CUBE Datacard

CUBE

(Cultural **Be**nchmark for T2I models)

Data Card Authors: Redacted for anonymity

CUBE is created to evaluate the cultural competency of text-to-image models. CUBE consists of CUBE-CSpace which is a broad collection of ~300k cultural artifacts along with their national association and domain, spanning 8 countries and 3 cultural domains. A subset of popular artifacts from CUBE-CSpace is made into T2I prompts, called CUBE-1K, intended to evaluate cultural awareness of T2I models. CUBE (CUBE-CSpace and CUBE-1K) is intended to be used for evaluation purposes only with a potential to guide future cultural data collection.

Data Card

Publishing Status

- Internal Only: No external references available
- Internal Only: Dataset is externally unpublished but external references are available
- Internal Version: A version of the dataset is externally published
- External Version: This is an externally published version of an internal dataset
- External Only: Dataset is external only

Team(s)

Redacted for anonymity

Contact Detail(s)

Redacted for anonymity

Data Subject(s)

- Sensitive Data about people
- Non-Sensitive Data about people
- Data about natural phenomena
- Data about places and objects
- Synthetically generated data
- Data about systems or products and their behaviors
- Unknown
- Others:
- Please Specify

Dataset Snapshot

CUBE-CSpace

Dataset Characteristics	Value
Size of dataset	88.8MB
Number of Instances	295,287
Number of Fields	8

Above: Dataset Statistics of CUBE-CSpace

Field Description:

Description of Content

CUBE-CSpace

A collection of ~300k cultural artifacts that are nodes in the WikiData Knowledge Base.

The artifacts were collected by traversing the KB from manually selected parent nodes along the 'instance of' (P31) and 'subclass of' (P279) edges.

Each cultural artifact additionally contains a country association that is present in field 'country-of-origin' (P495) and/or 'country' (P17)

Field 1. P31	Wikidata property 'instance of' of the artifact node
Field 2. P279	Wikidata property subclass of' of the artifact node
Field 3. P495	Wikidata property 'country-of-origin' of the artifact node
Field 4. P17	Wikidata property 'country' of the artifact node
Field 5. P361	Wikidata property 'part of' of the artifact node
Field 6. id	Wikidata id of the artifact node e.g. Q920940
Field 7. parent node name	Name of parent node in Wikidata. Indicates node from where the traversal was started to reach the artifact node.
Field 8. name	Name of artifact node in Wikidata

CUBE-1K

Dataset Characteristics	Value
Size of dataset	321 KB
Number of Instances	1000
Number of Fields	5

Above: Dataset Statistics of CUBE-1K

Field Description:

Field 1. country	Country of association of the

The obtained cultural artifacts were further filtered and extended using GPT- 4-Turbo by employing self-refinement and self-critiquing strategies

CUBE-1K

CUBE-1K consists of 1000 text-to-image model prompts that are created by selecting a subset of artifacts from CUBE-CSpace based on popularity estimates from Google Search API.

Prompts in CUBE-1K are created using simple templates for different domains. Each prompt has a mention of a single cultural artifact.

Countries covered: Brazil, India, Japan, Nigeria, Turkey, Italy, United States, France

Domains: Cuisine, Landmarks, Art

	artifact	
Field 2. domain	Domain of artifact	
Field 3. artifact name	Name of artifact	
Field 4. prompt	T2I prompt containing the artifact	

Additional Notes: CUBE-1K contains a subset of 1000 artifacts from CUBE-CSpace. Both datasets cover 8 countries

Primary Data Modality	Example Data Point	Data Fields

- Image Data
- Text Data
- Tabular Data
- Audio Data
- Video Data
- Time Series
- Graph Data
- Geospatial Data
- Multimoda
  - Please specify
- Unknown
- Others
  - Please specify

Field	Value
P31	['Q2625877', 'Q15846908']
P279	['Q178', 'Q27555297', 'Q192874']
P495	['Q38']
P17	['Q38']
P361	-
id	Q20026
parent node name	dish
name	Spaghetti

Above: Example datapoint in CUBE-CSpace

Additional Note: Values starting Q38 is the Italy node in WikiData. All values starting with 'Q' represent WikiData nodes

Field	Value
country	Italy
domain	Cuisine

Field name	Example field values	Description
P31	'Q2095', 'Q7802'	Property 'instance of'
P279	'Q13270'	Property 'subclass of'
P495	'Q8646', 'Q155'	Property 'country-of-origin'
P17	'Q155', Q38	Property 'country'
P361	'Q614394'	Property 'part of'
id	Q16493207	Wiki ID of artifact
parent node name	type of dance, dish	Name of parent node
name	Odissi, Osechi	Artifact name

Above: Field description and example field values in CUBE-CSpace

Additional Note: Values starting with 'Q' represent WikiData node IDs. For example, Q155 is the Brazil node

Field name	Example field values	Description
country	Brazil, India	Artifact country

artifact	Spaghetti
prompt	A high resolution image of Spaghetti from Italian cuisine, realistic

Above: Example datapoint in CUBE-1K

		of association
domain	Cuisine, Landmark	Artifact domain
artifact	Dosa, Eiffel Tower	Artifact name
prompt	A high resolution image of Dosa from Indian cuisine, realistic, A panoramic view of Palazzo Eiffel Tower in France, realistic	T2I prompt

Above: Field description and example field values in CUBE-1K

Purpose(s)	Domain(s) of application	Primary Motivation
• Monitoring • Research • Production • Others ○ Please specify	**Domains** Text-to-image models, fairness, cultural inclusivity, data collection **Problem Space** Measuring cultural competency in existing text-to-image models.	• Existing T2I benchmarks evaluate aspects such as image-text alignment, fidelity and aesthetics. Lack of a framework or benchmark to evaluate cultural competence of T2I models • Lack of any existing broad geo-coverage of cultural artifact collection to serve as grounding. • Data sources like these will serve as a great starting point and can compliment participatory approaches for cultural data collection

Dataset Use(s)	INTENDED AND/OR SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
• Safe for production use • Safe for research use • Conditional use- some unsafe applications • Only approved use • Others ○ Please specify	• Evaluate cultural awareness of T2I models • Serve as grounding to evaluate cultural diversity of T2I models • Compliment future cultural data collection approaches	• As a resource for training production systems • As a benchmark for assessing fairness or lack of fairness in T2I models • As an exhaustive resource of cultural artifacts across domains and countries considered

SAFETY OF USE WITH OTHER DATA	ACCEPTABLE TRANSFORMATIONS	BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET

Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others* (Please specify)	Joining with other datasets Subsampling and splitting Filtering Joining input sources Cleaning missing values Anomaly detection Grouping and summarizing Scaling and reducing Statistical transformations Redaction or Anonymization Others (please specify)	N/A (we have not attempted to use this dataset with other datasets, but we do not anticipate any issues)

VERSION STATUS	DATASET VERSION	MAINTENANCE PLAN
Regularly Updated New versions of the dataset have been or will continue to be made available. Actively Maintained No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data. Limited Maintenance The data will not be updated, but any technical issues will be addressed. Deprecated This dataset is obsolete or is no longer being maintained.	**Current Version** 1.0 **Last Updated** 06/2024 **Release Date** 06/2024	• We might add harder (less popular) artifacts to CUBE-1K intended to evaluate T2I models. • We might scale our automated cultural artifact extraction from KB to other domains. • We will address any potential issues that may arise during the usage of our dataset

ACCESS	RETENTION	WIPEOUT
The data will be accessible under CC BY-NC-SA 4.0 License	**N/A**	**N/A**

Provenance | Collection | Method(s) used	Provenance | Collection | Methodology detail(s)	Provenance | Collection | Data Processing

• API
• Artificially Generated
• Crowdsourced - Paid
• Crowdsourced - Volunteer
• Vendor Collection Efforts
• Scraped or Crawled
• Survey, forms or polls
• Taken from other existing datasets
• Unknown
• To be determined
• Others
  ○ Please Specify
	Source: The data is obtained from WikiData (https://www.wikidata.org/wiki/Wikidata:Introduction). WikiData dump of April 2024 was used.
	
	Methodology in Brief:
	We traverse the WikiData knowledge graph and extract nodes (and corresponding properties) that have cultural artifacts associated with a country
	
	Is this source considered sensitive or high-risk?
	No
	Dates of Collection: [04 2024 - 04 2024]
	
	Primary modality of collected data:
	• Image Data
	• Text Data
	• Tabular Data
	• Audio Data
	• Video Data
	• Time Series
	• Graph Data
	• Geospatial Data
	
	Update frequency for collected data:
	• Yearly
	• Quarterly
	• Monthly
	• Biweekly
	• Weekly
	• Daily
	• Hourly
	• Static
	• Others
	  ○ Please specify
		Simple text-based data filtering was used to remove noisy data. Non-latin node names were filtered. Nodes with inconsistent formatting were filtered out. GPT-4-Turbo was employed to remove incorrectly annotated (country association) artifacts. Existing techniques such as self-refinement and self-critiquing were performed with GPT-4-Turbo to fill in missing artifacts from a country and domain.

Provenance | Collection Criteria | Data Selection	**Provenance | Collection Criteria | Inclusion Criteria**	**Provenance | Collection Criteria | Exclusion Criteria**
Wikidata is a structured knowledge base that can be freely edited by humans and machines. It offers a broad geo-coverage of cultural artifacts we are interested in. An algorithmic approach to extract data from a KB offers additional benefits	A cultural artifact node must be a child of one of the manually curated parent nodes in Wikipedia. A cultural artifact node must also contain one or both of 'country-of-origin' and 'country' property. Nodes with inconsistent format are not included.	Nodes with inconsistent format in Wikidata were filtered out. The artifacts were passed through GPT-4-Turbo to test if the properties or national association were indeed correct. Erroneous nodes were filtered out.

such as ease of scaling to other domains LLMs are a great automated tool to refine data and filter data that have a reasonable amount of world knowledge.	Inclusion Criteria for CUBE-1K: Popularity of the artifacts was estimated using the number of search results. This was used to guide the artifact selection to be made into T2I prompts	
SENSITIVE HUMAN ATTRIBUTES	**SOURCE(S) OF HUMAN ATTRIBUTES**	**RATIONALE FOR COLLECTING HUMAN ATTRIBUTES**
Race Gender Ethnicity Socio-economic status **Geography** Language Sexual Orientation Religion Age **Culture** Disability Experience or Seniority Others (please specify)	**Geography:** Each cultural artifact within the collection is associated with a specific country. This association is determined using WikiData and subsequently verified through an LLM. However, it is important to acknowledge that the dataset may contain potentially inaccurate or debatable geographical associations for certain artifacts. **Culture:** The artifacts comprising the collection represent nodes within Wikidata, specifically identified as children of manually selected nodes deemed to contain cultural information. It is acknowledged that this process may introduce noise, potentially resulting in the inclusion of artifacts lacking significant cultural relevance or the omission of niche artifacts. Therefore, this method of cultural data extraction, while serving as a valuable starting point, should not be considered exhaustive.	In line with recent literature, we focus on cultures demarcated by national boundaries. Hence we collect the country association of each node. The cultural artifact collection approach helps in building a broad-coverage collection of cultural artifacts from regions around the globe which we believe is a valuable resource for evaluation of models. Such a resource is non-existent currently and can foster future data collection strategies.
TRANSFORMATIONS APPLIED	**FIELDS TRANSFORMED**	**LIBRARIES AND METHODS USED**
Anomaly Detection Cleaning Mismatched Values Cleaning Missing Values Converting Data Types Data Aggregation Dimensionality Reduction Joining Input Sources Redaction or Anonymization Others* (*Please specify)	• Data Points with incorrect country association or incorrect information was dropped • GPT-4-Turbo was used to fill in for the missing values. These artifacts only had the 'name' and 'id' field with rest of the fields empty	• Publically available GPT-4-Turbo was used for filtering and expanding artifacts. • Publicly available Google Search API was used to aid artifact sampling for CUBE-1K
SAMPLING METHOD(S)	**SAMPLING CHARACTERISTIC(S)**	• **SAMPLING CRITERIA**

Cluster Sampling Haphazard Sampling Multi-stage Sampling Random Sampling Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others* (*Please specify)	1000 artifacts were sampled to create CUBE-1K prompts.	• An artifact's popularity is estimated using the number of Google search results returned when a local user searches for it. To collect relevant and popular artifact prompts, CUBE-1K artifacts were sampled based on these popularity estimates using weighted sampling, where the probability of an artifact being selected is proportional to its popularity score.

ANNOTATOR TYPE	ANNOTATOR BREAKDOWN	ANNOTATOR DESCRIPTION
N/A (No manual annotation was performed on the dataset)	N/A	N/A

VALIDATION METHOD(S)	VALIDATION BREAKDOWN	DESCRIPTION OF VALIDATION

Data Type Validation
Range and Constraint Validation
Code/cross-reference Validation
Structured Validation
Consistency Validation
Not Validated
Others*
(*potential missing values)
	• Validation involved removing or adding entire data points
	• We perform 2 major steps of validation:
	  ○ 1) Incorrectly annotated data was **filtered** out using GPT-4-Turbo
	  ○ 2) Artifacts that may be **missing** from the KB collection for a particular country and domain, are filled in by GPT-4-Turbo
		1) The first step of validation is to ensure the artifact node adhered to the expected data format and contained readable information
		2) The second step of validation includes verifying if an artifact indeed belonged to a particular domain. for e.g., "Is the Statue of Liberty a landmark?", and verifying if the artifact associates with the country, for e.g, "Is Sushi part of Japanese cuisine". This validation step is performed using GPT-4-Turbo and erroneous points are removed
		3) Potential missing artifacts from a domain and country are identified and added to the collection by GPT-4-Turbo.

ML APPLICATION(S)	EVALUATION - RESULTS	EVALUATION - PROCESS

Evaluating cultural competency in T2I models 1) The prompts in CUBE-1K can be used to test if T2I models can correctly depict cultural artifacts unique to different cultures 2) CUBE-CSpace serves as the grounding to evaluate the diversity of T2I models to cultural prompts	N/A	N/A

Terms of Art

Concepts and Definitions referenced in this Data Card

Cultural Artifacts	Domains	Self Refinement & Self-Critiquing
Definition: A cultural artifact is any tangible or intangible object, creation, or expression that provides information about the culture For e.g, Sushi, Kimono, and Jagannath temple are example of artifacts	Definition: A domain is a distinct category of cultural artifacts For e.g, cuisine, landmarks and clothing	Definition: Strategies in literature To iteratively refine and improve LLM responses (https://arxiv.org/abs/2303.17651) (https://arxiv.org/abs/2310.16523)

Reflections on Data

Ingrained Biases in Google Search Results and LLM validation	Since our data is collected and verified automatically, our benchmark may contain biases ingrained in tools used to create this benchmark.
Cultural artifacts not captured by the dataset	We obtain artifacts from a large database WIkiData and use GPT-4-Turbo to fill in gaps. There could exist cultural artifacts not captured in our dataset, despite our best efforts on the contrary.