

Sina Ahmadi (sina.ahmadi@uzh.ch) - University of Zurich

This presents a contribution of 100 instances to the Multilingual Physical Commonsense Reasoning Datasets shared task, specifically for Central Kurdish (ckb), also known as Sorani. Central Kurdish is a variety of the Kurdish languages, an Indo-European language spoken by over 30 million speakers across the Middle East. Many regions where Kurdish is spoken lack institutionalized support, and speakers often struggle to access formal education in their native language. Despite these challenges, there exists significant enthusiasm for Kurdish language preservation and development, particularly in the field of Kurdish NLP. Kurdish remains a low-resource language with substantial opportunities for advancement in language and speech technologies.

The dataset construction process began with careful consideration of culturally authentic domains that would effectively test physical commonsense reasoning while remaining true to Kurdish cultural contexts. Rather than adapting existing English datasets or relying on translation, I opted for a completely original approach that draws from traditional Kurdish life and culture.

The primary focus centers on village life and traditional practices, domains where Kurdish possesses an exceptionally rich and nuanced vocabulary. I assume that these traditional domains would present unique challenges for language models, as they involve culture-specific phenomena that are often poorly documented in digital resources and unlikely to be well-represented in typical training corpora. Unlike modern technological and scientific concepts, which are increasingly documented in Kurdish Wikipedia and other digital platforms, traditional Kurdish practices and their associated terminology remain largely within oral tradition and community knowledge.

All 100 instances were manually created from scratch without the use of LLMs, translation tools, or adaptation from existing datasets. This approach ensures cultural authenticity and linguistic accuracy while creating evaluation items that require genuine command of both the Kurdish language and Kurdish cultural knowledge.

The dataset encompasses diverse aspects of traditional Kurdish life, distributed across the following domains (provided as an additional column in the tsv file):

- Culinary practices (26 instances): Traditional cooking methods, food preparation techniques, and kitchen practices
- Handcrafts (17 instances): Traditional crafting techniques, tool usage, and artisanal practices
- Agriculture (16 instances): Farming methods, crop cultivation, and agricultural tools
- Village life (12 instances): Daily activities, community practices, and rural living
- Animal husbandry (11 instances): Livestock care, animal behavior, and pastoral practices
- Customs and traditions (8 instances): Cultural ceremonies, social practices, and community rituals
- Traditional games (5 instances): Folk games, recreational activities, and play practices
- Know-how (2 instances): General life skills and practical problem-solving
- Architecture (1 instance): Traditional building methods and construction practices

This thematic distribution reflects the rich cultural heritage of Kurdish communities while providing a comprehensive evaluation framework for physical reasoning capabilities in culturally specific contexts. Each instance requires not only linguistic competence but also deep cultural knowledge that extends beyond surface-level translation or general world knowledge, making this dataset particularly valuable for evaluating genuine multilingual and multicultural AI capabilities.