Package: cat.survey (via r-universe)

June 4, 2026

Title Survey Response Classification with LLMs

Version 0.1.2

Description R interface to the Python cat-survey package. Classifies, extracts, and explores open-ended survey responses using LLMs. A thin domain wrapper around cat.stack that adds the survey_question parameter for survey-specific context.

License GPL (>= 3)

URL https://christophersoria.com/cat-llm/cat.survey/, https://github.com/chrissoria/cat-llm

BugReports https://github.com/chrissoria/cat-llm/issues

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

SystemRequirements Python (>= 3.9), pip

Imports reticulate (>= 1.28), cat.stack (>= 0.1.0)

Suggests testthat (>= 3.0.0), knitr, rmarkdown

VignetteBuilder knitr

Config/testthat/edition 3

Config/pak/sysreqs libpng-dev python3

Repository https://chrissoria.r-universe.dev

Date/Publication 2026-06-04 16:16:50 UTC

RemoteUrl https://github.com/chrissoria/cat-llm

RemoteRef main

RemoteSha f2d83209be8d621fceb422d434fb5b3b98fe301b

RemoteSubdir r-package/cat.survey

Contents

`classify` *Classify survey responses using LLMs*

Description

Wraps the Python `cat_survey.classify()` function. Adds `survey_question` context to the base cat.stack classification engine.

Usage

```
classify(
  input_data,
  categories,
  survey_question = "",
  description = "",
  add_other = "prompt",
  check_verbosity = TRUE,
  api_key = NULL,
  user_model = "gpt-4o",
  mode = "image",
  creativity = NULL,
  safety = FALSE,
  chain_of_verification = FALSE,
  chain_of_thought = FALSE,
  step_back_prompt = FALSE,
  context_prompt = FALSE,
  thinking_budget = 0L,
  example1 = NULL,
  example2 = NULL,
  example3 = NULL,
  example4 = NULL,
  example5 = NULL,
  example6 = NULL,
  filename = NULL,
  save_directory = NULL,
  model_source = "auto",
  max_categories = 12L,
  categories_per_chunk = 10L,
  divisions = 10L,
  research_question = NULL,
```

```
  models = NULL,
  consensus_threshold = "unanimous",
  use_json_schema = TRUE,
  max_workers = NULL,
  fail_strategy = "partial",
  max_retries = 5L,
  batch_retries = 2L,
  retry_delay = 1,
  row_delay = 0,
  pdf_dpi = 150L,
  auto_download = FALSE,
  prompt_tune = NULL,
  tune_iterations = 1L,
  tune_ui = "browser",
  tune_optimize = "balanced"
)
```

Arguments

input_data	A character vector, list, or data.frame column of survey responses to classify.
categories	A character vector of category names, or "auto" to infer categories from the data.
survey_question	Character. The survey question text. Default "".
description	Character. Additional context for the classification task. Default "".
add_other	Logical or "prompt". Controls addition of an "Other" category. Default "prompt".
check_verbosity	Logical. Check category descriptions. Default TRUE.
api_key	API key for the model provider (single-model mode).
user_model	Character. Model name. Default "gpt-4o".
mode	Character. PDF processing mode. Default "image".
creativity	Numeric or NULL. Temperature. Default NULL.
safety	Logical. Save progress after each item. Default FALSE.
chain_of_verification	Logical. Default FALSE.
chain_of_thought	Logical. Default FALSE.
step_back_prompt	Logical. Default FALSE.
context_prompt	Logical. Default FALSE.
thinking_budget	Integer. Extended thinking budget. Default 0L.
example1, example2, example3, example4, example5, example6	Optional few-shot examples.

`filename` Character or `NULL`. Output CSV filename.

`save_directory` Character or `NULL`. Output directory.

`model_source` Character. Provider hint. Default `"auto"`.

`max_categories` Integer. Max categories for auto mode. Default `12L`.

`categories_per_chunk` Integer. Default `10L`.

`divisions` Integer. Default `10L`.

`research_question` Character or `NULL`. Optional research context.

`models` List of model specs for ensemble mode.

`consensus_threshold` Character or numeric. Default `"unanimous"`.

`use_json_schema` Logical. Default `TRUE`.

`max_workers` Integer or `NULL`. Default `NULL`.

`fail_strategy` Character. Default `"partial"`.

`max_retries` Integer. Default `5L`.

`batch_retries` Integer. Default `2L`.

`retry_delay` Numeric. Default `1.0`.

`row_delay` Numeric. Default `0.0`.

`pdf_dpi` Integer. Default `150L`.

`auto_download` Logical. Default `FALSE`.

`prompt_tune` Integer or `NULL`. Rows sampled per APO correction round. Default `NULL`.

`tune_iterations` Integer. APO optimization passes. Default `1L`.

`tune_ui` Character. Correction UI: `"browser"` or `"terminal"`. Default `"browser"`.

`tune_optimize` Character. Metric to optimize: `"balanced"`, `"sensitivity"`, or `"precision"`. Default `"balanced"`.

Value

A `data.frame` with classification results.

Examples

```
## Not run:
results <- classify(
  input_data      = c("Took a new job in Chicago",
                      "Wanted to be closer to grandkids",
                      "Couldn't afford rent in the Bay Area"),
  categories      = c("Job/school", "Family", "Cost of living", "Other"),
  survey_question = "Why did you move?",
  api_key         = Sys.getenv("OPENAI_API_KEY"),
  user_model      = "gpt-4o-mini"
)

## End(Not run)
```

explore *Explore raw categories in survey response data*

Description

Wraps the Python `cat_survey.explore()` function. Returns every category string extracted from every chunk across every iteration – with duplicates intact. Useful for analysing category stability and saturation.

Usage

```
explore(
  input_data,
  api_key,
  survey_question = "",
  description = "",
  max_categories = 12L,
  categories_per_chunk = 10L,
  divisions = 12L,
  user_model = "gpt-4o",
  creativity = NULL,
  specificity = "broad",
  research_question = NULL,
  filename = NULL,
  model_source = "auto",
  iterations = 8L,
  random_state = NULL,
  focus = NULL,
  chunk_delay = 0
)
```

Arguments

`input_data`	A character vector, list, or `data.frame` column of survey responses.
`api_key`	Character. API key for the model provider.
`survey_question`	Character. The survey question text. Default `""`.
`description`	Character. Additional context. Default `""`.
`max_categories`	Integer. Max categories per chunk. Default `12L`.
`categories_per_chunk`	Integer. Default `10L`.
`divisions`	Integer. Number of data chunks. Default `12L`.
`user_model`	Character. Model name. Default `"gpt-4o"`.
`creativity`	Numeric or `NULL`. Temperature. Default `NULL`.

specificity	Character. "broad" or "specific". Default "broad".
research_question	Character or NULL. Optional research context.
filename	Character or NULL. Output CSV filename.
model_source	Character. Provider hint. Default "auto".
iterations	Integer. Number of passes. Default 8L.
random_state	Integer or NULL. Random seed.
focus	Character or NULL. Optional focus.
chunk_delay	Numeric. Seconds between API calls. Default 0.0.

Value

A character vector of every category string extracted.

Examples

```
## Not run:
raw_categories <- explore(
  input_data      = df$open_response,
  survey_question = "What concerns you most about your community?",
  api_key         = Sys.getenv("OPENAI_API_KEY"),
  user_model      = "gpt-4o-mini",
  iterations      = 4L
)
table(raw_categories)

## End(Not run)
```

---

extract *Extract categories from survey responses using LLMs*

Description

Wraps the Python `cat_survey.extract()` function. Discovers and returns a normalised, deduplicated set of categories found in survey response data.

Usage

```
extract(
  input_data,
  api_key,
  survey_question = "",
  description = "",
  input_type = "text",
  max_categories = 12L,
  categories_per_chunk = 10L,
```

```
  divisions = 12L,
  user_model = "gpt-4o",
  creativity = NULL,
  specificity = "broad",
  research_question = NULL,
  mode = "text",
  filename = NULL,
  model_source = "auto",
  iterations = 8L,
  random_state = NULL,
  focus = NULL,
  chunk_delay = 0
)
```

Arguments

input_data	A character vector, list, or data.frame column of survey responses.
api_key	Character. API key for the model provider.
survey_question	Character. The survey question text. Default "".
description	Character. Additional context. Default "".
input_type	Character. Type of input. Default "text".
max_categories	Integer. Maximum final categories. Default 12L.
categories_per_chunk	Integer. Default 10L.
divisions	Integer. Number of data chunks. Default 12L.
user_model	Character. Model name. Default "gpt-4o".
creativity	Numeric or NULL. Temperature. Default NULL.
specificity	Character. "broad" or "specific". Default "broad".
research_question	Character or NULL. Optional research context.
mode	Character. Processing mode. Default "text".
filename	Character or NULL. Output CSV filename.
model_source	Character. Provider hint. Default "auto".
iterations	Integer. Number of passes. Default 8L.
random_state	Integer or NULL. Random seed.
focus	Character or NULL. Optional focus.
chunk_delay	Numeric. Seconds between API calls. Default 0.0.

Value

A named list with counts_df, top_categories, and raw_top_text.

Examples

```
## Not run:
result <- extract(
  input_data      = c("Took a new job in Chicago",
                      "Wanted to be closer to grandkids",
                      "Couldn't afford rent in the Bay Area"),
  survey_question = "Why did you move?",
  api_key         = Sys.getenv("OPENAI_API_KEY"),
  user_model      = "gpt-4o-mini"
)
print(result$top_categories)

## End(Not run)
```

Index