***lenahalube*** | jaz vas bom udaril |
| ***nahalanagihe*** | ker me boste vi udarili |
| ***lahala'ibe*** | vidva nas bosta udarila |
| ***nofutagihe*** | ker ga mi zabodemo |
| ***lenifilu'ibe*** | midva vas bova zabodla |
| ***noho'inagihe*** | ker ga midva udariva |

**(a)** Prevedi v slovenščino:

***nonifibe, halu'ibe, lifilatagihe, nokufune, nolahanagihe.***

**(b)** Prevedi v benabena:

- vidva ga udarita;
- mi te bomo zabodli;
- ker vas mi udarimo;
- ker ga boste vi zabodli.

⚠ Jezik benabena spada v trans-novogvinejsko družino. Govori ga približno 45.000 ljudi v Papui Novi Gvineji. —*Ivan Deržanski*

結構化後數據形式


```
{
"source": ["第十二届国际语言学奥林匹克竞赛"],
    "location": "中国北京",
    "date": "2014年7月21-25日",
    "name": "个人赛题目",
 "problem": {
        "problem_1": {
            "title": "第一题",
            "score": 20,
            "instructions": " 以下为贝纳贝纳语的一些动词形式及其汉语翻译:\n
nohobe = 我在打他 \n
kahalune = 我们将打你 \n
nokoho'ibe = 我们俩在打你 \n
nolenufu'inagihe = 因为我们俩在刺你们 \n
nolifi'ibe = 你们俩在刺我们 \n
nofunagihe = 因为我在刺他 \n
nofine = 你在刺他 \n
nifila'ibe = 你们俩将刺我 \n
nonahatagihe = 因为你在打我 \n
lenahalube = 我将打你们 \n
nahalanagihe = 因为你们将打我 \n
lahala'ibe = 你们俩将打我们 \n
nofutagihe = 因为我们在刺他 \n
lenifilu'ibe = 我们俩将刺你们 \n
noho'inagihe = 因为我们俩在打他",
            "note": "ㅿ! 贝纳贝纳语属于跨新几内亚语系. 在巴布亚新几内亚, 大约四万五千人使用该语 \n言. ——戴谊凡 (伊万·德尔然斯基)",
            "question":[
                {
                    "question": "翻译成汉语:",
                    "options": ["nonifibe", "halu'ibe", "lifilatagihe", "nokufune", "nolahanagihe"],
                    "answer": ["你们在刺我", "我们俩将打他", "因为你将刺我们", "我们在刺你", "因为你们在打我们"]
                },
                {
                    "question": "翻译成贝纳贝纳语:",
                    "options": ["你们俩在打他", "我们将刺你", "因为我们在打你们", "因为你们将刺他"],
                    "answer": ["noha'ibe", "kifilune", "nolenohotagihe", "filanagihe"]

                }
            ]
```

附錄二 不同模型在不同語言、情境學習上的原始得分數據表（以 chatgpt 為例）

| 模型名稱 | 數據集類型 | folder_name | file_name | problem | title | total_bleu | overall_bleu |
|---|---|---|---|---|---|---|---|
| chatgpt | 3shots | 中文 | iol-2014-indiv-prob.zh3shot-results.json | problem_1 | 第一題 | 0.95726 5663 | 4.37959 8312 |
| chatgpt | 3shots | 中文 | iol-2014-indiv-prob.zh3shot-results.json | problem_2 | 第四題 | 3.42233 2649 | 4.37959 8312 |
| chatgpt | 3shots | 中文 | iol-2015-indiv-prob.zh3shot-results.json | problem_1 | 第二題 | 1.84371 7676 | 2.63207 0909 |
| chatgpt | 3shots | 中文 | iol-2015-indiv-prob.zh3shot-results.json | problem_2 | 第四題 | 0.78835 3233 | 2.63207 0909 |
| chatgpt | 3shots | 中文 | iol-2016-indiv-prob.zh3shot-results.json | problem_1 | 第三題 | 0.78792 5611 | 1.69700 931 |
| chatgpt | 3shots | 中文 | iol-2016-indiv-prob.zh3shot-results.json | problem_2 | 第五題 | 0.90908 3699 | 1.69700 931 |
| chatgpt | 3shots | 中文 | iol-2017-indiv-prob.zh3shot-results.json | problem_1 | 第三題 | 0.77445 556 | 1.81883 2883 |
| chatgpt | 3shots | 中文 | iol-2017-indiv-prob.zh3shot-results.json | problem_2 | 第五題 | 1.04437 7323 | 1.81883 2883 |
| chatgpt | 3shots | 中文 | iol-2018-indiv-prob.zh3shot-results.json | problem_1 | 第二題 | 1.90778 8821 | 3.86886 5462 |
| chatgpt | 3shots | 中文 | iol-2018-indiv-prob.zh3shot-results.json | problem_2 | 第四題 | 1.96107 6641 | 3.86886 5462 |
| chatgpt | 3shots | 中文 | iol-2019-indiv-prob.zh3shot-results.json | problem_1 | 第一題 | 0.80456 3807 | 1.93453 3379 |
| chatgpt | 3shots | 中文 | iol-2019-indiv-prob.zh3shot-results.json | problem_2 | 第五題 | 1.12996 9572 | 1.93453 3379 |
| chatgpt | 3shots | 中文 | iol-2021-indiv-prob.zh3shot-results.json | problem_1 | 第三題 | 2.85062 6017 | 3.36572 9681 |
| chatgpt | 3shots | 中文 | iol-2021-indiv-prob.zh3shot-results.json | problem_2 | 第五題 | 0.51510 3663 | 3.36572 9681 |
| chatgpt | 3shots | 中文 | iol-2022-indiv-prob.zh3shot-results.json | problem_1 | 第一題 | 0.42314 1046 | 1.37751 3341 |
| chatgpt | 3shots | 中文 | iol-2022-indiv-prob.zh3shot-results.json | problem_2 | 第三題 | 0.95437 2295 | 1.37751 3341 |
| chatgpt | 3shot | 中文 | iol-2023-indiv-prob.zh3sh | problem_1 | 第二題 | 0.61270 | 1.96988 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pt | s | | ot-results.json | em_1 | 题 | 7578 | 0478 |
| chatg | 3shot | 中文 | iol-2023-indiv-prob.zh3sh | probl | 第三 | 0.78523 | 1.96988 |
| pt | s | | ot-results.json | em_2 | 题 | 5882 | 0478 |
| chatg | 3shot | 中文 | iol-2023-indiv-prob.zh3sh | probl | 第四 | 0.57193 | 1.96988 |
| pt | s | | ot-results.json | em_3 | 题 | 7017 | 0478 |
| chatg | 0shot | 中文 | iol-2014-indiv-prob.zh-re | probl | 第一 | 1.02858 | 2.66854 |
| pt | | | sults.json | em_1 | 题 | 8744 | 305 |
| chatg | 0shot | 中文 | iol-2014-indiv-prob.zh-re | probl | 第四 | 1.63995 | 2.66854 |
| pt | | | sults.json | em_2 | 题 | 4306 | 305 |
| chatg | 0shot | 中文 | iol-2015-indiv-prob.zh-re | probl | 第二 | 1.36929 | 2.80101 |
| pt | | | sults.json | em_1 | 题 | 8922 | 8998 |
| chatg | 0shot | 中文 | iol-2015-indiv-prob.zh-re | probl | 第四 | 1.43172 | 2.80101 |
| pt | | | sults.json | em_2 | 题 | 0076 | 8998 |
| chatg | 0shot | 中文 | iol-2016-indiv-prob.zh-re | probl | 第三 | 0.44611 | 1.30170 |
| pt | | | sults.json | em_1 | 题 | 3477 | 2286 |
| chatg | 0shot | 中文 | iol-2016-indiv-prob.zh-re | probl | 第五 | 0.85558 | 1.30170 |
| pt | | | sults.json | em_2 | 题 | 8809 | 2286 |
| chatg | 0shot | 中文 | iol-2017-indiv-prob.zh-re | probl | 第三 | 1.36654 | 2.00455 |
| pt | | | sults.json | em_1 | 题 | 7585 | 3898 |
| chatg | 0shot | 中文 | iol-2017-indiv-prob.zh-re | probl | 第五 | 0.63800 | 2.00455 |
| pt | | | sults.json | em_2 | 题 | 6313 | 3898 |
| chatg | 0shot | 中文 | iol-2018-indiv-prob.zh-re | probl | 第二 | 2.03898 | 4.01609 |
| pt | | | sults.json | em_1 | 题 | 6487 | 5526 |
| chatg | 0shot | 中文 | iol-2018-indiv-prob.zh-re | probl | 第四 | 1.97710 | 4.01609 |
| pt | | | sults.json | em_2 | 题 | 9039 | 5526 |
| chatg | 0shot | 中文 | iol-2019-indiv-prob.zh-re | probl | 第一 | 0.92900 | 2.81723 |
| pt | | | sults.json | em_1 | 题 | 611 | 9426 |
| chatg | 0shot | 中文 | iol-2019-indiv-prob.zh-re | probl | 第五 | 1.88823 | 2.81723 |
| pt | | | sults.json | em_2 | 题 | 3316 | 9426 |
| chatg | 0shot | 中文 | iol-2021-indiv-prob.zh-re | probl | 第三 | 3.31783 | 3.93238 |
| pt | | | sults.json | em_1 | 题 | 8151 | 7676 |
| chatg | 0shot | 中文 | iol-2021-indiv-prob.zh-re | probl | 第五 | 0.61454 | 3.93238 |
| pt | | | sults.json | em_2 | 题 | 9525 | 7676 |
| chatg | 0shot | 中文 | iol-2022-indiv-prob.zh-re | probl | 第一 | 0.38081 | 1.66185 |
| pt | | | sults.json | em_1 | 题 | 7571 | 7181 |
| chatg | 0shot | 中文 | iol-2022-indiv-prob.zh-re | probl | 第三 | 1.28103 | 1.66185 |
| pt | | | sults.json | em_2 | 题 | 961 | 7181 |
| chatg | 0shot | 中文 | iol-2023-indiv-prob.zh-re | probl | 第二 | 0.37443 | 1.86186 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| pt | | | sults.json | em_1 | 題 | 4353 | 26 |
| chatgpt | 0shot | 中文 | iol-2023-indiv-prob.zh-results.json | problem_2 | 第三題 | 0.85441 | 1.86186 |
| | | | | | | 9582 | 26 |
| chatgpt | 0shot | 中文 | iol-2023-indiv-prob.zh-results.json | problem_3 | 第四題 | 0.63300 | 1.86186 |
| | | | | | | 8665 | 26 |

**致謝**

行文於此，落筆為終。食欲 2023 年秋，終於 2025 年盛夏。

兩年的碩士生涯即將結束，兩年的時光如白駒過隙，對我來說是青春序章裏炙熱的而意義非凡一章。回首碩士時光，百感交集。

一朝沐杏雨，一朝念師恩。感謝求學路上遇到的每一位老師，更要感謝的是我的導師袁毓林教授對我畢業論文的悉心指導，從選題到論文完稿，每一個環節都離不開導師字斟句酌的點撥指導，愚鈍有時，導師也不曾責備。師恩難忘，銘記於心，希望每一位老師在未來的工作生活中學術長青，工作順利，幸福安康。

再者感謝我的父母，予我生命，教我道理，助我成長，二十餘年的養育之恩不是一朝一夕，讓我走的路格外輕鬆坦蕩。樹高千尺不忘根深沃土，養育之恩，無以為報，只願我的家人身體安康，歲歲歡愉，年年順意。

一路走來，幸得朋友、室友和同學們的陪伴。熱烈的青春中遇到你們知足且幸福。祝你們前程似錦，未來一路繁花。

以夢為馬，不負韶華，最後感謝一下一直在路上的自己，求學二十餘載，縱然前路漫漫坎坷，祝自己保持初心，自洽而內求，清醒且堅定。

執筆至此，感慨萬千，道阻且長，關關難過關關過，願我們前路漫漫亦燦燦。