# 以 899 種南島語言數詞評估開放權重大型語言模型的構詞分析表現初探
# Evaluating Morphological Analysis Performance of Open-Weight LLMs Using Numerals from 899 Austronesian Languages: A Preliminary Study

**呂乃崴 Nai-Wei Lu**
Independent Researcher, Taiwan
weiweiaakid@gmail.com

## 摘要

本研究以 899 種南島語言的基礎數詞（1–10）為測試集，評估五款開放權重大型語言模型（8B–120B）在 zero-shot 與 one-shot 條件下推導數詞構詞公式的表現，並將輸出與人工標記比對。研究有四項發現。其一，8B 模型過半回應回抄輸入詞形（Echo），未能執行分析；27B 以上模型的 Echo 率近乎為零，zero-shot 整體準確率為 90.0–93.8%。然而，語料中 86.7% 的數詞為單純詞，整體準確率受多數類別影響，不能直接反映模型分析構詞複雜數詞的能力。其二，one-shot 範例使 8B 模型的準確率提升 42.1 個百分點，主要歸因於 Echo 減少；對較大型模型則未帶來一致增益。將 Echo 與解析失敗分開辨識，可區分任務執行、格式合規與分析正確三個階段。其三，模型對構詞複雜數詞的偵測召回率為 88–97%（啟發式基線為 57%），精確率則為 52–66%。模型常過度分析單純詞，並混淆加法與減法或誤判公式基底。其四，透過跨模型一致性與構詞透明度篩選，得到 29 個金標準未收錄的潛在構詞分析，顯示 LLM 可協助發現待查的構詞分析；然而，這些候選尚未經語言學專家驗證，因此無法判定其為模型發現的新分析，或是現有金標準的遺漏。整體而言，LLM 適合作為構詞分析的候選生成工具，但其輸出仍需專家驗證，不能直接視為可靠的最終標註。

## Abstract

This study evaluates five open-weight large language models (8B–120B) on numerals 1–10 from 899 Austronesian languages under zero-shot and one-shot conditions, comparing outputs with human annotations. Four findings emerge. First, the 8B model echoed over half of input forms, while models of 27B or more nearly eliminated Echo and achieved 90.0–93.8% overall accuracy. However, 86.7% of items were morphologically simple. Second, one-shot prompting raised 8B accuracy by 42.1 percentage points, mainly by reducing Echo, but yielded no consistent gains for larger models. Separating Echo and parse failures distinguishes task execution, format compliance, and analytical correctness. Third, complex-numeral detection recall was 88–97% (57% for a heuristic baseline), while precision was 52–66%. Models often overanalyzed simple forms and confused addition with subtraction or misidentified formula bases. Fourth, cross-model agreement and morphological transparency yielded 29 potential analyses absent from the gold standard, suggesting that LLMs can help identify candidate morphological analyses for further investigation. However, these candidates have not been linguistically validated, so it remains unclear whether they represent model-discovered analyses or omissions in the gold standard. Overall, LLMs are useful for generating candidates for morphological analysis, but their outputs require expert validation and should not be treated as reliable final annotations.

**關鍵字：**大型語言模型、數詞系統、南島語言、語言類型學、形態分析、模型評估

***Keywords:*** Large Language Models, Numeral Systems, Austronesian Languages, Linguistic Typology, Morphological Analysis, Model Evaluation

## 1 引言

數詞是人類語言中具跨語言可比性的語義範疇，幾乎所有語言都有表達基礎自然數的詞彙，但不同語言在構成這些詞彙時採用截然不同的策略：有的語言 1–10 多為不可分析的單純詞，有的語言以 5 為基底進行加法（如巴宰語 7 *xaseb-i-dusa*，包含 5 *xasep* 與 2 *dusa* 的成分，即 5+2；Li, 2006），有的採用減法（如雅浦語 9 *me-rep*，前綴 *me-* 加 1 *rep*，即 10−1），還有的採用乘法（如邵語 6 *ka-turu*，前綴 *ka-* 附加在詞根 *turu*（=3）上，即 2×3；Li, 2006）。



*Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026)*, pages 127–135
November 6-7, 2026 ©2026 Association for Computational Linguistics

單一語言可同時運用多種策略：表 1的恩德語即以加法構成 6–7、乘法構成 8、減法構成 9。

| # | 詞形 | 構詞分析 |
|---|---|---|
| 1 | *assa* | 單純詞 |
| 2 | *rua* | 單純詞 |
| 3 | *táru* | 單純詞 |
| 4 | *wutu* | 單純詞 |
| 5 | *rima* | 單純詞 |
| 6 | *rima ăssa* | 5+1：*rima* + *assa* |
| 7 | *rima rua* | 5+2：*rima* + *rua* |
| 8 | *rua mbutu* | 2×4：*rua* × *wutu* |
| 9 | *tarăssa* | 10−1：含 *assa* |
| 10 | *samburu* | 單純詞 |

表 1: 恩德語（Ende，弗洛勒斯島）的數詞系統。*wutu* → *mbutu* 為前鼻化音變。（語料來源：*The Numbers List* (Rosenfelder)；構詞分析為本文作者標註。）

南島語系（Austronesian）分布從臺灣延伸至太平洋諸島及馬達加斯加，涵蓋超過一千兩百種語言。其數詞系統呈現高度多樣性，除十進位制外，另有五進位與其他進位，以及與進位制無關的特殊表達策略 (Greenberg, 1978; Hammarström, 2010; Lu, 2026)。這種跨語言且形式透明的變異，使數詞成為測試模型從詞形對應推斷構詞模式的適合領域。

近年來，大型語言模型（Large Language Models, LLMs）已廣泛應用於翻譯、語法分析、語義推理等自然語言處理任務。然而，當 LLM 面對一組陌生語言的數詞時，它是否能正確從詞形的形態相似性推斷出構詞邏輯？

本研究旨在探討以下問題：

1. **任務表現**：不同規模的 LLM 從跨詞形對應推斷南島語言數詞構詞模式時，表現如何？

2. **樣本提示**：Zero-shot 與 one-shot 對模型表現有何影響？

3. **模式差異**：AI 對哪些構詞模式（加法、減法、乘法）的辨識表現較強或較弱？答題表現較傾向於過度分析或是過度簡化？

4. **輔助價值**：AI 能否為人工未標記的構詞模式提供候選？

為此，本研究以 899 種南島語言的基礎數詞（1–10）為測試集，選用五款參數量公開的開放權重模型（8B、27B、31B、70B、120B），操控 zero-shot 與 one-shot 兩種提示詞策略，以人工標記為金標準逐詞比對，量化 LLM 的數詞分析表現。

## 2 相關研究

數詞系統的跨語言比較是語言類型學的經典議題。Greenberg (1978) 奠定了加法、減法、乘法等數詞構詞模式的分類框架；Comrie (2005, 2013) 在 WALS 中整理了全球數詞基底類型及其分布；Hammarström (2010) 以大規模語料進一步調查數詞上限、混合基數與減法等特殊構詞類型的跨語言分布。針對南島語言，Blust (2009) 從比較語言學角度討論南島語數詞系統及其歷時發展。Lu (2026) 以 900 種南島語言數詞資料，系統標記加法、減法、乘法等構詞策略的分布，提出減法與乘法構詞的蘊含層級；本研究的金標準即源自該標記。

在數詞標注方法論方面，Rubehn et al. (2025) 提出標準化的數詞標注方案（CoSiNuS）。該研究運用 EDICTOR 工具進行手動標注與雙人同行審查機制（至少包含一名標注者與一名審查者），細緻切分形素邊界並對齊表層與底層形態結構，經跨語言資料格式（CLDF）驗證後建立金標準資料集。研究測試自動化形態切分模型，發現詞素變體（allomorphy）為切分失敗的主因。雖然該研究提供嚴謹的金標準標注協議，但其樣本僅涵蓋 25 種語言，完全未包含南島語系語言。兩者在研究策略上具高度互補性：Rubehn et al. 以少量語言進行深度標注（涵蓋形素邊界切分）；本研究則以大規模語言樣本（899 種）換取廣度，補足南島語系的缺口，並聚焦於基礎數詞 1–10 之構詞公式辨識，不處理具體的形素邊界切分。

在 LLM 方面，Mahowald et al. (2024) 系統性討論大型語言模型的語言能力與語言使用，指出模型能力並不等同於人類語言能力。數詞構詞分析是一項形態推理任務，要求模型僅根據有限詞形辨識跨詞彙的形態對應關係；而本研究所測試的南島語言多屬低資源語言，因此適合用來測試模型在缺乏語言特定監督下的構詞模式推斷能力。

## 3 研究方法

### 3.1 資料來源

本研究使用的數詞詞形取自 Rosenfelder 的線上數詞資料庫 *The Numbers List* (Rosenfelder)，其南島語系部分涵蓋 899 種語言的 1–10 基礎數詞（以國際音標或正字法轉寫）。

#### 3.1.1 金標準（Gold Standard）

構詞模式的金標準取自 Lu (2026) 的人工標記，扣除原始南島語後共 899 種語言：僅依據詞形間的形態對應關係，不參考語言背景知識或外部文獻，判定每個數詞是否可分析為其他數詞的組合，並記錄其構詞公式。

899 種語言中，583 種（65%）的 1–10 全為單純詞，316 種至少有一個構詞複雜數詞。表 2彙整各運算類型構詞複雜數詞的出現次數：五進位加法（$5+n$）最為普遍（214–236 種語言），其次為乘法（$2\times n$，共 82 次）與減法（$10-n$，共 77 次）；某些罕見模式（如 9=8+1: 2 種、8=7+1: 1 種）僅見於少數語言。

| 運算類型 | 子模式 | 格子次數 |
|---|---|---|
| 加法 | $5+n$（$n$=1∼4） | 899 |
| | 以 2 連加*（2+2+⋯） | 66 |
| | 其他（10=5+5 等） | 27 |
| 減法 | $10-n$（$n$=1∼4） | 77 |
| 乘法 | $2\times n$（$n$=3∼5） | 82 |

表 2: 各運算類型構詞複雜數詞的出現次數（僅列 ≥4 者）。同一語言若具有多個構詞複雜數詞（如 5+1 至 5+4），則分別計數。

*「以 2 連加」指以 2 為基底反覆相加的表達策略（如 3=2+1、4=2+2）、7=2+2+2+1。

表 3則列出各數詞的可能構詞公式。數詞 6–9 的公式最為多樣，反映加法、減法、乘法策略的運用可能性。

| 數詞 | 可能的構詞公式 |
|---|---|
| 3 | 2+1 |
| 4 | 2+2, 2×2 |
| 5 | 4+1 |
| 6 | 2×3, 4+2, 5+1, 10−4 |
| 7 | 2×3+1, 4+3, 5+2, 6+1, 8−1, 10−3 |
| 8 | 2×4, 5+3, 6+2, 7+1, 10−2 |
| 9 | 2×4+1, 5+2+2, 5+4, 8+1, 10−1 |
| 10 | 5+5, 2×5 |

表 3: 各數詞的可能構詞公式*。數詞 1 與 2 在本語料中皆為單純詞；受篇幅限制僅呈現主要模式。

* 以 2 連加的表達策略未另行列入；出現於 ≤2 種語言的罕見公式（如 7=2×4−1、10=8+2）亦省略。

針對語料中的一部分語言，*The Numbers List* 採取遞迴方式記錄（如語料記錄 7=[6][1]，而 6 本身又可分解為 2+2+2），本研究決定採取最展開的形式：操作元若本身為構詞複雜詞，則代入展開至基礎數詞（如語料記錄 4=[2][2] 且 9=[5][4] 則 9 的金標準則記為 5+2+2）。

29 個數詞（19 種語言）的詞形僅透明顯示一層分解（如西薩諾語 7 = *[6] [1]* 記為 2+2+2+1）；由於提示詞未示範遞迴展開，模型缺乏判斷依據，故此 29 格不納入評分。連同語料空缺格（201 格）及標註者存疑格（54 格），共 284 格不計分，實際計分格數為 8,706。判斷準則以子字串匹配與算術驗證為主，人工判斷空間有限；第 4.4節的跨模型篩選間接檢核標記品質。

### 3.2 AI 任務定義

任務要求 LLM 接收一組 10 個數詞（1–10 的詞形），輸出每個數詞的構詞公式。若數詞為不可分析的單純詞，輸出該數詞本身（如 “1”、“5”）；若數詞可分析為其他數詞的組合，輸出以算術運算表示的數詞構詞公式（如 “5+1”、“2×3”、“10−1”）。本文所稱「構詞公式」是以算術運算表示數詞形成模式，而非完整的形態素切分或語法分析。

**任務範例** 輸入（賽德克語 Seediq）：*1: kiŋal, 2: daha, 3: teru, 4: sepac, 5: rima, 6: mumuteru, 7: mupitu, 8: mumusepac, 9: muŋari, 10: maxal*。預期輸出：`1, 2, 3, 4, 5, 2×3, 7, 2×4, 9, 10`。此例中，6（mumuteru）包含 mu-mu-teru（teru = 3），推導為 2×3；8（mumusepac）包含 mu-mu-sepac（sepac = 4），推導為 2×4。

### 3.3 模型選擇

作為初步性研究（pilot study），本研究選用五款參數量公開的開放權重（open-weight）模型，五款模型涵蓋約一個數量級的規模跨度（8B → 27B → 31B → 70B → 120B）。其中 Llama 8B 與 70B 同屬 Meta Llama 系列，構成同系列的規模對照；Qwen 27B 與 Gemma 31B 為參數量相近但系列不同的模型，提供跨系列的參照；GPT-OSS 120B 則為稀疏架構（MoE）的代表。

| 模型 | 參數量 | 架構 |
|---|---|---|
| Llama 3.1 8B | 8B | Dense |
| Qwen 3.6 27B | 27B | Dense |
| Gemma 4 31B | 31B | Dense |
| Llama 3.3 70B | 70B | Dense |
| GPT-OSS 120B | 120B | Sparse (MoE) |

表 4: 核心實驗模型。推論平台：Groq（Llama 8B、Qwen 27B、Llama 70B）、Cerebras（Gemma 31B、GPT-OSS 120B）。

五款模型各在 zero-shot 與 one-shot 兩條件下對 899 種語言進行推論。

模型均透過各平台的 OpenAI 相容 API 呼叫，生成溫度設為 0。具備推理機制的模型將 `reasoning_effort` 設為最低，以維持跨模型的輸出可比性並控制 token 用量。

### 3.4 實驗條件

本研究操控兩個實驗條件：**Zero-shot** 僅提供數詞詞形，不附語言名稱或其他身分線索，測試模型從跨詞形對應推斷構詞模式的能力；**one-shot** 加入一個賽德克語範例，觀察範例對準確率的影響。兩條件均採用中性版提示詞

（見第 3.4.2節）。每次輸入僅包含單一語言的 1–10 數詞，因此不會受到資料表前後其他語言資料的上下文影響。

#### 3.4.1 提示詞設計

Zero-shot 條件使用英文提示詞模板：提供 10 個數詞詞形，要求模型以逗號分隔輸出每個數詞的構詞公式（單純詞輸出數字本身，複合詞輸出算術式如 5+1、2×3）。One-shot 條件加入一個賽德克語的完整範例（輸入詞形＋正確公式＋形態分析說明）。完整提示詞見附錄 A。

#### 3.4.2 提示詞設計注意事項

先導測試初步觀察到，提示詞中的格式範例對模型輸出有啟動效應（priming effect）：初版包含 “5+1, 2×3, 10−1” 等落在 1–10 範圍內的格式示範，導致部分模型模仿範例模式而非獨立分析詞形。為此，本研究設計了中性版提示詞，將符號說明改用範圍外的數字（如 10+2、2×10、20−1）。2 個模型 × 20 種語言的先導對照顯示，中性版可降低已具備任務理解力之模型的偽陽性率；而 8B 模型因回抄率過高（第 4.2節），兩版提示詞均無法引出有效分析，故難以評估啟動效應。本文全部主實驗採用中性版。

### 3.5 評分標準

#### 3.5.1 逐詞比對

本研究借用統計學的真偽陰陽性框架：「陽性」指模型判定該數詞為合成詞，即由其他數詞組合而成；「陰性」指判定為不可分析的單純詞。每個數詞的 AI 輸出與金標準逐一比對，判定為表 5所列五類之一。

| 判定 | 金標準 | AI 輸出 | 說明 |
|---|---|---|---|
| TP（真陽性） | 5+2 | 5+2 | 正確辨識合成詞 |
| TN（真陰性） | 1 | 1 | 正確辨識單純詞 |
| FP（偽陽性） | 6 | 5+1 | 單純詞誤判為合成詞 |
| FN（偽陰性） | 5+4 | 9 | 合成詞漏判為單純詞 |
| Wrong（公式錯誤） | 5+4 | 10−1 | 合成公式辨識錯誤 |

表 5: 評分判定類別。

**判定示例** 表 6以北艾法特語（N Efate，Gemma 31B, zero-shot）為例，展示逐詞判定的其中四種結果。AI 正確辨識 7 *la-rua* 含 *rua*（= 2）（TP），但將單純詞 6 *latesa* 誤判為 5+1（FP），並將 9 *lo-βeti* 的 5+4 誤認為 10−1（Wrong），整體 8/10。

公式字串在比對前經正規化：統一乘號、移除空白、符號表達順序合併[1]。

[1] 5+1 與 1+5 視為等價，2×3 與 3×2、5+5 與 2×5 亦同，因其區別未在提示詞中明示。此情況在 zero-shot 五模型合計 47 筆。

| # | 詞形 | 金標準 | AI 輸出 | 判定 |
|---|---|---|---|---|
| 1 | *sikai* | 1 | 1 | TN |
| 2 | *ndua* | 2 | 2 | TN |
| 3 | *ndolu* | 3 | 3 | TN |
| 4 | *pati* | 4 | 4 | TN |
| 5 | *lima* | 5 | 5 | TN |
| 6 | *latesa* | 6 | 5+1 | FP |
| 7 | *la-rua* | 5+2 | 5+2 | TP |
| 8 | *la-tolu* | 5+3 | 5+3 | TP |
| 9 | *lo-βeti* | 5+4 | 10−1 | Wrong |
| 10 | *ⁿdualima* | 2×5 | 2×5 | TP |

表 6: 北艾法特語（N Efate）逐詞判定示例（Gemma 31B, zero-shot）。

除以上五類之外，本研究另定義兩個獨立類別：**Echo（回抄）**指模型直接複製輸入詞形，顯示未執行要求的分析任務；**解析失敗**指回應經三層寬容解析仍無法擷取公式列表，該語言全部數詞計錯。上述類別共同區分任務執行、格式合規、分析正確三個階段（見第 4.2節）；Echo 與解析失敗不計入 P/R/F1 分母。

#### 3.5.2 報告指標

報告指標包括：整體準確率 Acc、構詞複雜詞的精確率 P、召回率 R 與 F1、逐數詞準確率、錯誤類型分布、多數類基線與字串匹配啟發式基線，以及偽陽性中經過濾後符合候選篩選條件的比例。

整體準確率（Accuracy, Acc）衡量模型在所有數詞（含單純詞）上的整體表現；精確率（precision, P）衡量模型判定為合成詞的數詞中，有多少確實正確；召回率（recall, R）衡量金標準中的合成詞，有多少被模型找出。兩者的分母分別為模型判定的合成詞總數（TP + FP + Wrong）與金標準的合成詞總數（TP + FN + Wrong）：

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} + \mathrm{Wrong}} \tag{1}$$

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{Wrong}} \tag{2}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{Wrong}} \tag{3}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

Wrong 同時出現在 P 與 R 的分母，是因為它表示模型成功偵測到構詞複雜性，但未找出金標準公式。

寬鬆偵測召回率 $\mathrm{R}_{\mathrm{det}}$ 僅要求模型將該數詞判為合成詞，不要求公式正確；因此，Wrong

在偵測層級屬於正確偵測，但在精確公式評估中同時不構成正確預測。

$$\mathrm{R}_{\mathrm{det}} = \frac{\mathrm{TP} + \mathrm{Wrong}}{\mathrm{TP} + \mathrm{FN} + \mathrm{Wrong}} \tag{5}$$

### 3.6 自動化流程

實驗透過 Python 腳本自動化執行：逐筆組裝提示詞、呼叫 API，並以三層遞進策略解析回應：(1) 尋找含恰好 10 個逗號分隔項的單行（指令要求的格式）；(2) 若無符合者，匹配逐行編號格式（如 "1: 5+1"），收集 10 筆；(3) 仍失敗則將全文以逗號切分，檢查是否恰為 10 段。三層均無法擷取 10 個公式時判定為解析失敗。比對金標準後以 JSONL 格式儲存。若模型回應因長度上限截斷（`finish_reason` 為 `length`），亦視為解析失敗。

## 4 結果與討論

本節依序回答四個研究問題。

### 4.1 任務表現：不同規模的模型表現

27B 以上模型之整體準確率介於 90.0–93.8%；而 8B 模型因過半回應為 Echo（見第 4.2節）準確率僅 14.1%。

**基線比較** 表 7比較兩個非 AI 基線與核心模型。多數類基線將所有數詞判為單純詞，因 86.7% 的數詞確實為單純詞而直接達 86.7%。字串匹配啟發式則檢查詞形是否包含較小數詞的子字串（≥3 字元），達 89.9%。

| | Acc% | P% | R% | F1% | $\mathrm{R}_{\mathrm{det}}$% |
|---|---|---|---|---|---|
| 多數類基線 | 86.7 | – | 0 | – | 0 |
| 字串匹配啟發式 | 89.9 | – | 57 | – | 57 |
| Qwen 27B | 90.0 | 52 | 78 | 63 | 97 |
| Gemma 31B | 93.8 | 66 | 78 | 72 | 95 |
| Llama 70B | 90.5 | 52 | 54 | 53 | 88 |
| GPT-OSS 120B | 91.4 | 58 | 84 | 69 | 97 |

表 7: 基線與核心模型比較（zero-shot，Llama 8B 因過半 Echo 不列入）。P/R/F1 為構詞複雜詞的嚴格指標；$\mathrm{R}_{\mathrm{det}}$ 為寬鬆偵測召回率。

四款大模型的整體準確率介於 90–94%，但辨識模式截然不同：啟發式基線的偵測召回率 $\mathrm{R}_{\mathrm{det}}$ 僅 57%，大模型達 88–97%，幾乎不遺漏構詞複雜詞，精確公式的精確率 P 仍有限（嚴格標準下 52–66%）。換言之，LLM 的附加價值集中於**構詞複雜性的偵測**，而精確公式的精確率仍為瓶頸（第 4.4節）。值得注意的是，整體準確率相近的四款模型 F1 差距達 19 點，這是無法單純從整體準確率看出的差異。

**逐數詞分析** 圖 1以構詞複雜詞的精確率（P）與召回率（R）呈現逐數詞表現。數詞 6–9 為進位制的核心區間，四款模型的 R 普遍高於 P，顯示主要瓶頸在**偽陽性**（將單純詞誤判為合成詞）而非遺漏。數詞 10 尤為突出：多款模型的 P 驟降至 30% 以下，顯示模型傾向將 10 過度分解。

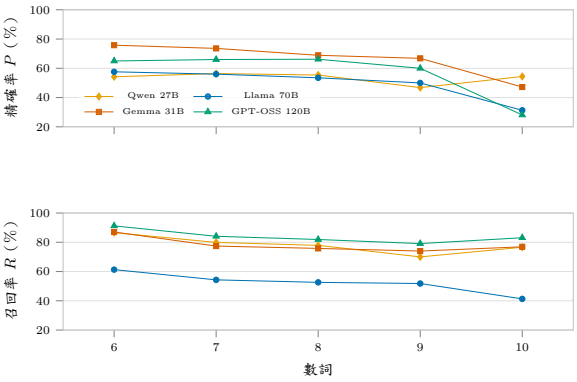


圖 1: 四款大模型逐數詞精確率（$P$）與召回率（$R$），僅顯示構詞複雜詞密集的數詞 6–10（%，zero-shot）。

### 4.2 樣本提示：one-shot 效果與失敗階段

對四款大模型而言，one-shot 範例的增益微弱甚至轉負，因未經顯著性檢定，無法判斷方向：27B +4.5、31B +0.6、120B +0.2 個百分點，Llama 70B 的 one-shot 效果為負（降低 1.4 個百分點，90.5% → 89.1%）：zero-shot 殘餘的 87 個 Echo 全數消除，但 FP 從 174 暴增至 474（增幅 172%），與模型輸出行為一致的是，該範例伴隨 FP 大幅增加，顯示範例可能促使模型更傾向輸出構詞公式。該模型在 zero-shot 錯誤以 Wrong 為主（47%）而 FP 佔比最低（21%），範例使模型將大量原本正確的單純詞判定轉為錯誤公式[2]。

相對而言，圖 2說明了 8B 模型的 one-shot 帶來 42.1 個百分點的準確率提升（14.1% → 56.2%），幾乎完全來自 Echo 的消除（從 4,545 降至 397 個，減少 91%）。然而 FP 反而從 2,091 增至 2,356，顯示模型較能遵循輸出格式，但分析結果未同步改善；單純詞區段因 Echo 消除而大幅回升，但構詞複雜數詞幾乎未改善。

錯誤型態的分布隨模型規模分化（圖 3）。8B 模型在 zero-shot 條件下的主要失敗是 Echo（回抄）：可解析回應的 8,549 個數詞中有 4,545 個（53.2%）直接複製輸入詞形，完全未執行分析。27B 以上的模型 Echo 幾乎消失（70B 有

[2] 在此需要說明的是，唯一 one-shot 範例為賽德克語（乘法型），也可能引入偏差。

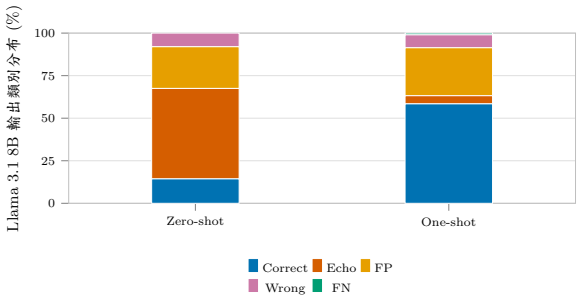


圖 2: Llama 3.1 8B 在 zero-shot 與 one-shot 條件下的輸出類別分布。

87 個，佔 1.0%；27B、31B、120B 均為零），顯示這些模型較能穩定進入要求的分析輸出階段。解析失敗在五款模型均為邊緣現象（8B: 23 筆、70B: 6 筆、31B: 6 筆、27B 與 120B 各 1 筆）。

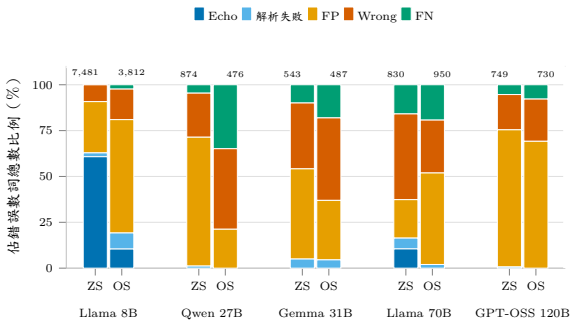


圖 3: 錯誤型態組成（ZS = zero-shot；OS = one-shot）。各柱正規化為該模型、該條件的錯誤數詞總數的 100%，柱頂數字為錯誤數詞總數。

此對比顯示三個遞進的能力階段：任務執行（8B 的主要失敗）→ 格式合規（非瓶頸，≤2.6%）→ 分析正確（大模型間的主要差異所在）。

### 4.3 模式差異：系統性公式辨識錯誤

四款大模型整體準確率相近（90–94%），但錯誤分布分為兩型：27B、31B、120B 相似，70B 呈現不同的錯誤分布。圖 4顯示難度差異：減法（$10-n$）錯誤率均超 50%，高於加法（$5+n$, 5–43%）與乘法（12–64%）；加法內部錯誤率隨操作元增大而遞增。舉例而言，5+1 為最常見的構詞公式（224 種語言），四款模型的錯誤率分別為 GPT-OSS 120B 5.4%、Qwen 27B 與 Gemma 31B 各 8.9%、Llama 70B 36.0%，後者約為其餘模型的四至七倍。混淆矩陣（附錄 B）呈現運算類型的分類傾向。

**系統性公式錯誤** Wrong 的主要來源為兩類。其一為**運算方向錯認**：減法（$10-n$）與加法

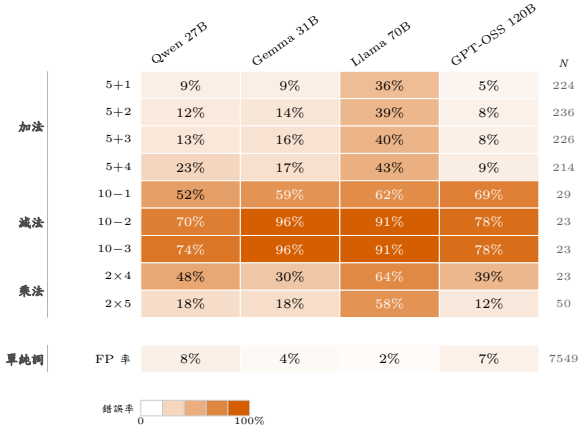


圖 4: 各公式逐模型錯誤率熱力圖（zero-shot）。格內數字為錯誤率；$N$ 為該公式出現次數。

（$5+n$）的混淆呈雙向且因模型而異（附錄 B），70B 有 64% Wrong 將加法誤判為減法，31B 與 120B 則相反。此混淆有構詞學基礎：5+4 與 10−1 結果相同，詞形同時與 5 和 10 部分相似時，模型無法僅從詞形判定方向。其二為**基底錯認**：選擇算術等值但基底不正確的分解（如金標準 5+3 被輸出為 6+2），在 31B 與 120B 各佔排除方向錯認後的 51%。可以看到模型判斷數詞是否為合成詞的能力，優於判斷具體的基底與運算類型。

### 4.4 輔助價值：候選篩選

核心模型的偽陽性中，部分可能對應金標準未標記的構詞模式。為從偽陽性中篩出符合條件的潛在候選，本研究設計三階段過濾：(1) **跨模型一致**，僅取 Gemma 31B 與 GPT-OSS 120B 同時判為構詞複雜的位置；(2) **構詞透明**，成分數詞的詞形須在目標詞形中可見（詞根匹配），並排除自我指涉（成分即目標本身）與類別前綴（出現於半數以上數詞的共用前綴）；(3) **回核金標準**，檢視標準欄位與特殊自由文字欄位（53 種語言使用），排除已記錄者。不要求算術一致（分解須等於該數字）：有些候選的表面形式符合多個數詞成分的匹配條件，卻不符合標準算術運算（如 9 同時匹配 2 與 8，但 $2+8=10$）。

在 zero-shot 條件下，由 94 個雙模型一致位置進一步篩得 29 個值得進一步檢查的構詞分析候選，其中 3 筆雖不符合算術關係，仍因通過形式匹配條件而予以保留；one-shot 條件下則篩得 23 個。這些候選可作為後續複合詞辨識與構詞分析的檢查參考，協助語言學者檢視金標準與模型分析之間的差異，但由於尚未經語言學專家或其他獨立方法驗證，仍不能確認其是否為金標準的漏標分析。表 8列舉數例。

| 語言 | 詞 | 詞形 | AI | 成分 |
|---|---|---|---|---|
| S Raglai | 9 | *salapat* | 1+8 | *sa+lapat* |
| Tandia | 10 | *marusibě* | 2×5 | *arusi+márási* |
| W Uvea | 10 | *limaona-tupu* | 5+5 | *lima* |

表 8: AI 辨識之構詞候選（金標準未標記，節選）。

**候選例示與篩選限制**　表 8列出通過篩選的形式例子。以南拉格萊語 9（*salapat*）為例，兩款模型均提出 1+8 分析，且表面形式分別可匹配 *sa* 與 *lapat*；然而，本研究未進一步查核其詞素分析或相關語言資料，故此例僅能視為待驗證候選，不能據此判定金標準漏標。相對地，馬西旺語（Masiwang）的數詞分類詞前綴 *ʔain-* 可能被 Gemma 誤解為 1 而拆成 $1{+}n$ 系統；GPT-OSS 未作相同判定，因而未通過跨模型一致性條件。這說明一致性篩選可排除部分模型特有的分析，但不能取代語言學複核。

**非算術構詞**　少數語言的構詞組成雖形態透明，卻不對應標準算術運算：回語（Tsat）與 Haroi 語的 9 皆含 2 與 8 的詞素，但 $2{+}8{=}10{\neq}9$，無任何標準運算成立。模型面臨形態正確與算術正確不可兼得的困境。本研究將此類案例排除計分：2 格以 & 明確標記，另有數格因相同性質歸入前述標註者存疑格（54 格）。整體影響極少，但日後若將框架擴展至非算術構詞模式可作參考。

**AI 輔助標記的定位**　LLM 的偵測召回率為 88–97%（啟發式基線 57%），但精確率僅 52–66%。本研究的結果顯示，LLM 輸出經形式條件篩選後可產生待查候選；由於候選尚未經獨立語言學審核，本研究不能據此評估其作為資料庫品質檢核工具的實際效益。

**質性案例：基底切分的歧義**　文馬維斯語（Vinmavis）的 6–9 提供了一個較複雜的案例。從表面形式觀察，7、8、9 分別包含與 2、3、4 對應的成分，因此一種自然的分析是將其分別切分為 5＋2、5＋3、5＋4；在此分析下，6 可分析為 5+，即「1」未明寫，因加法結構已由詞形中的加號標示。此分析即為本研究採用的金標準。

然而，若優先考慮詞形之間的形式對應，而非其數值關係，亦可提出另一種分析：將 7、8、9 分別切分為 6＋2、6＋3、6＋4。雖然這些式子的數值並不成立，但此分析保留了 7–9 與 6 之間的共同詞形成分，因此可視為一種具有形式依據的基底分析。Qwen 的輸出呈現了這一模式。

Gemma 與 GPT-OSS 兩者均將 6 作為 7–9 的基底，但僅 Gemma 的後三項與數值遞增關係一致。GPT-OSS 在 9 的分析中保留了與 4 對應的詞形成分，與 Qwen 的 6＋4 分析相同。由於本研究未取得模型的 chain-of-thought，無法進一步判定這些輸出背後的分析過程。相比之下，Llama 的輸出則呈現 1 與 4 的成分對應錯置。

此案例顯示，數詞構詞分析可能同時涉及詞形切分與數值對應兩個層面。部分模型雖未產生符合金標準的公式，卻輸出了具有形式依據的替代分析。因此，以單一金標準公式評估模型時，可能將部分具有語言學依據的分析歸為錯誤，顯示模型評估中的「錯誤」與「合理替代分析」之間仍需區分。

| # | 詞形 | 金標準 | Qwen | Llama | Gemma | GPT-OSS |
|---|---|---|---|---|---|---|
| 1 | *seβax* | 1 | 1 | 1 | 1 | 1 |
| 2 | *i-ru* | 2 | 2 | 2 | 2 | 2 |
| 3 | *i-tl* | 3 | 3 | 3 | 3 | 2+1 |
| 4 | *i-βah* | 4 | 4 | 4 | 4 | 2+2 |
| 5 | *i-lim* | 5 | 5 | 5 | 5 | 2+3 |
| 6 | *čou* | 5+1 | 6 | 10−4 | 6 | 6 |
| 7 | *ču-ru* | 5+2 | 6+2 | 7 | 6+1 | 6+1 |
| 8 | *ču-tl* | 5+3 | 6+3 | 8 | 6+2 | 6+2 |
| 9 | *ča-βah* | 5+4 | 6+4 | 10−1 | 6+3 | 6+4 |
| 10 | *βuŋaβil* | 10 | 10 | 10 | 10 | 10 |

表 9: 文馬維斯語（Vinmavis）四模型判定（zero-shot）。底線表示錯誤。

## 5　結論與未來展望

本研究以 899 種南島語言基礎數詞評估五款開放權重 LLM 的數詞構詞推斷能力。主要發現如下：(1) 8B 模型過半回應為回抄（Echo），未能執行分析；27B 以上模型的 Echo 率降至近零，較能穩定進入要求的分析輸出階段。(2) 逐步排除 Echo 與解析失敗後，可區分任務執行、格式合規與分析正確三個能力階段；one-shot 範例使 8B 的準確率提升 42 個百分點，主要歸因於 Echo 減少，對較大型模型則未帶來一致增益。(3) 模型對構詞複雜數詞的偵測召回率為 88–97%，高於啟發式基線的 57%，但精確率僅 52–66%；大模型的錯誤較傾向於過度分析，將單純詞誤判為合成詞，並出現加法與減法混淆及公式基底誤判。(4) 以跨模型一致性與構詞透明度過濾偽陽性後，得到 29 個金標準未標記的潛在構詞候選，顯示 LLM 輸出可經形式條件篩選為待查的構詞分析線索；然而，這些候選尚未經語言學專家驗證，因此無法判定其為模型發現的新分析，或是現有金標準的漏標。

本研究為初探性研究，後續研究可從幾個方向檢驗上述結果的穩定性與適用範圍。首先，可建立由多位語言學者獨立標記並經複核的

金標準，以降低單一標記者、子字串匹配與算術驗證可能造成的系統性遺漏，並對 29 個候選進行獨立語言學驗證。其次，可擴展數詞範圍與語言樣本，並考慮親緣語言間的非獨立性，以評估模型表現於不同資料條件下的穩定性；同時對模型間差異進行統計顯著性檢定。第三，可比較不同 one-shot 範例、few-shot 範例、模型規模與 reasoning effort，並進一步區分模型訓練階段的記憶與推論階段的分析對結果的影響；本研究僅使用一個賽德克語範例，且將具推理機制的模型設為最低 reasoning effort，因此相關結果仍可能低估其他設定下的分析能力。最後，可在更完整的資料與專家驗證下，評估 LLM 作為構詞候選生成與初步篩選工具的實際價值。

## 資料與程式碼可取用性

本研究使用的數詞詞形取自 Rosenfelder 的線上資料庫 (Rosenfelder)。實驗程式碼與模型輸出結果已公開於 https://github.com/WilliamLu1218and0630/austronesian-numeral-llm-2026。所有推論於 2026 年 7 月 14–25 日透過 Groq 與 Cerebras 平台免費 API 完成。

## References


Robert Blust. 2009. *The Austronesian Languages*. Pacific Linguistics, Canberra.

Bernard Comrie. 2005. Numeral bases. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Language Structures*, pages 530–533. Oxford University Press, Oxford.

Bernard Comrie. 2013. Numeral bases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Joseph H. Greenberg. 1978. Generalizations about numeral systems. In Joseph H. Greenberg, editor, *Universals of Human Language*, volume 3, pages 249–295. Stanford University Press, Stanford.

Harald Hammarström. 2010. Rarities in numeral systems. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking Universals: How Rarities Affect Linguistic Theory*, pages 11–59. De Gruyter Mouton, Berlin.

Paul Jen-kuei Li. 2006. Numerals in Formosan languages. *Oceanic Linguistics*, 45(1):133–152.

Nai-Wei Lu. 2026. Beyond decimal: Typological patterns and hierarchies in numeral formation (1–10) across 900 Austronesian languages. Oral presentation at the Emory Undergraduate Linguistics Conference (EULC), online.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Mark Rosenfelder. Numbers from 1 to 10 in over 5000 languages. https://www.zompist.com/numbers.shtml. Accessed: 2026-07-14.

Arne Rubehn, Christoph Rzymski, Luca Ciucci, Katja Bocklage, Alžběta Kučerová, David Snee, Abishek Stephen, Kellen Parker van Dam, and Johann-Mattis List. 2025. Annotating and inferring compositional structures in numeral systems across languages. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025)*, pages 29–42, Vienna, Austria. Association for Computational Linguistics.


## A 提示詞全文

以下為實驗使用的完整提示詞。$\{w_i\}$ 代表各語言的第 $i$ 個數詞詞形，由程式自動填入。

### A.1 Zero-shot 提示詞

```
Given the following numeral forms
(1--10) of a language, analyze the
morphological structure of each
numeral and determine its formation
pattern.

For each numeral, write its
decomposition formula:
-- If the numeral is a basic/atomic
word (not derived from other
numerals), write the number itself
(for instance, a language whose word
for eleven is a single unanalyzable
word would get ``11'')
-- If the numeral is derived from
other numerals, write the formula
using +, -, × (for instance, a word
for twelve built from `ten' plus
`two' would be ``10+2''; a word
for twenty built from `two' times
`ten' would be ``2×10''; a word for
nineteen meaning `twenty minus one'
would be ``20-1'')

Numerals:
1: {w1}
2: {w2}
…
10: {w10}

Reply with ONLY the 10 formulas,
comma-separated, in order from 1 to
10.
```

### A.2 One-shot 提示詞

```
Given the following numeral forms
(1--10) of a language, analyze the
```

morphological structure of each numeral and determine its formation pattern.

For each numeral, write its decomposition formula:
-- If the numeral is a basic/atomic word (not derived from other numerals), write the number itself (for instance, a language whose word for eleven is a single unanalyzable word would get ``11'')
-- If the numeral is derived from other numerals, write the formula using +, -, × (for instance, a word for twelve built from `ten' plus `two' would be ``10+2''; a word for twenty built from `two' times `ten' would be ``2×10''; a word for nineteen meaning `twenty minus one' would be ``20-1'')

Example:
Input: 1: kiŋal, 2: daha, 3: teru, 4: sepac, 5: rima, 6: mumuteru, 7: mupitu, 8: mumusepac, 9: muŋari, 10: maxal
Output: 1, 2, 3, 4, 5, 2×3, 7, 2×4, 9, 10
(6 ``mumuteru'' contains mu-mu-teru where teru=3, suggesting 2×3; 8 ``mumusepac'' contains mu-mu-sepac where sepac=4, suggesting 2×4)

Now analyze this language:
1: {$w_1$}
2: {$w_2$}
…
10: {$w_{10}$}

Reply with ONLY the 10 formulas, comma-separated, in order from 1 to 10.

## B 運算類型混淆矩陣

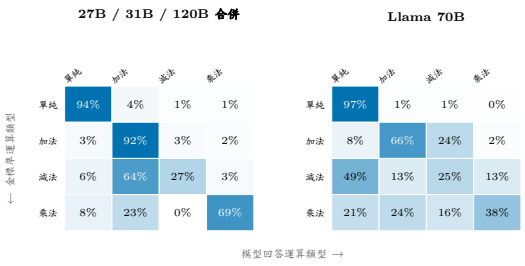


圖 5: 運算類型混淆矩陣（zero-shot）。格內數字為該運算類型的 Wrong 數量。