

# LLM-Extraction Project

## Data Preparation

Johannes Schwenke

2025-11-05

### Table of contents

1 Setup .....	1
2 Download Data from REDCap .....	1
3 Format Data for Analysis .....	2

### 1 Setup

```
library(tidyverse)
library(marginaleffects)
library(REDCapR)
library(lme4)
library(brms)
library(tinytable)
library(irr)
library(glue)
library(scales)
library(gt)
library(patchwork)
library(writexl)
library(here)
library(arrows)
```

### 2 Download Data from REDCap

```
redownload <- TRUE

if(redownload){
  data <- redcap_read(
    redcap_uri = Sys.getenv("redcap_fxdb_url"),
    token = Sys.getenv("llm_radio_api"),
    raw_or_label = "label"
  )$data

  # saveRDS(data, "./data/oncology/redcap_export_complete.rds")
} else {
  data <- readRDS("./data/oncology/redcap-2025-10-09-labels.rds")$data
}

logs <- redcap_log_read(
  redcap_uri = Sys.getenv("redcap_fxdb_url"),
  token = Sys.getenv("llm_radio_api"),
```

```
log_begin_date = as.Date("2025-05-01"),
log_end_date = Sys.Date(),
record = NULL,
user = NULL,
http_response_encoding = "UTF-8",
locale = readr::default_locale(),
verbose = TRUE,
config_options = NULL,
handle_httr = NULL
)
```

### 3 Format Data for Analysis

```
write_parquet(ground_truth_long, sink = here( "data", "oncology",
"ground_truth_long.parquet"))
write_parquet(analysis_data, sink = here("data","oncology","analysis_data.parquet"))
```