SCCT Annotation Guidelines

Hello, dear annotator, and thank you for your time!
Here is a brief background for the task and the specific annotation guidelines. This task is a part of the DharmaBench paper and evaluation benchmark for Sanskrit and Tibetan.

Introduction

Task: Classifying canonical literature (SCCT)

Goal: To distinguish between Tibetan allochthonous literature into the two categories of scripture and non-scripture.

Background: This task classifies Tibetan canonical literature (allochthonous) into two primary categories: “scriptures” and “non-scriptures.” Its primary objectives are to (a) evaluate LLM performance in canonical literature classification into these two broad categories, and (b) illuminate the evolutionary processes of works that are considered scriptures by the tradition (allochthonous texts) as well as para-canonical works through (works whose provenance is debated) linguistic analysis. The task aims at identifying distinctive features characteristic of each category, including compositional styles, syntax, vocabulary, and other linguistic markers. In particular, such an analysis should help reveal recurring stylistic patterns and compositional strategies that a scripture creator possibly employed to achieve two critical goals: canonical acceptance and popular success.

Real-world relevancy: By tracing and mapping these linguistic patterns, this task has the potential to provide insights into how textual authority was constructed and maintained within Indic and Tibetic literary traditions.

Example

1. Text: “རྒྱལ་པོའི་ཁབ་ཀྱི་གྲོང་ཁྱེར་ཆེན་པོར་བྱོན་ནས་དེ་རྒྱལ་པོའི་ཁབ་ཀྱི་གྲོང་ཁྱེར་ཆེན་པོ་ན་བྱ་རྐོད་ ཀྱི་ཕྱང་པོའི་རི་ལ་བཞུགས་ཤ། འདི་ཕྱར་བཅོམ་ལྡན་འདས་དེ་ནི་བཞིན་གཤེགས་པ་དགྲ་བཅོམ་པ་ཡང་དག་པར་རྫོགས་པའི་སངས་རྒྱས་རིག་པ་དང་ཞབས་སུ་ལྡན་པ། བདེ་བར་གཤེགས་པ་འཇིག་རྟེན་མཁྱེན་པ། སྐྱེས་བུ་འདུལ་བའི་ཁ་ལོ་སྒྲུར་བ། ཐུ་ན་མེད་པ། ལྷ་དང་མི ་རྣམས་ཀྱི་སྟོན་པ། སངས་རྒྱས་བཅོམ་ལྡན་འདས་ཡིན་ཏེ། དེ་ཚོས་སྟོན་པ་ནི་ཚངས་པར་སྤྱོད་པ། ཐོག་མར་དགེ་བ། བར་དུ་དགེ་བ། ཐ་མར་དགེ་བ། དོན་བཟང་པོ། ཚིག་འབྲུ་བཟང་པོ། མ་འདྲེས་པ་ཡོངས་སུ་རྫོགས་པ། ཡོངས་སུ་དག་པ། ཡོངས་སུ་བྱང་བ་རབ་ཏུ ་སྟོན་ཏོ་ཞེས་དགེ་སློང་གོ་ཤ་དེའི་དགེ་བ་གྲགས་པའི་སྒྲ་ཚིགས་སུ་བཅད་པ་རྒྱ་ཆེན་པོ་དེ་ལྟ་བུ་ཡང་མངོན་པར་བྱང་བ་ཁྱིམ་བདག་ཚོང་དཔོན་བཟང་སྐྱོང་གིས་ཐོས་ནས་དེ་འདི་སྙམ་དུ་སེམས་ཏེ། དེ་བཞིན་གཤེགས་པ་དགྲ་བཅོམ་པ་ཡང་དག་པར་རྫོགས་པའི་སངས་རྒྱས་དེ་ལྟ་བུ་དག་མཐོང་ན་ཡེགས་པར་འགྱུར་གྱིས་བདག་དགེ་སློང་གོ་ཤ་དེ་བལྟ་

བའི་ཕྱིར་འགྲོའོ་སྐྱམ་མོ། །དེ་ནས་ཁྱིམ་བདག་ཚོང་དཔོན་བཟང་སྐྱོང་ཁྱིམ་བདག་ལྟ་བརྒྱ་ཙམ་དང་ལྡན་ཅིག་ཏུ་བཅོམ་ལྡན་འདས་ལ་བལྟ་བའི་ཕྱིར་རྒྱལ་པོའི་ཁབ་ཀྱི་གྲོང་ཁྱེར་ཆེན་པོ་ནས་བྱུང་ངོ། །" Class: SCR

2. Text: "སངས་རྒྱས་ལ་མ་དད་ཅེས་བྱ་བ་ནི་ཡིད་མ་ཆེས་པས་བརྟགས་ཤིང་མ་དད་པའི་དོན་ཏོ། །སྐྱ་གཞན་ལ་བརྟེན་ཞེས་བྱ་བ་ནི་འཇིག་རྟེན་པའི་སྐྱ་ལ་བརྟེན་ཅིང་སྐུ་སྟེགས་པའི་ཕྱོགས་ འཇིན་པའོ། །ང་ཡི་གསང་སྔགས་བརླག་ན་ཕྱང་བར་འགྱུར། །ཞེས་བྱ་བ་ནི་འཆི་བ་དང་འཇིག་རྟེན་ལ་རོལ་དུ་ཡང་དན་སོང་དུ་སྐྱེ་བར་འགྱུར་རོ། །བསམ་པ་ཞན་ཅིང་ཞེས་བྱ་བ་ནི་སེམས་ཤུམ་པའོ། །ཉམས་ཀྱིས་མི་སྡུག་པས་ཞེས་བྱ་བ་ནི་ཡོ་བྱད་དང་མི་ལྡན་ཞེས་བྱ་བའི་དོན་ཏོ། །ཁྱིམ་བདག་ཚོང་དཔོན་བཟང་སྐྱོང་མ་འགྲོར་ཉེ་སེམས་ཤུམ་ནས། དཀྱིལ་འཁོར་འཇིག་ཚེན་འཇུག་མ་ཐུས་ན་ཡང་། ཞེས་བྱ་བའི་དོན་ཏོ། །གཅིག་ཙམ་ཞིག་ཀྱང་དད་པས་མཐོང་སྐྱར་ན། །ཞེས་བྱ་བ་ནི་ཡོ་ བྱད་མེད་པའི་སྐྱབ་ས་པོ་དབུལ་པོས་ཀྱང་དཀྱིལ་འཁོར་དེ་དག་ཆམས་ལས་ཉམས་ཀྱིས་གང་སྡུག་པ་གཅིག་ནི་དེས་པར་འདྲི་དགོས་པར་གསུངས་ཏེ། དོན་དེ་ཉིད་རབ་ཏུ་གྲུབ་པར་བྱེད་པའི་ཚུལ་ཚེན་པོ་ལས་ཀྱང་ཇི་སྐད་དུ།" Class: NSRC

Annotation guidelines

1. The samples will consist of heuristically cut sections of text from canonical and paracanonical sources (ACIP & rKTs). They will be provided in a CSV. file with a pre-assigned label. Read the sample of the text carefully and determine if the label is correct.
2. If the sample contains any material that is not part of the original text, such as table of contents, editorial notes, translator's remarks, and so forth, reject the sample.
3. If the sample is too short (i.e., shorter than two full sentences), reject the sample.
4. If there are any conversion errors in the Tibetan (due to erroneous input in the original material, for example), if the error is minor, correct it. If there are many conversion errors that cannot be resolved easily, reject the sample.

Edge cases and common mistakes

1. In cases of ambiguity, please consult with the team and your supervisors.
    a. If you can't reach an agreement, discard the sample.

Something is off?

Reach out to us.

Thank you very much!