```
# ----------------------------
# 1. CORE IMPORTS
# ----------------------------
import pandas as pd
import numpy as np
import duckdb
import gdown
from tqdm import tqdm

# HuggingFace translation
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM, pipeline

# Display settings
pd.set_option("display.max_colwidth", None)

print("Setup complete.")
```

```
Setup complete.
```

```
import gdown

file_id = "1RxeikLQHjYEJCewMtxpbNVlQ1FcQcC9s"
url = f"https://drive.google.com/uc?id={file_id}"

output = "/content/farmer_questions.csv"

gdown.download(url, output, quiet=False)

print("Download complete:", output)
```

```
Downloading...
From (original): https://drive.google.com/uc?id=1RxeikLQHjYEJCewMtxpbNVlQ1FcQcC9s
From (redirected): https://drive.google.com/uc?id=1RxeikLQHjYEJCewMtxpbNVlQ1FcQcC9s&confirm=t&uuid=311a30
To: /content/farmer_questions.csv
100%|██████████| 7.25G/7.25G [01:53<00:00, 63.9MB/s]Download complete: /content/farmer_questions.csv
```

```
import duckdb

con = duckdb.connect("/content/producers_direct.duckdb")

con.execute("""
    CREATE OR REPLACE VIEW raw_questions AS
    SELECT * FROM read_csv_auto('/content/farmer_questions.csv');
""")

print("View ready.")
```

```
View ready.
```

```
con.execute("SHOW TABLES").df()
```

	name
0	raw_questions

```
con.execute("PRAGMA table_info('raw_questions')").df()['name'].tolist()
```

```
['question_id',
 'question_user_id',
 'question_language',
 'question_content',
 'question_topic',
 'question_sent',
 'response_id',
 'response_user_id',
 'response_language',
 'response_content',
 'response_topic',
 'response_sent',
 'question_user_type',
 'question_user_status',
 'question_user_country_code',
 'question_user_gender',
 'question_user_dob',
 'question_user_created_at',
 'response_user_type',
 'response_user_status',
 'response_user_country_code',
 'response_user_gender',
 'response_user_dob',
 'response_user_created_at']
```

```
con.execute("SELECT * FROM raw_questions LIMIT 5").df()
```

	question_id	question_user_id	question_language	question_content	question_topic	question_sent	res
0	3849056	519124	nyn	E ABA WEFARM OFFICES ZABO NIZISHANGWA NKAHI?	None	2017-11-22 12:25:03+00:00	
1	3849061	521327	eng	Q this goes to wefarm. is it possible to get for us market for our product. thax	None	2017-11-22 12:25:05+00:00	
2	3849077	307821	nyn	E ENTE YANJE EZAIRE ENYENA YASHOBERA. \nOBWIRE BWOKUZARA BUBAIRE BWAHIKIRE EZAIRE AKANYENA KAKYE KAMARAHO ENDAKIKA ITANO KAFA .KANDI NKAZINJE .OBWO NIBURWIREKI??	cattle	2017-11-22 12:25:08+00:00	
3	3849077	307821	nyn	E ENTE YANJE EZAIRE ENYENA YASHOBERA. \nOBWIRE BWOKUZARA BUBAIRE BWAHIKIRE EZAIRE AKANYENA KAKYE KAMARAHO ENDAKIKA ITANO KAFA .KANDI NKAZINJE .OBWO NIBURWIREKI??	cattle	2017-11-22 12:25:08+00:00	
4	3849077	307821	nyn	E ENTE YANJE EZAIRE ENYENA YASHOBERA. \nOBWIRE BWOKUZARA BUBAIRE BWAHIKIRE EZAIRE AKANYENA KAKYE KAMARAHO ENDAKIKA ITANO KAFA .KANDI NKAZINJE .OBWO NIBURWIREKI??	cat	2017-11-22 12:25:08+00:00	

5 rows × 24 columns

```
con.sql("""
    SELECT
        COUNT(*) AS total_rows,

        -- core fields
        SUM(question_id IS NULL) AS missing_question_id,
```

```
        SUM(question_content IS NULL) AS missing_question_content,
        SUM(question_language IS NULL) AS missing_question_language,
        SUM(question_topic IS NULL) AS missing_question_topic,
        SUM(question_user_id IS NULL) AS missing_question_user_id,

        -- challenge 4 fields
        SUM(question_user_country_code IS NULL) AS missing_country_code,
        SUM(question_sent IS NULL) AS missing_question_sent

    FROM raw_questions;
""").df()
```

	total_rows	missing_question_id	missing_question_content	missing_question_language	missing_question
0	20304843	0.0	0.0	0.0	353

```
con.sql("""
    CREATE OR REPLACE VIEW question_level AS
    SELECT DISTINCT
        question_id,
        question_content,
        question_language,
        question_topic,
        question_user_country_code,
        question_sent
    FROM raw_questions;
""")
```

```
con.sql("""
    SELECT COUNT(*) AS n_question_level_rows
    FROM question_level;
""").df()
```

	n_question_level_rows
0	6627409

```
con.sql("""
    SELECT
        COUNT(*) AS total_questions,
        SUM(question_topic IS NULL) AS missing_topics
    FROM question_level;
""").df()
```

	total_questions	missing_topics
0	6627409	1672009.0

```
con.sql("""
    SELECT question_topic, COUNT(*) AS n
    FROM question_level
    GROUP BY question_topic
    ORDER BY n DESC;
```

```
""").df()
```

	question_topic	n
0	None	1672009
1	maize	595779
2	chicken	493791
3	cattle	462713
4	tomato	353851
...	...	...
144	blackberry	27
145	setaria	25
146	mulberry	22
147	purple-vetch	12
148	cranberry	4

149 rows × 2 columns

```
con.sql("""
    SELECT DISTINCT question_topic
    FROM question_level
    ORDER BY 1;
""").df()
```

	question_topic
0	acacia
1	african-nightshade
2	amaranth
3	animal
4	apple
...	...
144	vetch
145	watermelon
146	wheat
147	yam
148	None

149 rows × 1 columns

```
con.sql("""
    SELECT DISTINCT question_language
    FROM question_level;
""").df()
```

	question_language
0	nyn
1	eng
2	swa
3	lug

```
con.sql("""
    SELECT question_id, COUNT(*)
    FROM question_level
    GROUP BY question_id
    HAVING COUNT(*) > 1
    ORDER BY COUNT(*) DESC
    LIMIT 20;
""").df()
```

	question_id	count_star()
0	33636032	15
1	44092721	15
2	8811829	14
3	36617024	13
4	36717802	13
5	34835109	13
6	24015389	12
7	31138296	12
8	42214400	12
9	43450645	12
10	5233828	11
11	34764414	11
12	26529316	11
13	20433220	11
14	19636378	11
15	23316811	11
16	19636223	11
17	19635914	11
18	27272879	11
19	22928035	11

```
con.sql("""
    SELECT SUM(occurrences - 1) AS total_duplicate_question_rows
    FROM (
        SELECT question_id, COUNT(*) AS occurrences
```

```
        FROM question_level
        GROUP BY question_id
        HAVING COUNT(*) > 1
    );
""").df()
```

	total_duplicate_question_rows
0	761590.0

```
con.sql("""
    CREATE OR REPLACE VIEW question_level AS
    SELECT DISTINCT
        question_id,
        question_content,
        question_language,
        question_topic,
        question_user_country_code,
        question_sent
    FROM raw_questions;
""")
```

```
con.sql("""
    SELECT COUNT(*) AS n_question_level_rows
    FROM question_level;
""").df()
```

	n_question_level_rows
0	6627409

```
con.sql("""
    SELECT
        COUNT(*) AS total_rows,
        SUM(question_content IS NULL) AS missing_question_content,
        SUM(question_language IS NULL) AS missing_question_language,
        SUM(question_topic IS NULL) AS missing_question_topic,
        SUM(question_user_country_code IS NULL) AS missing_country_code
    FROM question_level;
""").df()
```

	total_rows	missing_question_content	missing_question_language	missing_question_topic	missing_count
0	6627409	0.0	0.0	1672009.0	

table	column	issue	row_count	magnitude	solvable?	
question_level	question_topic	missing topic labels	1,672,009	25.23%	N	leave as is – no way to
question_level	question_content	no missing values	0	0.00%	N/A	no action needed
question_level	question_language	no missing values	0	0.00%	N/A	no action needed
question_level	question_user_country_code	no missing values	0	0.00%	N/A	no action needed
raw_questions	question_id	duplicates collapse into fewer rows	14,439,024 dupes	71.1%	Y	resolved by creating qu

N — Note & Document

In this final step, I documented every decision made during the data cleaning process. I completed the issues log with row counts and percentages, explained how each issue was handled, and recorded which issues could not be fixed. I also kept a clear paper trail by noting the code used to remove duplicates, the columns kept for analysis, and the reasoning behind leaving missing topics unchanged. This documentation provides transparency, shows how the dataset was transformed, and ensures that anyone reviewing the project can follow the cleaning process from beginning to end.

Below is the final issues log summarizing all identified problems, their magnitude, and the actions taken:

```
ISSUES LOG

table           column                        issue                                  row_count     magnitu
--------------- ----------------------------- -------------------------------------- ------------- ------
question_level  question_topic                 missing topic labels                     1,672,009      25.2
question_level  question_content               no missing values                        0              0.00%
question_level  question_language              no missing values                        0              0.00%
question_level  question_user_country_code     no missing values                        0              0.00%
raw_questions   question_id                    duplicate question rows in raw data      14,439,024     ~71%
```

These notes complete the data cleaning documentation. All cleaning steps and decisions were recorded to maintain clarity, support reproducibility, and provide a transparent audit trail for downstream translation and topic analysis.

```
# Row count: should be ~6.6M unique questions
con.sql("""
    SELECT COUNT(*) AS n_question_level_rows
    FROM question_level;
""").df()

# Quick sanity check on columns and values
con.sql("""
    SELECT *
    FROM question_level
    LIMIT 5;
""").df()
```

	question_id	question_content	question_language	question_topic	question_user_country_code	question
0	3849077	E ENTE YANJE EZAIRE ENYENA YASHOBERA. \nOBWIRE BWOKUZARA BUBAIRE BWAHIKIRE EZAIRE AKANYENA	nyn	cattle	ug	2017 12:25:08+

```
sample = con.sql("""
    SELECT
        question_id,
        question_language,
```

```
        question_content
    FROM question_level
    WHERE question_language != 'eng'
    LIMIT 5;
""").df()

sample
```

	question_id	question_language	question_content
0	3937171	swa	sungura wangu wako na upele niwape ndawa gani?
1	3937182	swa	S;nmefuka samaki na nmekoxa soko.Where wll I get?.
2	3937212	nyn	E# nimbuza ngu ebihimba hati biri arizingahi?.
3	3937230	swa	s napaswa kula nyama ya ngombe ambaye alikufa usiku.?
4	3937270	swa	S Niko Na Maragwe Gunia Tanu Na Ta Futa Shoko Kama Unataka Napati Kana

```
from transformers import pipeline

translator = pipeline(
    "translation",
    model="facebook/nllb-200-distilled-600M",
    device_map="auto"
)

def translate_row(text, src_lang_code):
    return translator(
        text,
        src_lang=src_lang_code,     # e.g. "nyn_Latn", "swh_Latn"
        tgt_lang="eng_Latn",
        max_length=400
    )[0]["translation_text"]

sample["translated_en"] = sample.apply(
    lambda r: translate_row(r["question_content"], r["question_language"]),
    axis=1
)

sample
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/se
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
```

config.json: 100% 846/846 [00:00<00:00, 45.4kB/s]

pytorch_model.bin: 100% 2.46G/2.46G [01:28<00:00, 73.0MB/s]

model.safetensors: 100% 2.46G/2.46G [01:06<00:00, 69.5MB/s]

generation_config.json: 100% 189/189 [00:00<00:00, 11.4kB/s]

tokenizer_config.json: 100% 564/564 [00:00<00:00, 57.6kB/s]

sentencepiece.bpe.model: 100% 4.85M/4.85M [00:01<00:00, 4.19MB/s]

tokenizer.json: 100% 17.3M/17.3M [00:00<00:00, 21.8MB/s]

special_tokens_map.json: 3.55k/? [00:00<00:00, 232kB/s]

```
Device set to use cuda:0
```

	question_id	question_language	question_content	translated_en
0	3937171	swa	sungura wangu wako na upele niwape ndawa gani?	What medication should I give my baby?
1	3937182	swa	S;nmefuka samaki na nmekoxa soko.Where wll I get?.	S;nmefuka fish and nmekoxa soko.Where will I get?.
2	3937212	nyn	E# nimbuza ngu ebihimba hati biri arizingahi?.	E# nimbuza by ebihimba that are arizingahi?.

```
batch = con.sql("""
    SELECT *
    FROM question_level
    WHERE question_language != 'eng'
    ORDER BY question_id
    LIMIT 5000
    OFFSET 0
""").df()
```

```
def translate_row(text, src_lang_code):
    return translator(
        text,
        src_lang=src_lang_code,
        tgt_lang="eng_Latn",
        max_length=400,
    )[0]["translation_text"]
```

```
from tqdm.auto import tqdm
tqdm.pandas()  # lets us use progress_apply on DataFrames
```

```
batch["translated_en"] = batch.progress_apply(
    lambda r: translate_row(r["question_content"], r["question_language"]),
    axis=1
)
```

100% 5000/5000 [57:39<00:00, 1.72it/s]

```
batch[["question_id", "question_language", "question_content", "translated_en"]].head()
```

	question_id	question_language	question_content	translated_en
0	3849056	nyn	E ABA WEFARM OFFICES ZABO NIZISHANGWA NKAHI?	Where are the WEFARM offices used?
1	3849077	nyn	E ENTE YANJE EZAIRE ENYENA YASHOBERA. \nOBWIRE BWOKUZARA BUBAIRE BWAHIKIRE EZAIRE AKANYENA KAKYE KAMARAHO ENDAKIKA ITANO KAFA .KANDI NKAZINJE .OBWO NIBURWIREKI??	My wife, Ezaire, is a good-for-nothing woman, and I'm a good-for-nothing woman, but I'm a good-for-nothing woman, and I'm a good-for-nothing woman, and I'm a good-for-nothing woman, and I'm a good-for-nothing woman, and I'm a good-for-nothing woman, and I'm a good-for-nothing woman, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and I'm good-for-nothing, and

```
# Any empty translations?
batch[batch["translated_en"].isna() | (batch["translated_en"].str.strip() == "")]
```

	question_id	question_content	question_language	question_topic	question_user_country_code	quest
1555	3886291	s muembe uzaa	swa	mango	ke	2

```
# How many rows per language in this batch
batch["question_language"].value_counts()
```

question_language	count
swa	2877
nyn	1927
lug	196

dtype: int64

```
# language distribution
con.sql("""
    SELECT
        question_language,
        COUNT(*) AS n_rows
    FROM question_level
    GROUP BY question_language
    ORDER BY n_rows DESC;
""").df()

# total non-English rows
con.sql("""
    SELECT
        COUNT(*) AS non_english_rows
    FROM question_level
```

```
    WHERE question_language != 'eng';
""").df()
```

	non_english_rows
0	3242518

```
batch.to_csv("translated_batch_000.csv", index=False)

from google.colab import files
files.download("translated_batch_000.csv")
```

We have created a helper table of translated ids. This lets use skip anything we already translated.

```
import pandas as pd

df_batch = pd.read_csv("translated_batch_000.csv")
df_batch.head()
```

	question_id	question_content	question_language	question_topic	question_user_country_code	question
0	3849056	E ABA WEFARM OFFICES ZABO NIZISHANGWA NKAHI?	nyn	NaN	ug	2017 12:25:03+

Next steps: Generate code with df_batch | New interactive sheet

```
df_batch["src_len"] = df_batch["question_content"].str.len()
df_batch["tgt_len"] = df_batch["translated_en"].str.len()
df_batch["len_ratio"] = df_batch["tgt_len"] / df_batch["src_len"]

df_batch.sort_values("len_ratio", ascending=False).head(10)[[
    "question_id", "question_language", "src_len", "tgt_len", "len_ratio",
    "question_content", "translated_en"
]]
```

	question_id	question_language	src_len	tgt_len	len_ratio	question_cont
4462	3953872	swa	65	1590	24.461538	S nina kuku kienyeji ina miezi saba na hai mayai nifanye

```
bad = df_batch[df_batch["len_ratio"] > 4]
good = df_batch[df_batch["len_ratio"] <= 4]

print("Bad translations:", len(bad))
print("Good translations:", len(good))
```

```
Bad translations: 82
Good translations: 4918
```

```
good.sort_values("len_ratio", ascending=True)[
    ["question_id", "question_language", "len_ratio",
     "question_content", "translated_en"]
].head(20)
```

	question_id	question_language	len_ratio	question_content	translated_en
457	3859172	swa	0.378947	S ASANDE JOHN KWA JIBU LAKO. JE NDAMA ANAFAA APEWE MAZIWA KIASI NGANI KWA SIKI?.	And thank you John for your prayers.
458	3859172	swa	0.378947	S ASANDE JOHN KWA JIBU LAKO. JE NDAMA ANAFAA APEWE MAZIWA KIASI NGANI KWA SIKI?.	And thank you John for your prayers.
4021	3943907	swa	0.400000	NATAKA KUFUGA SUNGURA WA KIENYEJI TELL ME MORE ABOUT IT	Tell me more about it.
4019	3943907	swa	0.400000	NATAKA KUFUGA SUNGURA WA KIENYEJI TELL ME MORE ABOUT IT	Tell me more about it.
4020	3943907	swa	0.400000	NATAKA KUFUGA SUNGURA WA KIENYEJI TELL ME MORE ABOUT IT	Tell me more about it.
1447	3882623	lug	0.421053	E Mweziki Omutufu Ogwokusimbilamu Enyanya? Nekilala Waliwo Obubwa Ngabudugavu(budolodondo) Dagala Ki Eryokufuyira.	Is the music of the dead a symbol of friendship?
1626	3887295	swa	0.421053	S Je mtanisaije nipate Mbolea aina ya DAP NA CAN. MIMEA YANGU ILILIWA NA WANYAMA PORI NUSU EKA	I'm trying to find a DAP and CAN bottle.
1627	3887295	swa	0.421053	S Je mtanisaije nipate Mbolea aina ya DAP NA CAN. MIMEA YANGU ILILIWA NA WANYAMA PORI NUSU EKA	I'm trying to find a DAP and CAN bottle.
1464	3883095	lug	0.421053	E Mweziki Omutufu Ogwokusimbilamu Enyanya? Nekilala Waliwo Obubwa Ngabudugavu(budolodondo) Dagala Ki Eryokufuyira.	Is the music of the dead a symbol of friendship?

```
langs = ["swa", "lug", "nyn"]

sample = (
    good[good["question_language"].isin(langs)]
    .groupby("question_language", group_keys=False)
    .apply(lambda g: g.sample(10, random_state=42))
)

sample
```

```
/tmp/ipython-input-3030559252.py:6: DeprecationWarning: DataFrameGroupBy.apply operated on the grouping c
  .apply(lambda g: g.sample(10, random_state=42))
```

	question_id	question_content	question_language	question_topic	question_user_country_code	qu
935	3871651	E\nKawukaki Akaletera Ebikola Bya Kasooli Okufuka Ebyayero	lug	maize	ug	1
3802	3938460	E BWEMBA NE KILLO ZANGE LUKUMI (1000) EZAKASOLI NSOBOLA OKUFUNA OMUGUZI AMPA SHs 1000?	lug	NaN	ug	1
		E buyambi nze emma ekitoke kyange bwekilwala sikikulawawo wabula nkisimbako omumwanyi				

```
sample = sample.copy()
sample["quality"] = "unknown"   # default
```

okugwawo nga nayo

```
bad_ids = [
    3914673,  # "It's a crime against humanity."
    3905716,  # rabbit / pregnancy -> fine, you can keep or drop; I marked as ok below
    3877860,  # "Who needs a visa to attend school?"
    3887661,  # "best chicken feeders in the world."
    3929464,  # "What is the meaning of life?"
    3909714,  # "Equal Rights Amendment?"
]

sample.loc[sample["question_id"].isin(bad_ids), "quality"] = "bad"
```

piagara kurima kabegi

```
sample.loc[sample["quality"] == "unknown", "quality"] = "good"
```

E ENO E KAYUNGA

```
sample[["question_id","question_language","translated_en","quality"]]
```

554	3861031	E ddagala kyi eliita ebiwuka ebikossa kasooli	lug	maize	ug	0
355	3857476	E.ensawo ya green paper yameka emasaka.	lug	NaN	ug	1
2751	3909714	E Ecuzuöebihango nibirwaza ki?	nyn	NaN	ug	2
3877	3939714	E Ebinyobwa nibigura bita, INNOCENT RUBANDA.	nyn	NaN	ug	1
3497	3930780	E Ebafu negura sh 10000 zonka omuri kakumiro District.	nyn	NaN	ug	0
1649	3887661	E ebwa zagye zitirwe amahwa hati koreki ebitere byo kurya enkoko bidikubi	nyn	chicken	ug	1
440	3858911	E Ekyikoko Kyemituba	nyn	NaN	ug	

	question_id	question_language	translated_en	quality
935	3871651	lug	Kawukaki is responsible for the burning of fossil fuels.	good
3802	3938460	lug	What is the price of a thousand dollars a day or a thousand shillings a day?	good
2075	3896031	lug	The help I get when I'm sick is not enough, but it's enough to make me drink a glass of water!	good
3974	3942371	lug	Qlug MUTUBA IN RUNYANKOLE is located in RUNYANKOLE.	good
2975	3914673	lug	It's a crime against humanity.	bad
4120	3963555	lug	cattle E, Ekipwa Neklugiki.	good
536	3860554	lug	E sirina kisenyi but I have to live to grow kabegi kiklugaki kyenba using kubukalu ?	good
561	3861184	lug	That's the only way we're going to be able to survive!!.	good
554	3861031	lug	Egala kyi elita is known as ebikossa kasolugoli.	good
355	3857476	lug	E.lugensawo of green paper was tested in the mail.	good
2751	3909714	nyn	What do you think about the Equal Rights Amendment?	bad
3877	3939714	nyn	Ebinyobwa is called nibigura bita, INNOCENT RUBANDA.	good
3497	3930780	nyn	Ebafu negura sh 10000 is located in omuri kakumiro district.	good
1649	3887661	nyn	Their dogs are rumored to be the best chicken feeders in the world.	bad
440	3858911	nyn	E Ekyikoko Kyemituba Nikyiretwakyi is also known as Kyemituba Nikyiretwakyi.	good
1951	3894376	nyn	E EMbunzi Eyine moon Nombasakugiha We are Zigahi?	good
2054	3895758	nyn	If you have 250 pounds of weight, you can call 0700489699 ,0755788070 at any time.	good
3590	3933573	nyn	E#nyine,ente,eyimukize Orwoyangihemubaaziki, who is also the leader of the group.	good
4780	3964996	nyn	He hates the drunk who drinks the wine of wild beasts.	good
3450	3929464	nyn	What is the meaning of life? - Wellen Kasese	bad
2899	3912698	swa	I also don't want to know the price of the tomato seeds and the lifespan of harvesting it.	good
927	3871356	swa	What is the meaning of the term "craftsmanship"?	good
4491	3954646	swa	Where else can I buy coins, carrots and vegetables?	good
3296	3923425	swa	Q I want to cook chicken kienyeji WA Mayai in which space for 200 Chicken.	good
3782	3938018	swa	The seeds to be sown during the planting process are the same as the seeds to be sown during the planting process.	good

```
sample.to_csv("translation_quality_sample.csv", index=False)

from google.colab import files
files.download("translation_quality_sample.csv")
```

```
good_sample = sample[sample["quality"] == "good"].copy()
good_sample.head(2)
```

	question_id	question_content	question_language	question_topic	question_user_country_code	quest
935	3871651	E\nKawukaki Akaletera Ebikola Bya Kasooli Okufuka Ebyayero	lug	maize	ug	2 15:40
3802	3938460	E BWEMBA NE KILLO ZANGE LUKUMI (1000) EZAKASOLI NSOBOLA OKUFUNA OMUGUZI AMPA SHs 1000?	lug	NaN	ug	2 17:49

```
import re
from collections import Counter
import pandas as pd

def clean_text(text):
    if pd.isna(text):
        return ""
    text = text.lower()
    text = re.sub(r"[^a-z\s]", " ", text)
    return text

rows = []
for lang in ["swa", "lug", "nyn"]:
    texts = good_sample.loc[
        good_sample["question_language"] == lang, "translated_en"
    ].apply(clean_text)
    tokens = " ".join(texts).split()
    counts = Counter(tokens)
```