Historical Textbook Source

Stephen Return Riggs
Dakota Grammar &
Dictionary (1890)
665 pages, Internet Archive

Phase 1: VLM Extraction

Grammar Rules
Extraction
Images 31-92 (62 pages)
Claude Sonnet 4.5
1,036 rules identified

Dictionary Extraction
Images 93-440
(348 pages)
~10,000
{dakota:english} pairs

Phase 2: Data Processing

RL Task Generation
1 rule → 5.5 tasks
5,657 total tasks

Synthetic Data
Generation
Bi-directional pairs
{english:dakota}
Stoney Nakoda
methodology

Validation

Phase 3: RL Training Loop

Verification Rules

Grammar Gym
Environment
PrimeIntellect GRPO
Multi-turn verification
TOPLOC character validation

Compositional Rewards
Character preservation (40%)
Affix accuracy (40%)
Semantic correctness (20%)

Curriculum Learning
Easy (1,998
tasks) → 80%
Medium (2,155
tasks) → 75%
Hard (398
tasks) → 70%

Phase 4: Trained Model

Grammar-Aware
Dakota Model
Qwen2.5-7B-
Instruct + LoRA
Verified special
character handling
Morphologically
accurate