Package: cat.ademic (via r-universe)

June 4, 2026

Title Academic Paper Classification with LLMs

Version 0.1.1

Description R interface to the Python catademic package. Classifies, extracts, explores, and summarizes academic papers using LLMs. A thin domain wrapper around cat.stack that adds journal and topic sourcing parameters for academic literature analysis.

License GPL (>= 3)

URL https://christophersoria.com/cat-llm/cat.ademic/, 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.ademic

Contents

`classify` *Classify academic papers using LLMs*

Description

Wraps the Python `catademic.classify()` function. Adds journal and topic sourcing parameters to the base cat.stack classification engine.

Usage

```
classify(
  categories,
  input_data = NULL,
  api_key = NULL,
  journal_issn = NULL,
  journal_name = NULL,
  journal_field = NULL,
  topic_name = NULL,
  topic_id = NULL,
  paper_limit = 50L,
  date_from = NULL,
  date_to = NULL,
  polite_email = NULL,
  journal = NULL,
  field = NULL,
  research_focus = NULL,
  paper_metadata = NULL,
  description = "",
  filename = NULL,
  save_directory = 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,
  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,
  add_other = "prompt",
  check_verbosity = TRUE,
  prompt_tune = NULL,
  tune_iterations = 1L,
  tune_ui = "browser",
  tune_optimize = "balanced"
)
```

Arguments

categories	A character vector of category names, or "auto".
input_data	A character vector, list, or data.frame column, or NULL to fetch from academic sources. Default NULL.
api_key	Character or NULL. API key for the LLM provider.
journal_issn	Character or NULL. Journal ISSN to fetch papers from.
journal_name	Character or NULL. Journal name to fetch papers from.
journal_field	Character or NULL. Academic field to filter by.
topic_name	Character or NULL. Topic name to search for.
topic_id	Character or NULL. OpenAlex topic ID.
paper_limit	Integer. Max papers to fetch. Default 50L.
date_from	Character or NULL. Start date (YYYY-MM-DD).
date_to	Character or NULL. End date (YYYY-MM-DD).
polite_email	Character or NULL. Email for polite API pool.

journal Character or NULL. Alias for journal_name.

field Character or NULL. Alias for journal_field.

research_focus Character or NULL. Research focus filter.

paper_metadata Named list or NULL. Additional paper metadata.

description Character. Context description. Default "".

filename Character or NULL. Output CSV filename.

save_directory Character or NULL. Output directory.

user_model Character. Model name. Default "gpt-4o".

mode Character. 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. Default 0L.

example1, example2, example3, example4, example5, example6
Optional few-shot examples.

model_source Character. Provider hint. Default "auto".

max_categories Integer. 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.
add_other	Logical or "prompt". Default "prompt".
check_verbosity	Logical. Default TRUE.
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:
# Classify abstracts directly
results <- classify(
  categories = c("Methods", "Theory", "Review", "Other"),
  input_data = df$abstract,
  mode       = "text",
  api_key    = Sys.getenv("OPENAI_API_KEY"),
  user_model = "gpt-4o-mini"
)

# Fetch papers from a journal via OpenAlex
results <- classify(
  categories   = c("Empirical", "Theoretical", "Review"),
  journal_name = "American Sociological Review",
  paper_limit  = 100L,
  polite_email = "you@university.edu",
  api_key      = Sys.getenv("OPENAI_API_KEY")
)

## End(Not run)
```

---

explore *Explore raw categories in academic paper data*

Description

Wraps the Python catademic.explore() function. Returns every category string extracted from every chunk across every iteration – with duplicates intact.

Usage

```
explore(
  input_data = NULL,
  api_key = NULL,
  description = "",
  journal_issn = NULL,
  journal_name = NULL,
  journal_field = NULL,
  topic_name = NULL,
  topic_id = NULL,
  paper_limit = 50L,
  date_from = NULL,
  date_to = NULL,
  polite_email = NULL,
  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 NULL to fetch from academic sources. Default NULL.
api_key	Character or NULL. API key for the LLM provider.
description	Character. Context description. Default "".
journal_issn	Character or NULL. Journal ISSN.
journal_name	Character or NULL. Journal name.
journal_field	Character or NULL. Academic field.
topic_name	Character or NULL. Topic name.
topic_id	Character or NULL. OpenAlex topic ID.
paper_limit	Integer. Max papers to fetch. Default 50L.
date_from	Character or NULL. Start date (YYYY-MM-DD).
date_to	Character or NULL. End date (YYYY-MM-DD).
polite_email	Character or NULL. Email for polite API pool.
max_categories	Integer. Default 12L.

`categories_per_chunk`
: Integer. Default `10L`.

`divisions`
: Integer. Default `12L`.

`user_model`
: Character. Default `"gpt-4o"`.

`creativity`
: Numeric or `NULL`. Default `NULL`.

`specificity`
: Character. Default `"broad"`.

`research_question`
: Character or `NULL`.

`filename`
: Character or `NULL`.

`model_source`
: Character. Default `"auto"`.

`iterations`
: Integer. Default `8L`.

`random_state`
: Integer or `NULL`.

`focus`
: Character or `NULL`.

`chunk_delay`
: Numeric. Default `0.0`.

Value

A character vector of every category string extracted.

Examples

```
## Not run:
raw_cats <- explore(
  input_data = df$abstracts,
  api_key    = Sys.getenv("OPENAI_API_KEY"),
  user_model = "gpt-4o-mini",
  iterations = 4L
)
table(raw_cats)

## End(Not run)
```

---

extract *Extract categories from academic papers using LLMs*

Description

Wraps the Python `catademic.extract()` function. Discovers and returns a normalised, deduplicated set of categories from academic paper data.

Usage

```
extract(
  input_data = NULL,
  api_key = NULL,
  journal_issn = NULL,
  journal_name = NULL,
  journal_field = NULL,
  topic_name = NULL,
  topic_id = NULL,
  paper_limit = 50L,
  date_from = NULL,
  date_to = NULL,
  polite_email = NULL,
  journal = NULL,
  field = NULL,
  research_focus = NULL,
  paper_metadata = NULL,
  description = "",
  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 NULL to fetch from academic sources. Default NULL.
api_key	Character or NULL. API key for the LLM provider.
journal_issn	Character or NULL. Journal ISSN.
journal_name	Character or NULL. Journal name.
journal_field	Character or NULL. Academic field.
topic_name	Character or NULL. Topic name.
topic_id	Character or NULL. OpenAlex topic ID.
paper_limit	Integer. Max papers to fetch. Default 50L.
date_from	Character or NULL. Start date (YYYY-MM-DD).
date_to	Character or NULL. End date (YYYY-MM-DD).

polite_email	Character or NULL. Email for polite API pool.
journal	Character or NULL. Alias for journal_name.
field	Character or NULL. Alias for journal_field.
research_focus	Character or NULL. Research focus filter.
paper_metadata	Named list or NULL. Additional paper metadata.
description	Character. Context description. Default "".
max_categories	Integer. Default 12L.
categories_per_chunk	Integer. Default 10L.
divisions	Integer. Default 12L.
user_model	Character. Default "gpt-4o".
creativity	Numeric or NULL. Default NULL.
specificity	Character. Default "broad".
research_question	Character or NULL.
mode	Character. Default "text".
filename	Character or NULL.
model_source	Character. Default "auto".
iterations	Integer. Default 8L.
random_state	Integer or NULL.
focus	Character or NULL.
chunk_delay	Numeric. Default 0.0.

Value

A named list with counts_df, top_categories, and raw_top_text.

Examples

```
## Not run:
result <- extract(
  topic_name   = "climate change adaptation",
  paper_limit  = 200L,
  polite_email = "you@university.edu",
  api_key      = Sys.getenv("OPENAI_API_KEY"),
  user_model   = "gpt-4o-mini"
)
print(result$top_categories)

## End(Not run)
```

summarize *Summarize academic papers using LLMs*

Description

Wraps the Python `catademic.summarize()` function. Generates summaries of academic paper data. The Python function accepts `input_data` and passes all other arguments through via `**kwargs` to `cat_stack.summarize()`.

Usage

```
summarize(
  input_data,
  api_key = NULL,
  description = "",
  instructions = "",
  format = "paragraph",
  max_length = NULL,
  focus = NULL,
  user_model = "gpt-4o",
  model_source = "auto",
  mode = "image",
  input_mode = NULL,
  input_type = "auto",
  pdf_dpi = 150L,
  creativity = NULL,
  thinking_budget = 0L,
  chain_of_thought = TRUE,
  context_prompt = FALSE,
  step_back_prompt = FALSE,
  filename = NULL,
  save_directory = NULL,
  models = NULL,
  max_workers = NULL,
  parallel = NULL,
  auto_download = FALSE,
  safety = FALSE,
  max_retries = 5L,
  batch_retries = 2L,
  retry_delay = 1,
  row_delay = 0,
  fail_strategy = "partial",
  batch_mode = FALSE,
  batch_poll_interval = 30,
  batch_timeout = 86400
)
```

Arguments

input_data	A character vector, list, or data.frame column of paper abstracts or text.
api_key	Character or NULL. API key for the model provider.
description	Character. Context description. Default "".
instructions	Character. Specific instructions for the summary. Default "".
format	Character. Output format. Default "paragraph".
max_length	Integer or NULL. Max summary length. Default NULL.
focus	Character or NULL. Optional focus. Default NULL.
user_model	Character. Model name. Default "gpt-4o".
model_source	Character. Provider hint. Default "auto".
mode	Character. Processing mode. Default "image".
input_mode	Character or NULL. Explicit input mode. Default NULL.
input_type	Character. Input type. Default "auto".
pdf_dpi	Integer. DPI for PDFs. Default 150L.
creativity	Numeric or NULL. Temperature. Default NULL.
thinking_budget	Integer. Default 0L.
chain_of_thought	Logical. Default TRUE.
context_prompt	Logical. Default FALSE.
step_back_prompt	Logical. Default FALSE.
filename	Character or NULL. Output filename.
save_directory	Character or NULL. Output directory.
models	List of model specs for ensemble mode. Default NULL.
max_workers	Integer or NULL. Default NULL.
parallel	Logical or NULL. Default NULL.
auto_download	Logical. Default FALSE.
safety	Logical. Default FALSE.
max_retries	Integer. Default 5L.
batch_retries	Integer. Default 2L.
retry_delay	Numeric. Default 1.0.
row_delay	Numeric. Default 0.0.
fail_strategy	Character. Default "partial".
batch_mode	Logical. Default FALSE.
batch_poll_interval	Numeric. Default 30.0.
batch_timeout	Numeric. Default 86400.0.

Value

A `data.frame` with summarization results.

Examples

```
## Not run:
summaries <- summarize(
  input_data   = df$abstracts,
  description  = "Sociology journal abstracts",
  instructions = "Summarize the key findings in 2 sentences",
  format       = "paragraph",
  api_key      = Sys.getenv("OPENAI_API_KEY"),
  user_model   = "gpt-4o-mini"
)

## End(Not run)
```

Index