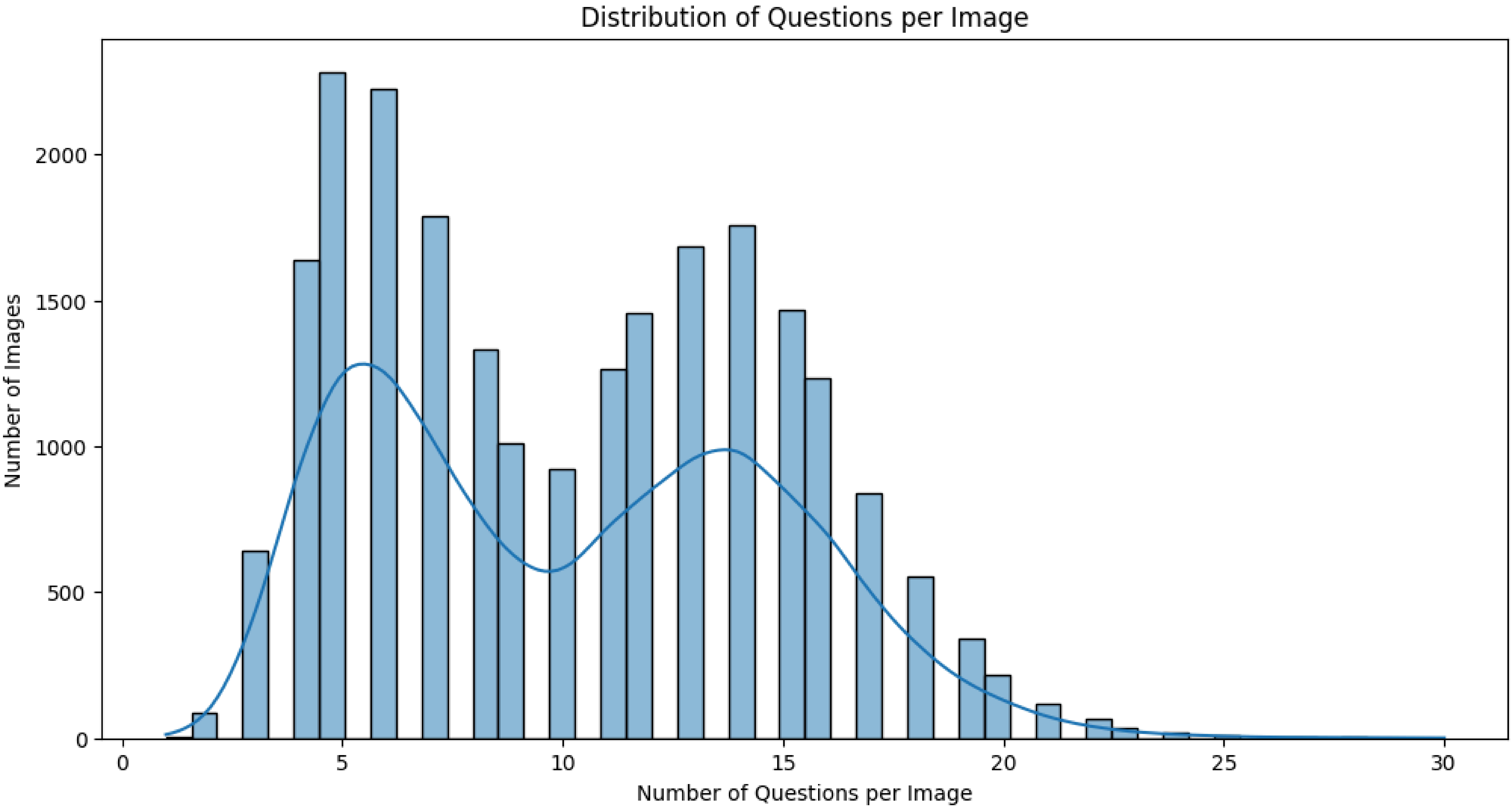
EDA

For VQA TextMining’s Project

Data info

Images: 23008
QA-pairs: 235421
Questions: 57300
Answers: 108


Distribution of Questions per Image
Number of Images
0
500
1000
1500
2000
0
5
10
15
20
25
30
Number of Questions per Image

Top 10 question

01

Top 10 Questions Leading to Yes/No Answers:
Are there any person in image? 12300
Are there any cup in image? 2152
Are there any bowl in image? 2080
Are there any vase in image? 1566
Are there any car in image? 1452
Are there any wine glass in image? 1418
Are there any dining table in image? 1371
Are there any dog in image? 1354
Are there any bird in image? 1338
Are there any bottle in image? 1331

02

Top 10 Questions Leading to Numeric Answers:
How many person in image? 12207
How many cup in image? 1950
How many bowl in image? 1879
How many vase in image? 1368
How many car in image? 1284
How many dog in image? 1182
How many wine glass in image? 1175
How many dining table in image? 1151
How many bottle in image? 1129
How many bird in image? 1114

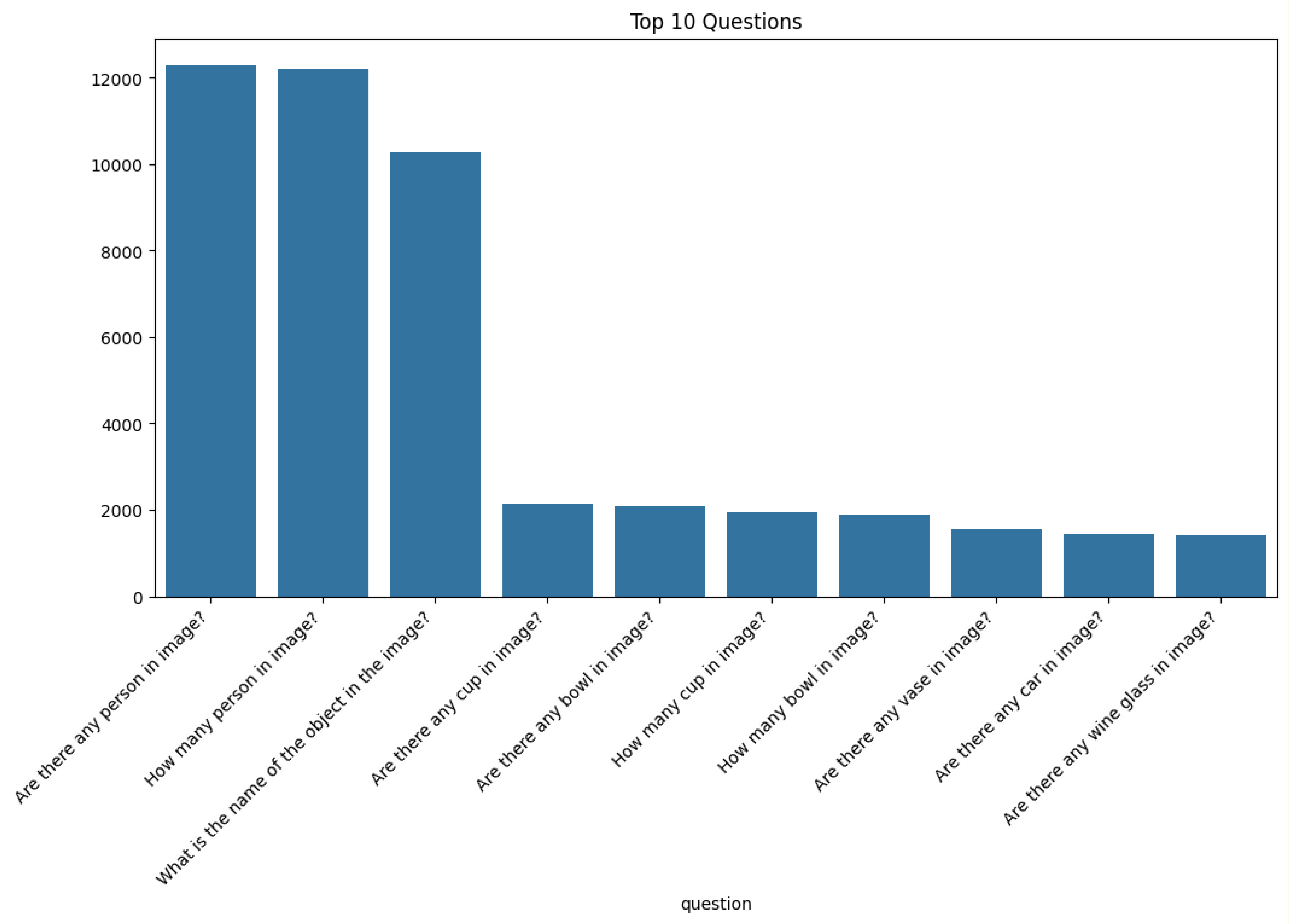
Top 10 question

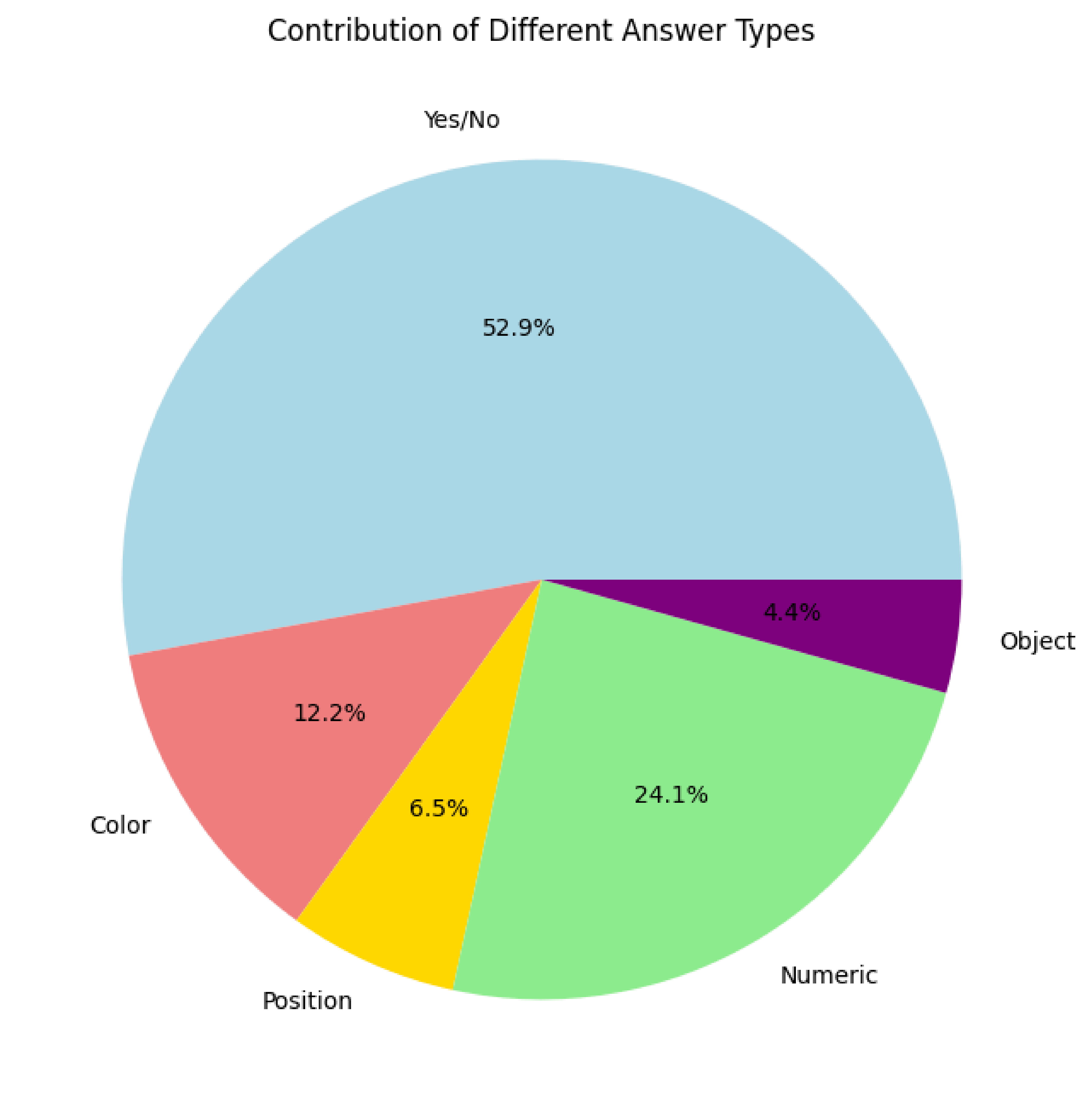
01

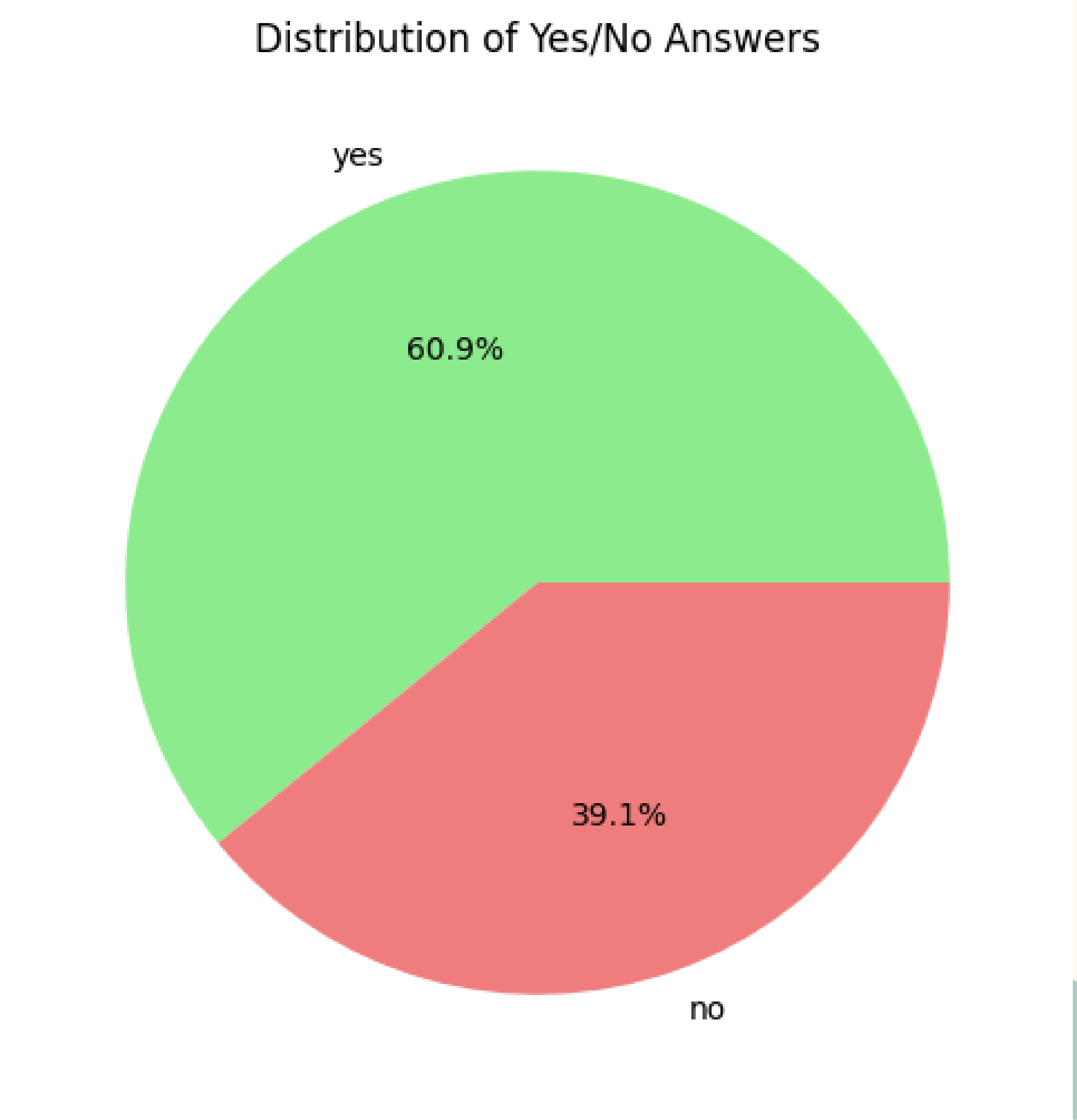
Top 10 Questions Leading to Color Answers:
What color is the sky? 562
What color is the background? 353
What color are the leaves? 263
What color are the flowers? 248
What color is the shirt? 228
What color is the grass? 217
What is the color of the sky? 205
What color is the jacket? 177
What color is the table? 167
What color is the water? 149

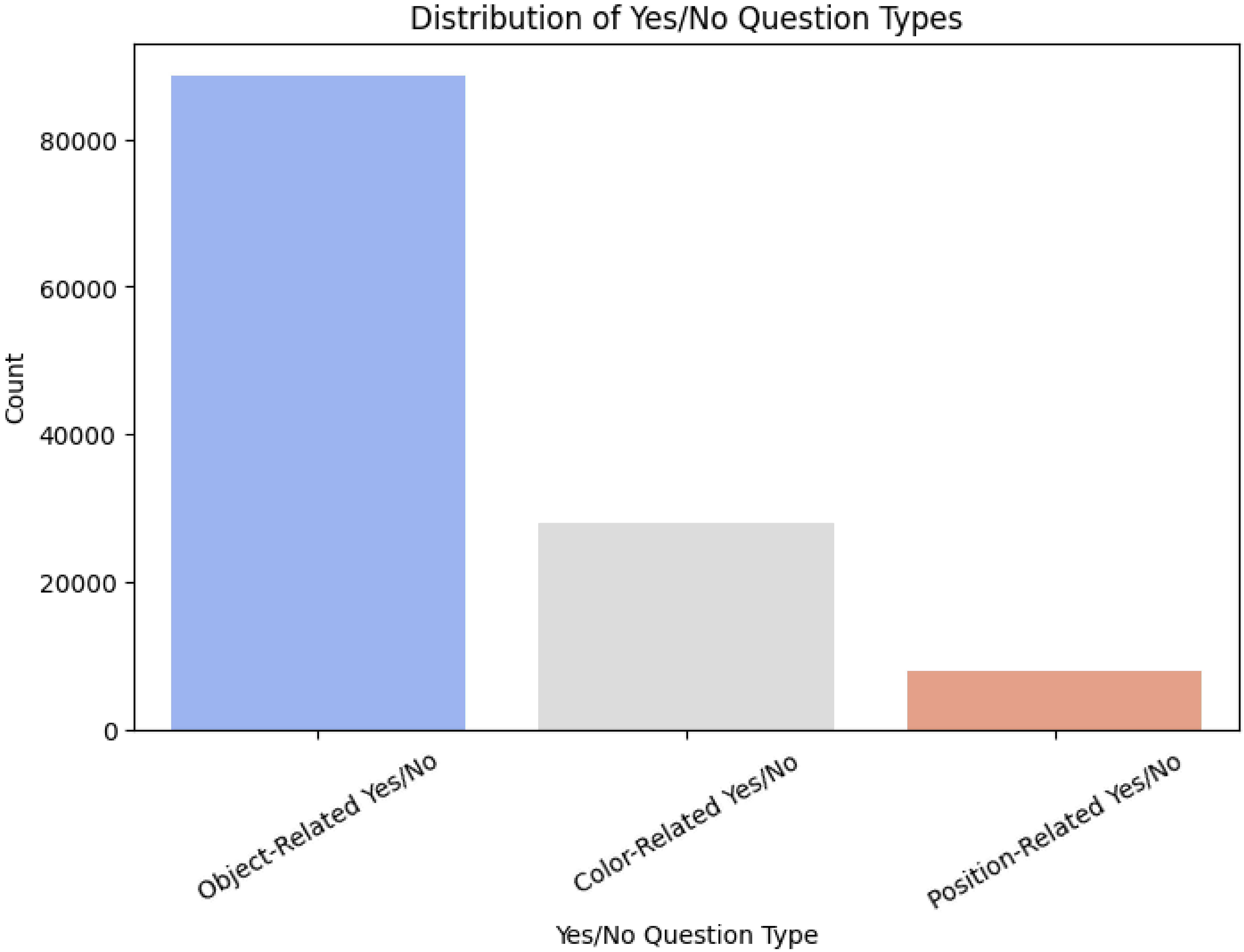
02

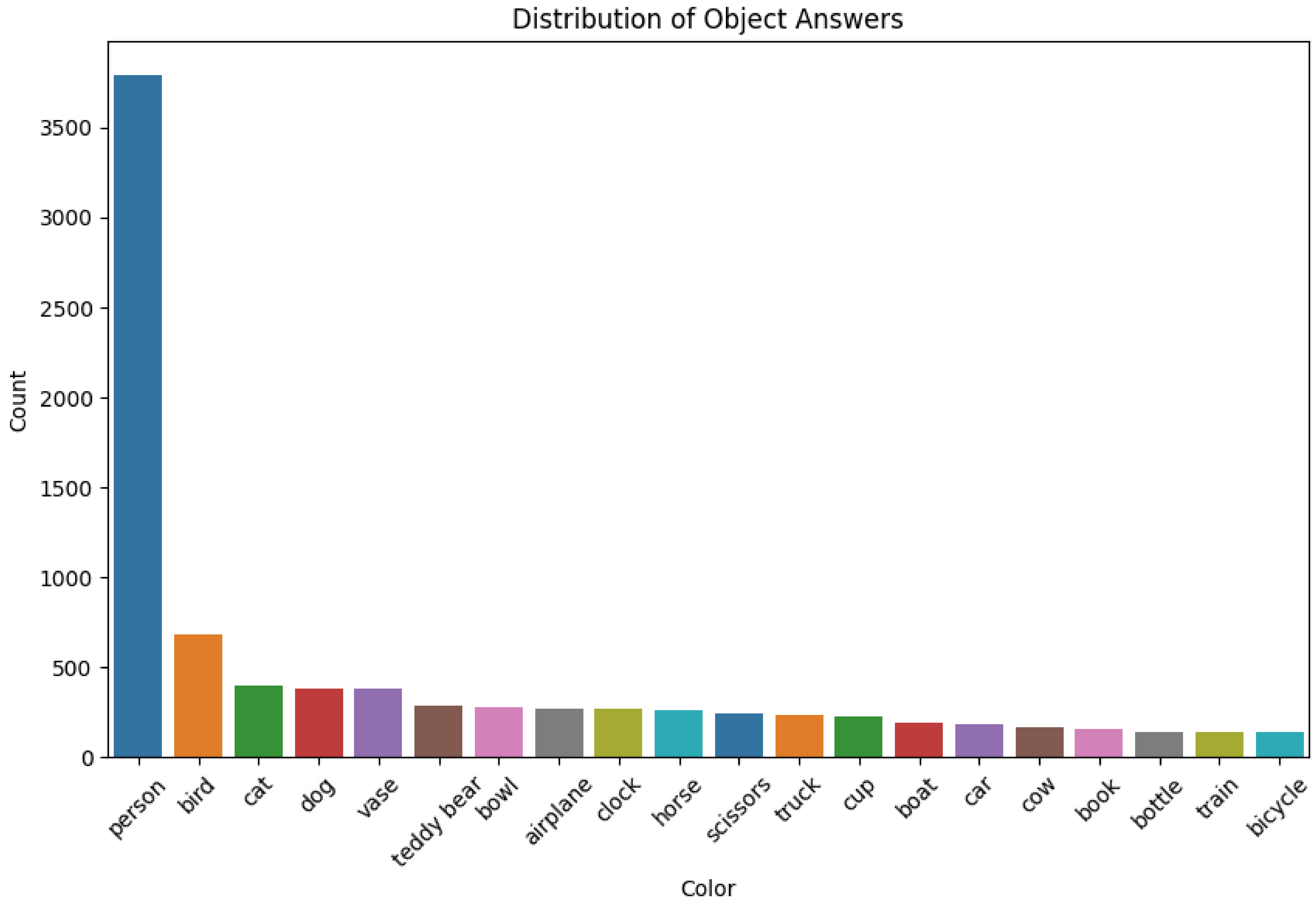
Top 10 Questions Leading to Position Answers:
Where is the vase? 73
Where is the teddy bear? 65
Where is the fork? 51
Where is the clock? 44
Where is the book? 35
are the flowers? 34
Where is the knife? 34
Where is the glass? 33
Where is the plant? 31
Where is the wine glass? 31

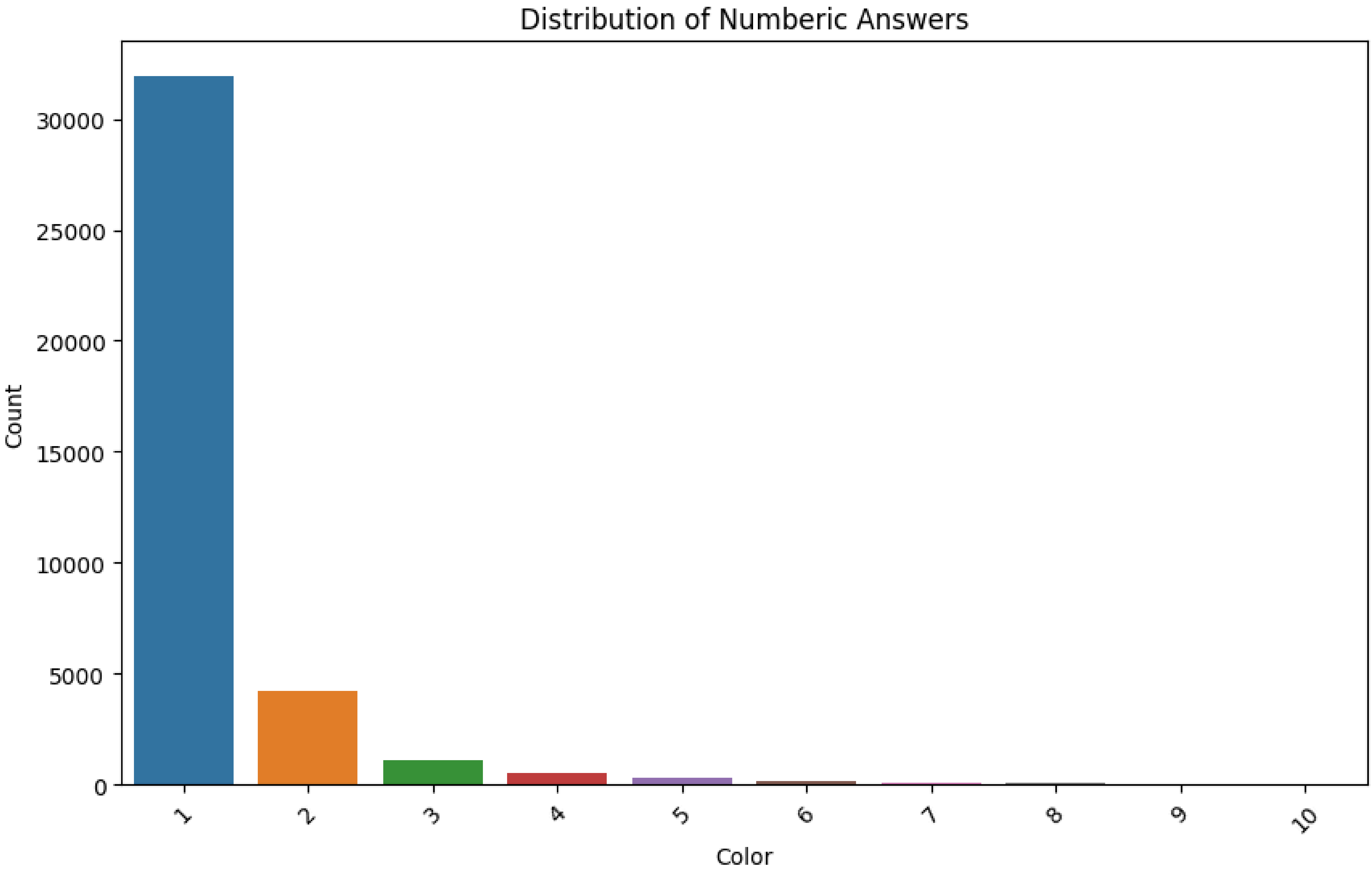

Top 10 Questions
12000
10000
8000
6000
4000
2000
0
Are there any person in image?
How many person in image?
What is the name of the object in the image?
Are there any cup in image?
Are there any bowl in image?
How many cup in image?
How many bowl in image?
Are there any vase in image?
Are there any car in image?
Are there any wine glass in image?
question

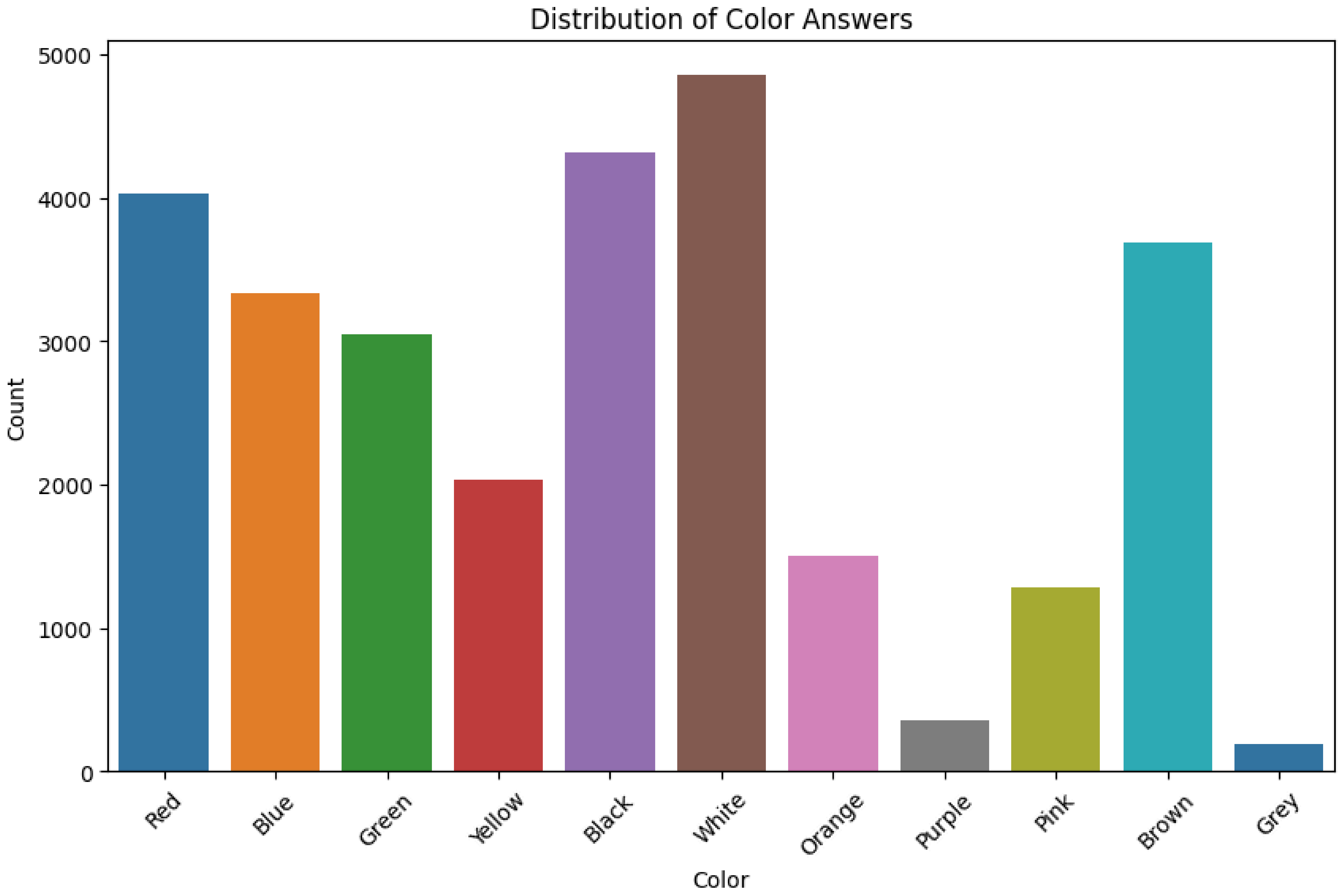

Contribution of Different Answer Types
Yes/No
52.9%
Color
12.2%
Position
6.5%
Numeric
24.1%
Object
4.4%

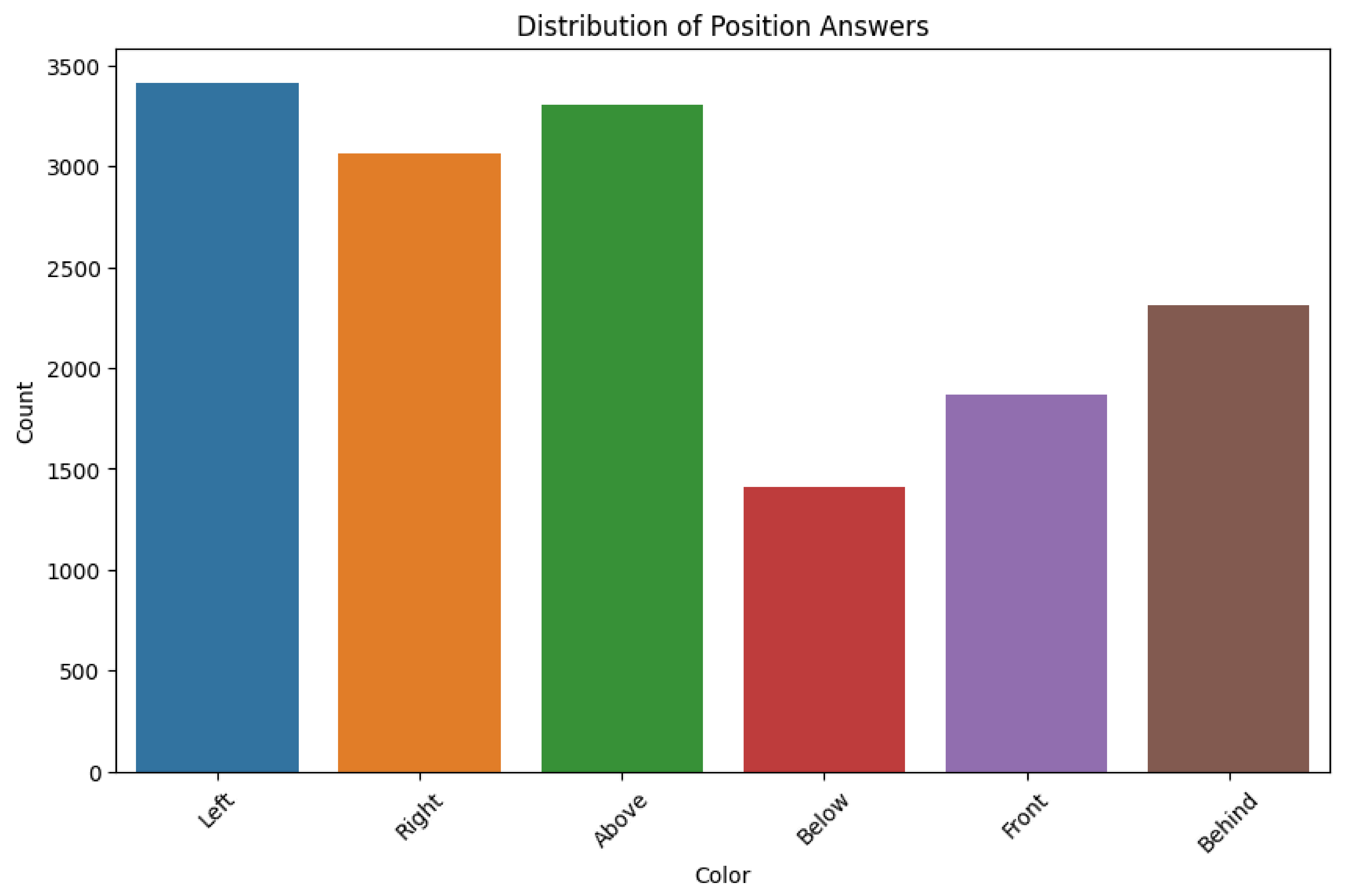

Distribution of Yes/No Answers

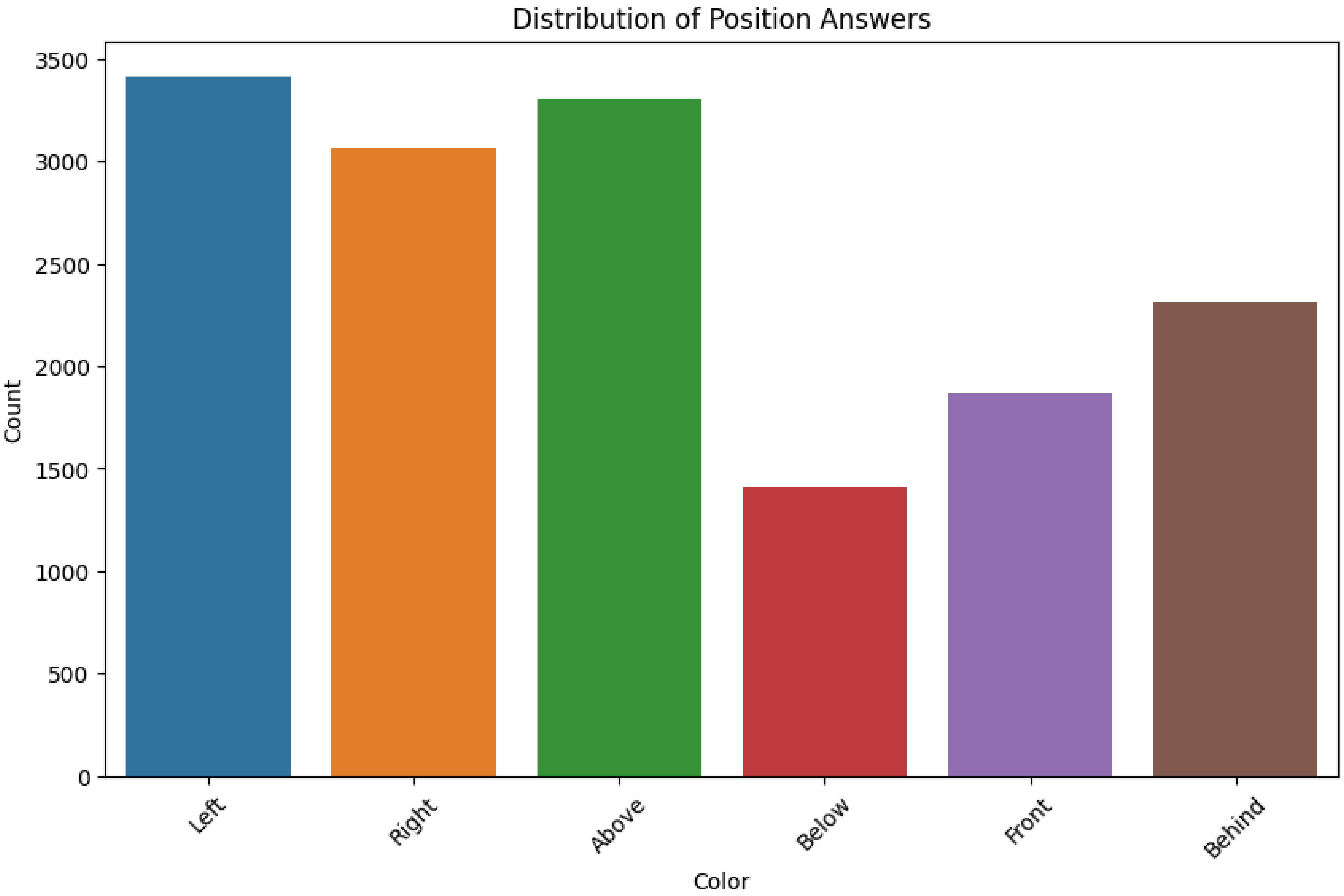

Distribution of Yes/No Question Types
Count
80000
60000
40000
20000
0
Object-Related Yes/No
Color-Related Yes/No
Position-Related Yes/No
Yes/No Question Type


Distribution of Object Answers
Count
0
500
1000
1500
2000
2500
3000
3500
person
bird
cat
dog
vase
teddy bear
bowl
airplane
clock
horse
scissors
truck
cup
boat
car
cow
book
bottle
train
bicycle
Color


Distribution of Numberic Answers
Count
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
Color


Distribution of Color Answers
Count
5000
4000
3000
2000
1000
0
Red
Blue
Green
Yellow
Black
White
Orange
Purple
Pink
Brown
Grey
Color


Distribution of Position Answers
3500
3000
2500
2000
1500
1000
500
0
Count
Left
Right
Above
Below
Front
Behind
Color


Distribution of Position Answers
Count
3500
3000
2500
2000
1500
1000
500
0
Left
Right
Above
Below
Front
Behind
Color

Test info

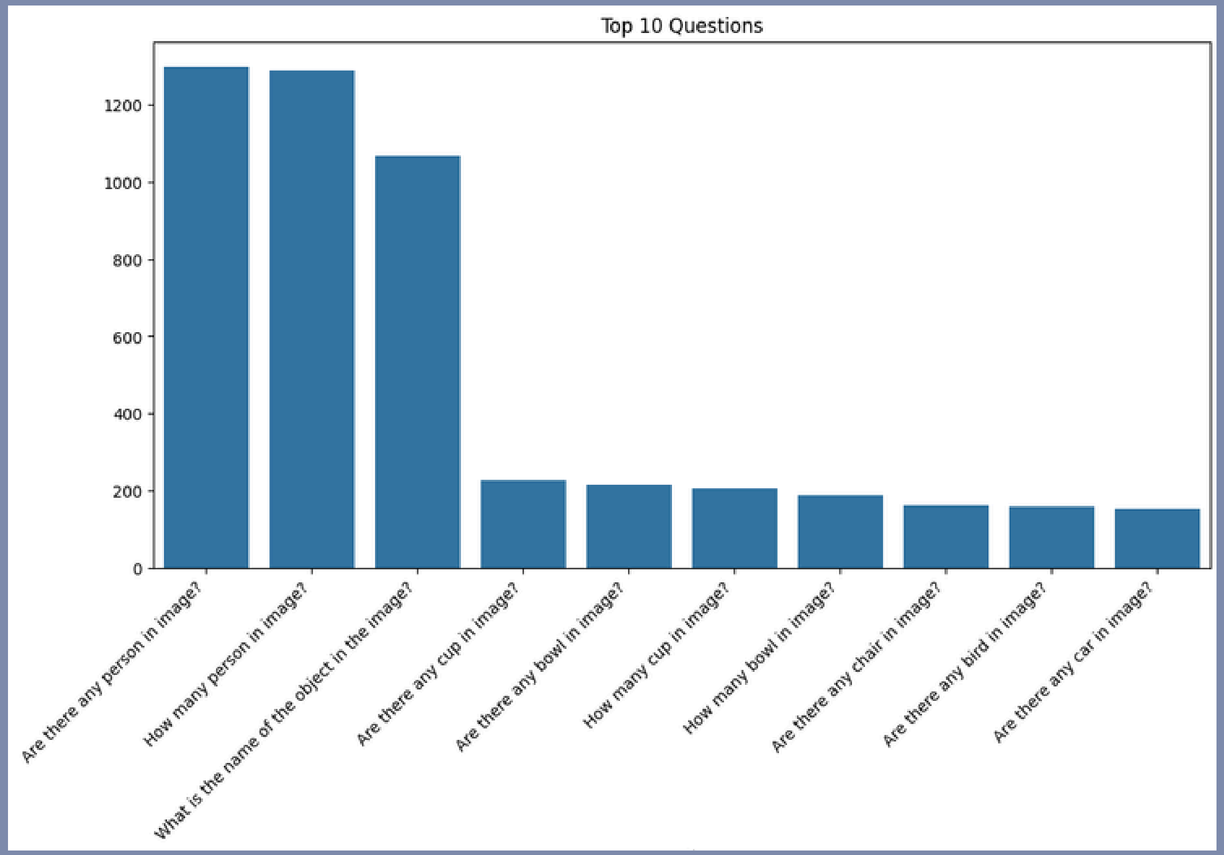
Images: 2417
Questions: 7961
Answers: 87

Eval info

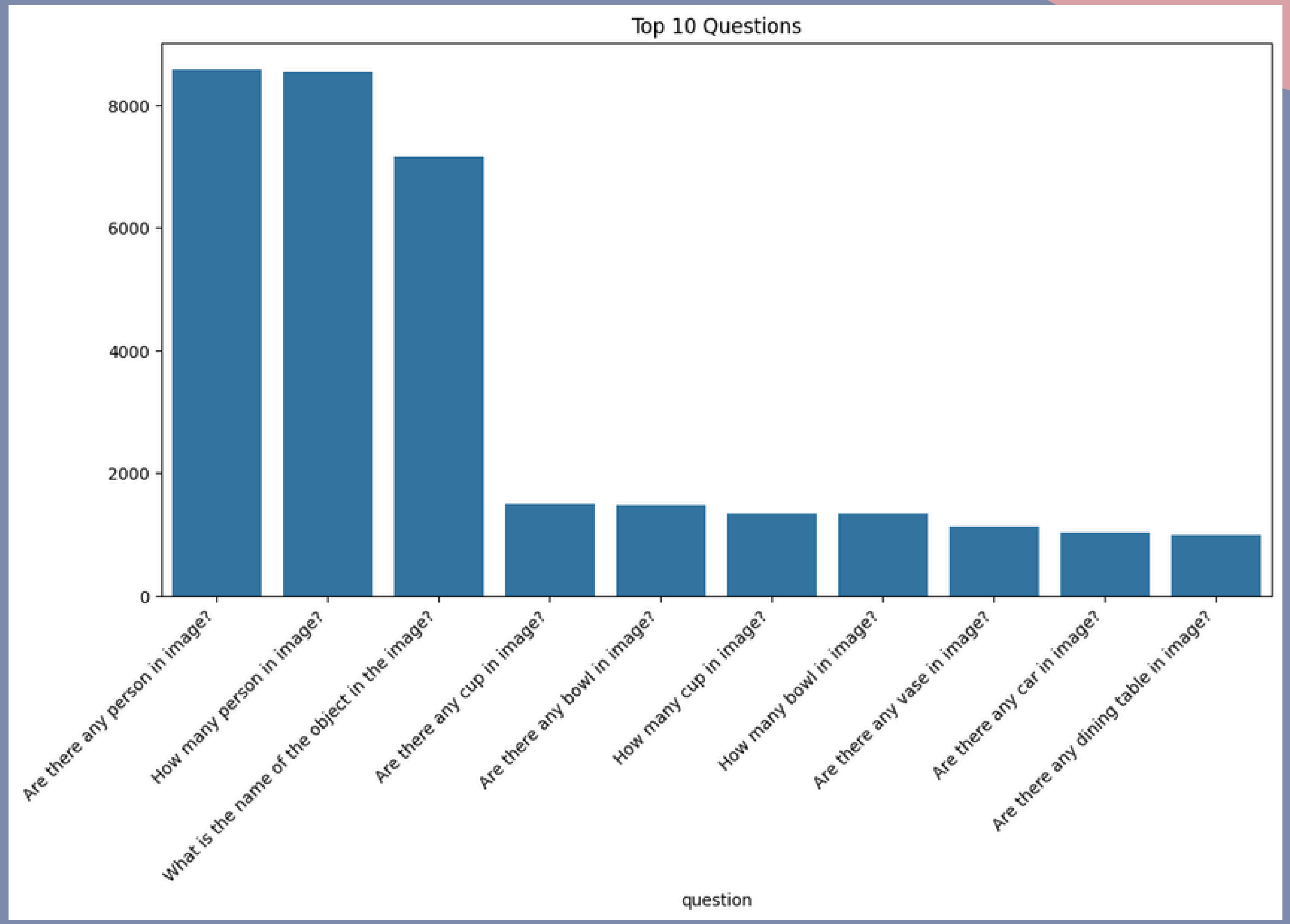
Images: 4486
Questions: 13460
Answers: 91

Train info

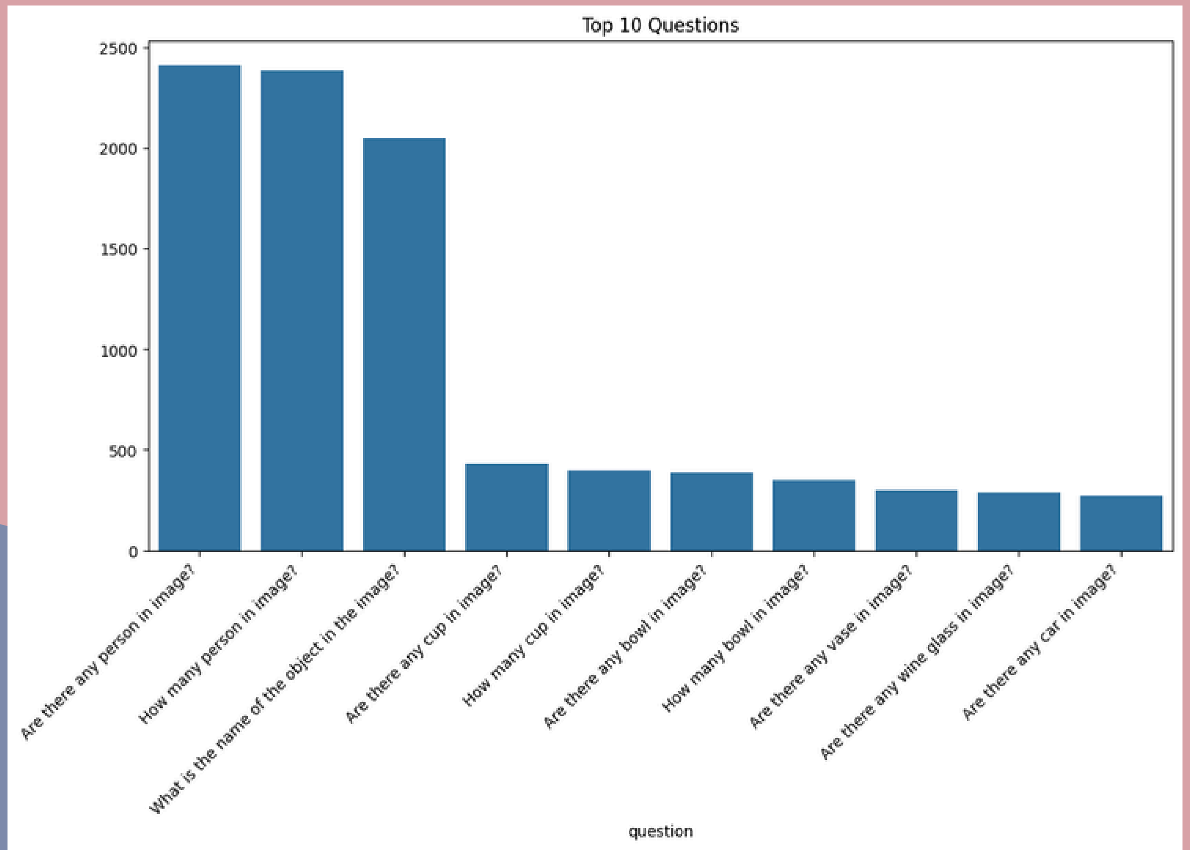
Images: 16105
Questions: 42073
Answers: 105



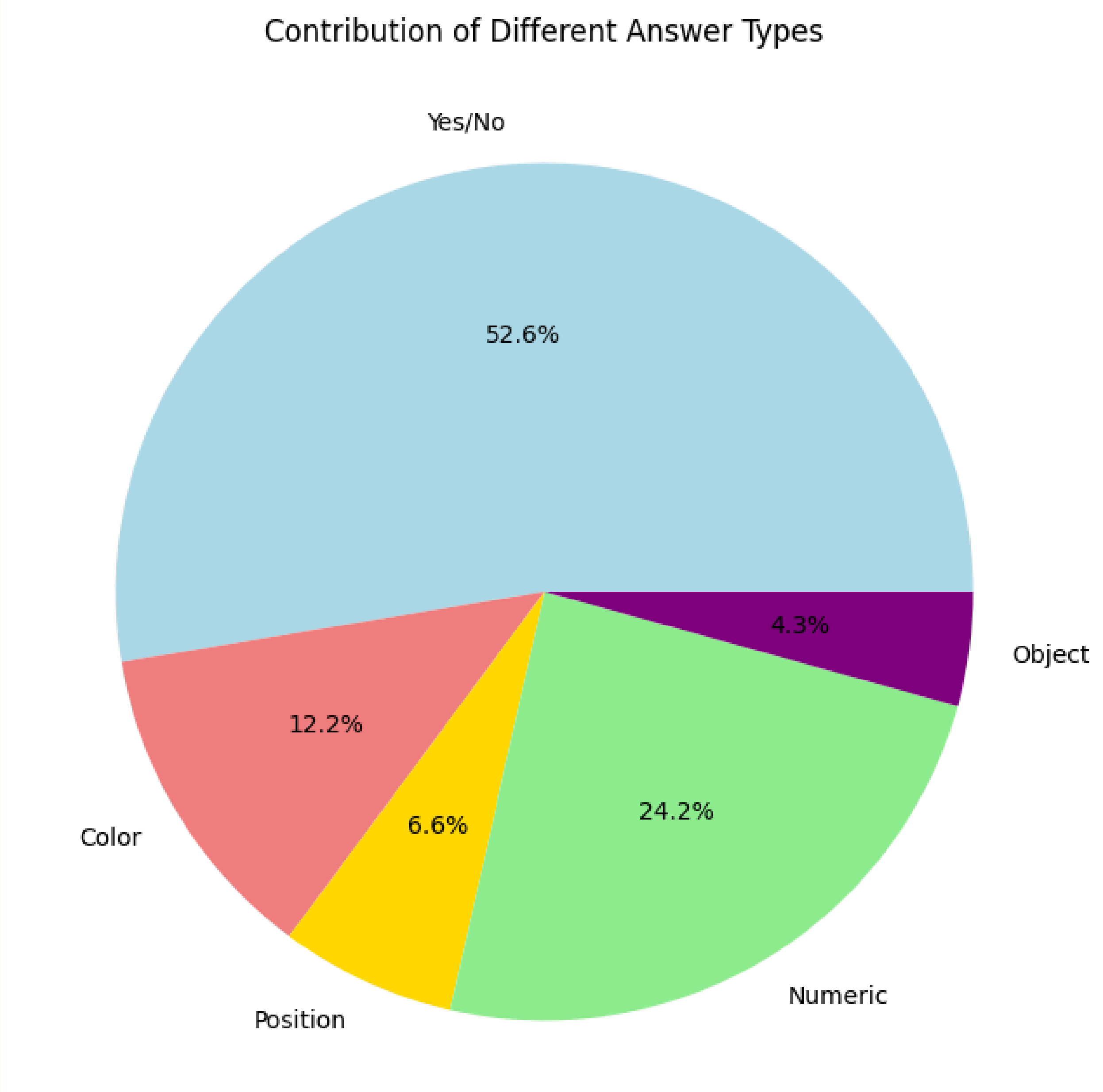
Test sets



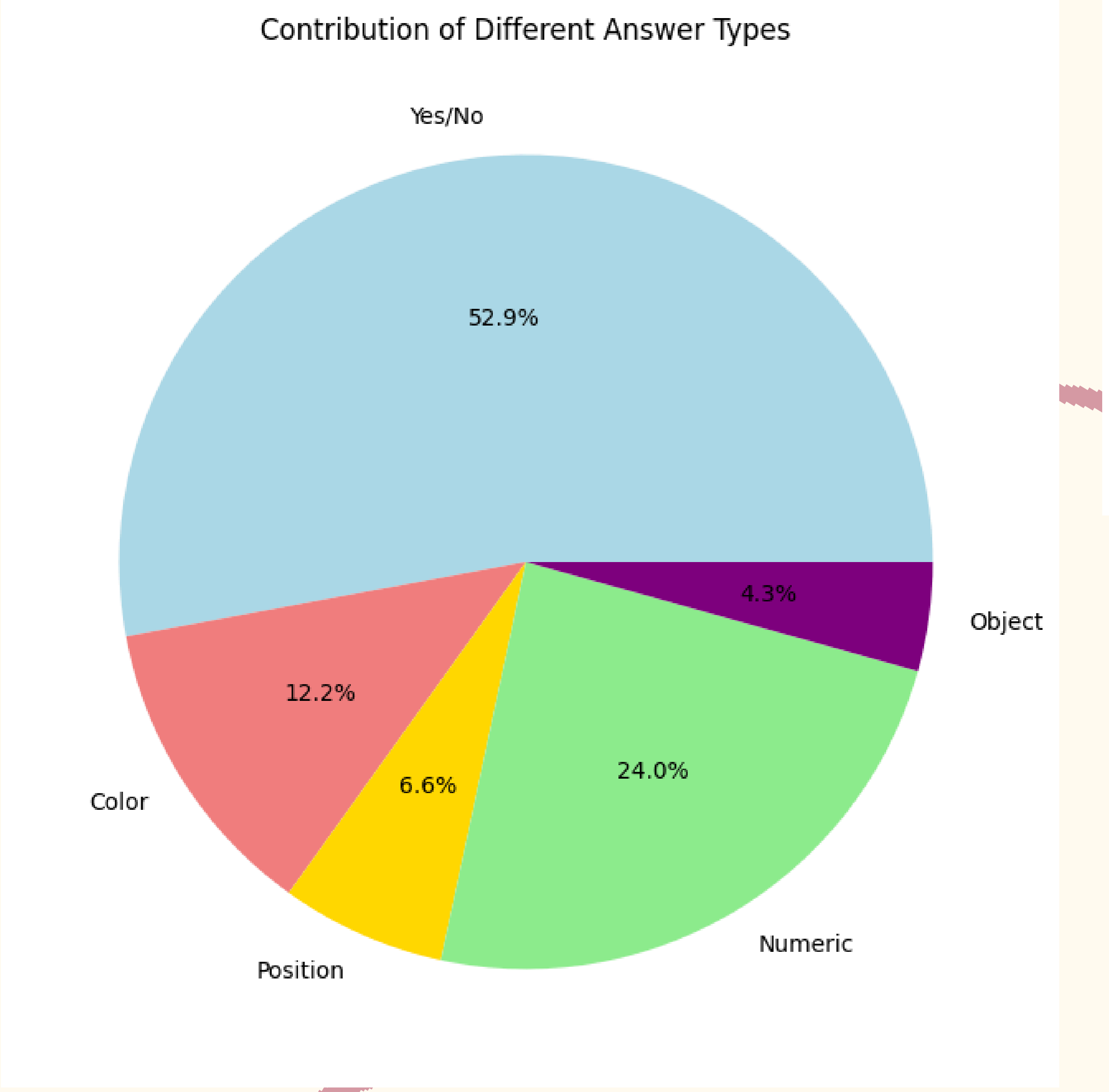
Train set



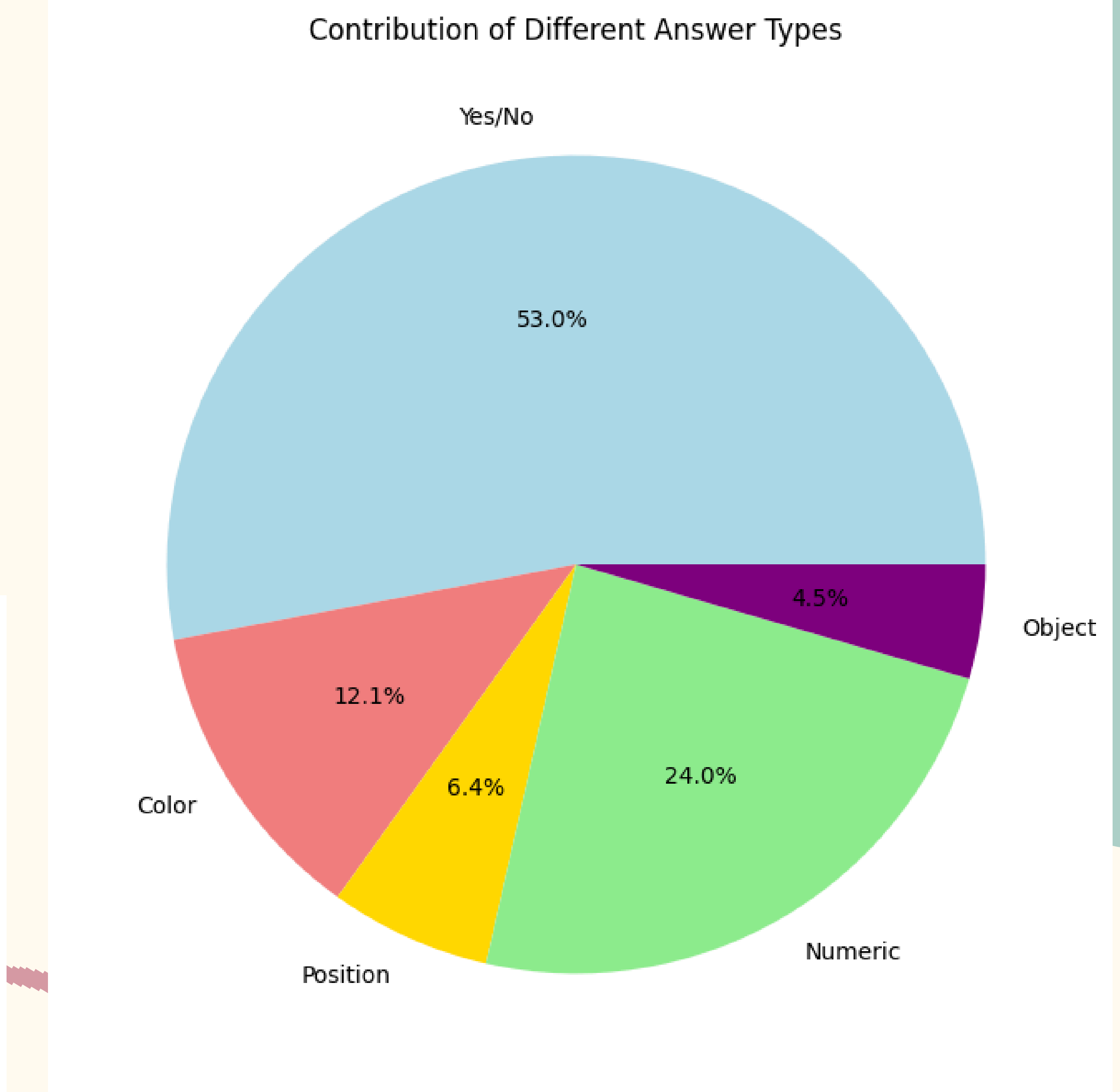
Eval set



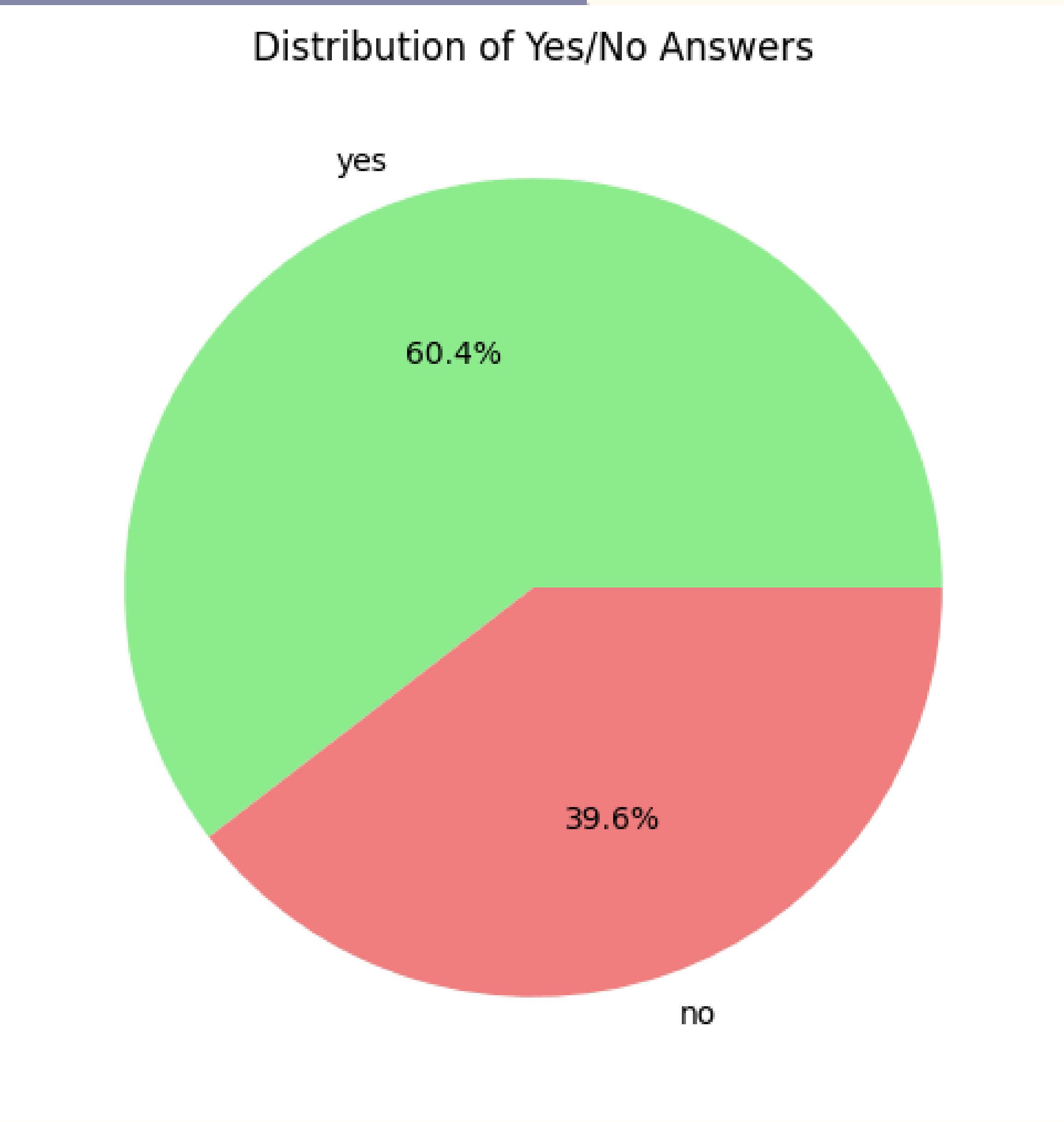
Test set



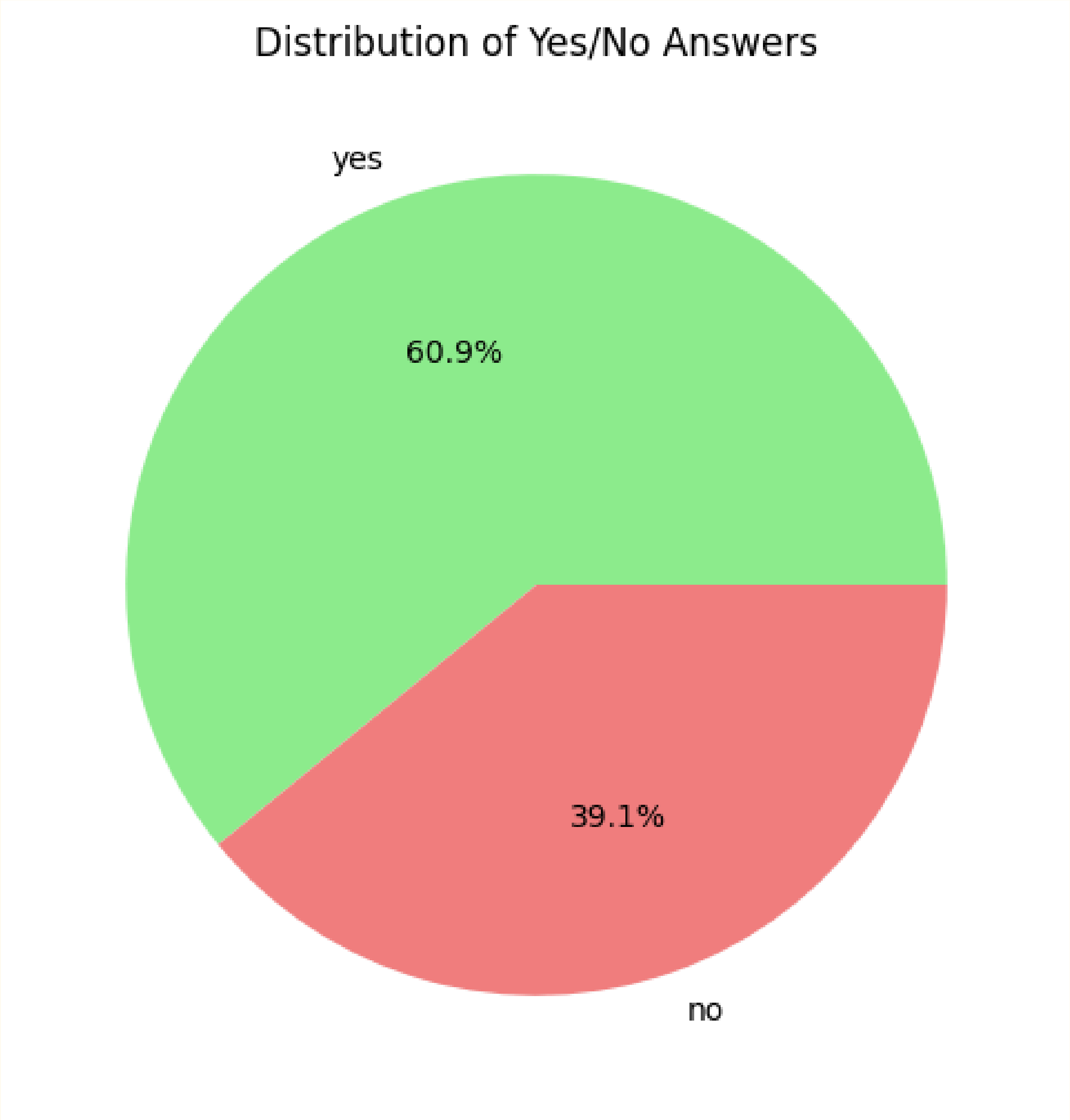
Train set



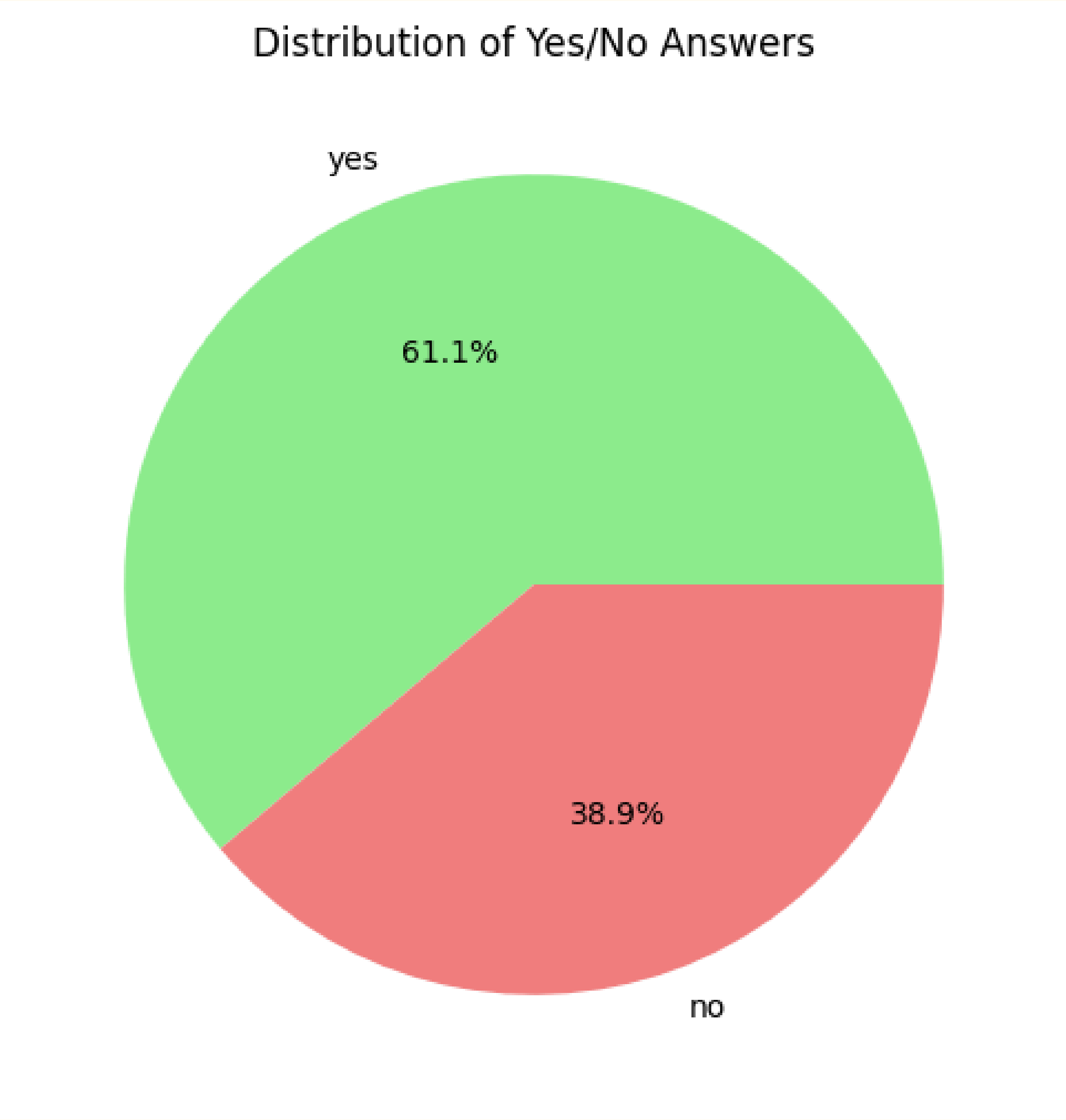
Eval set



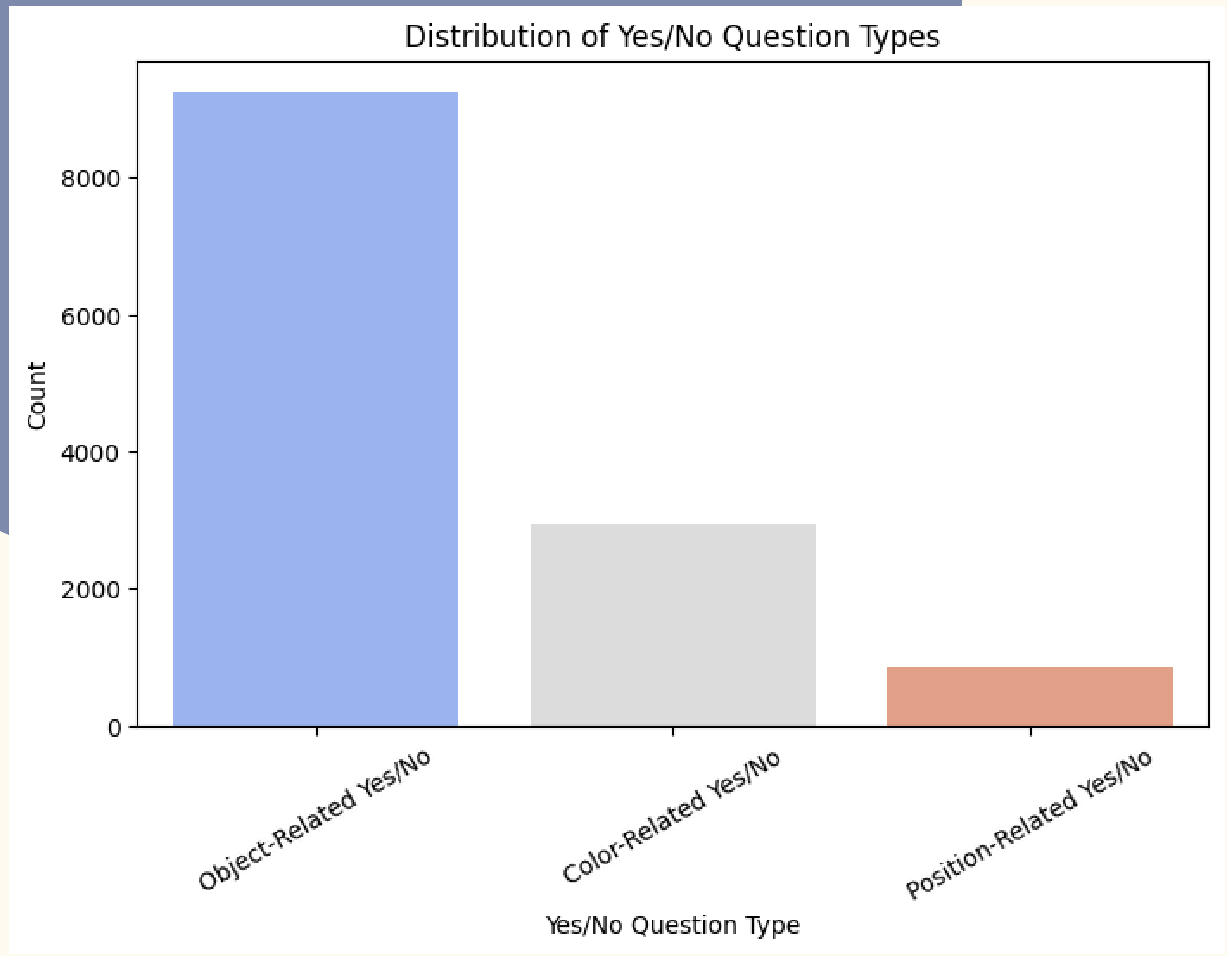
Test set



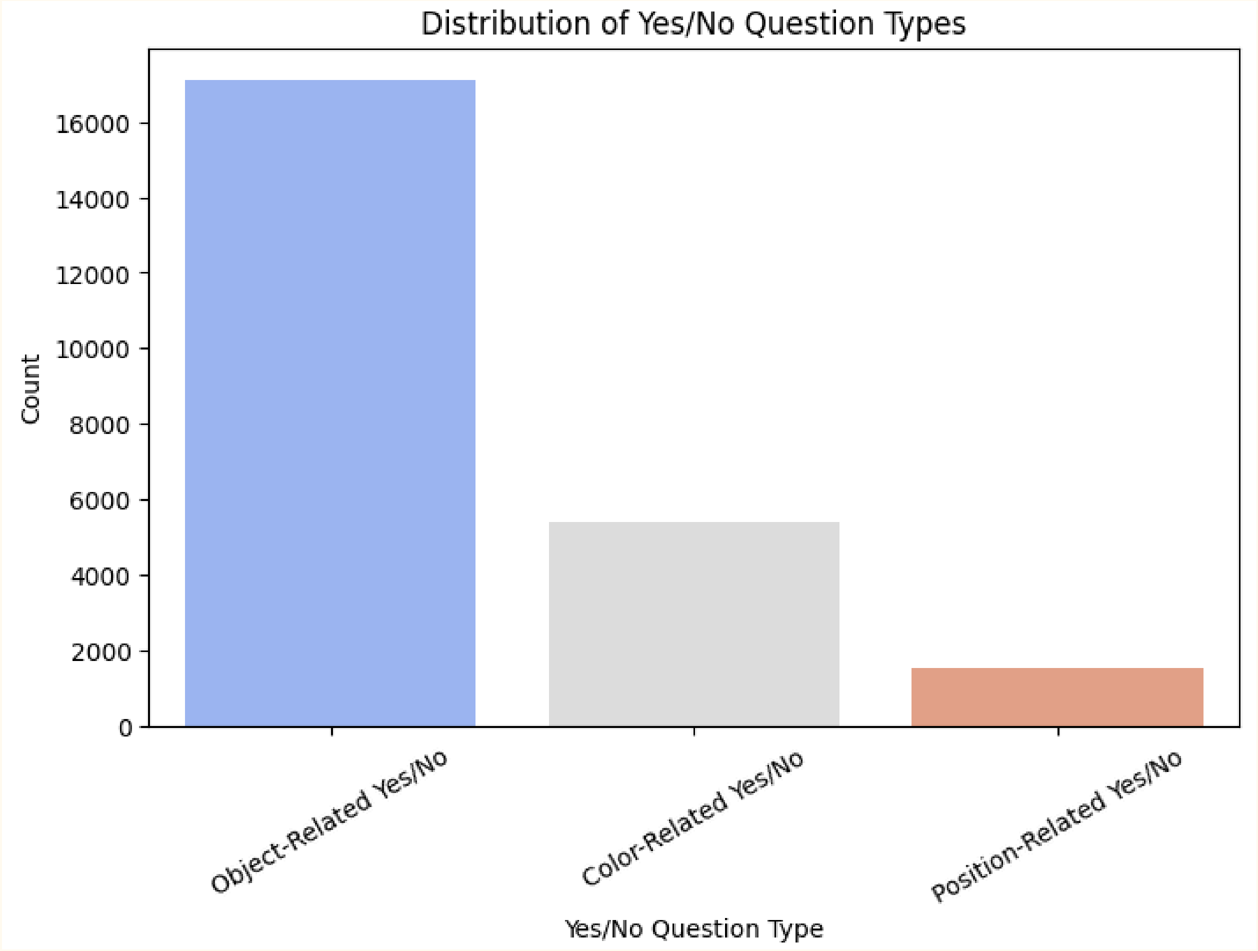
Train set



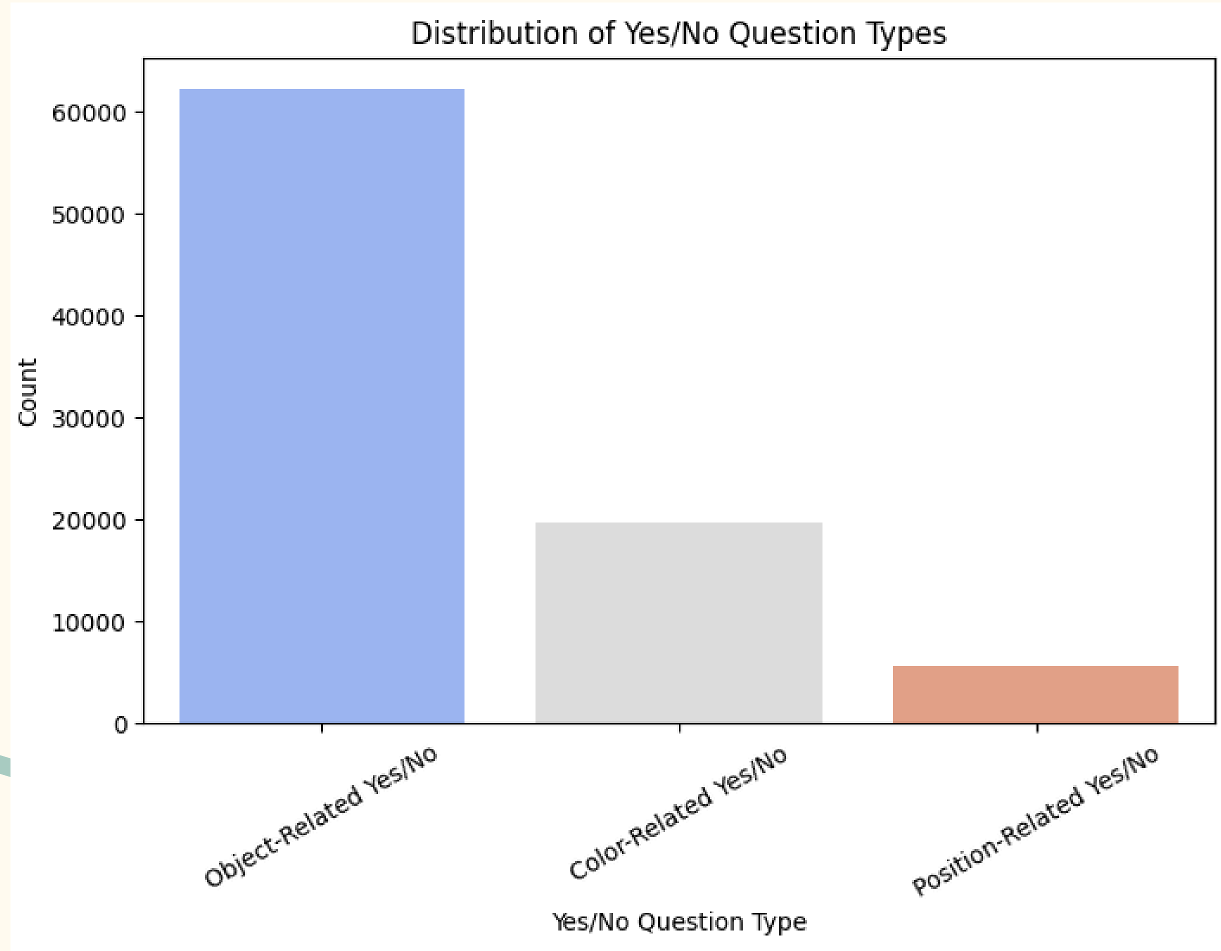
Eval set



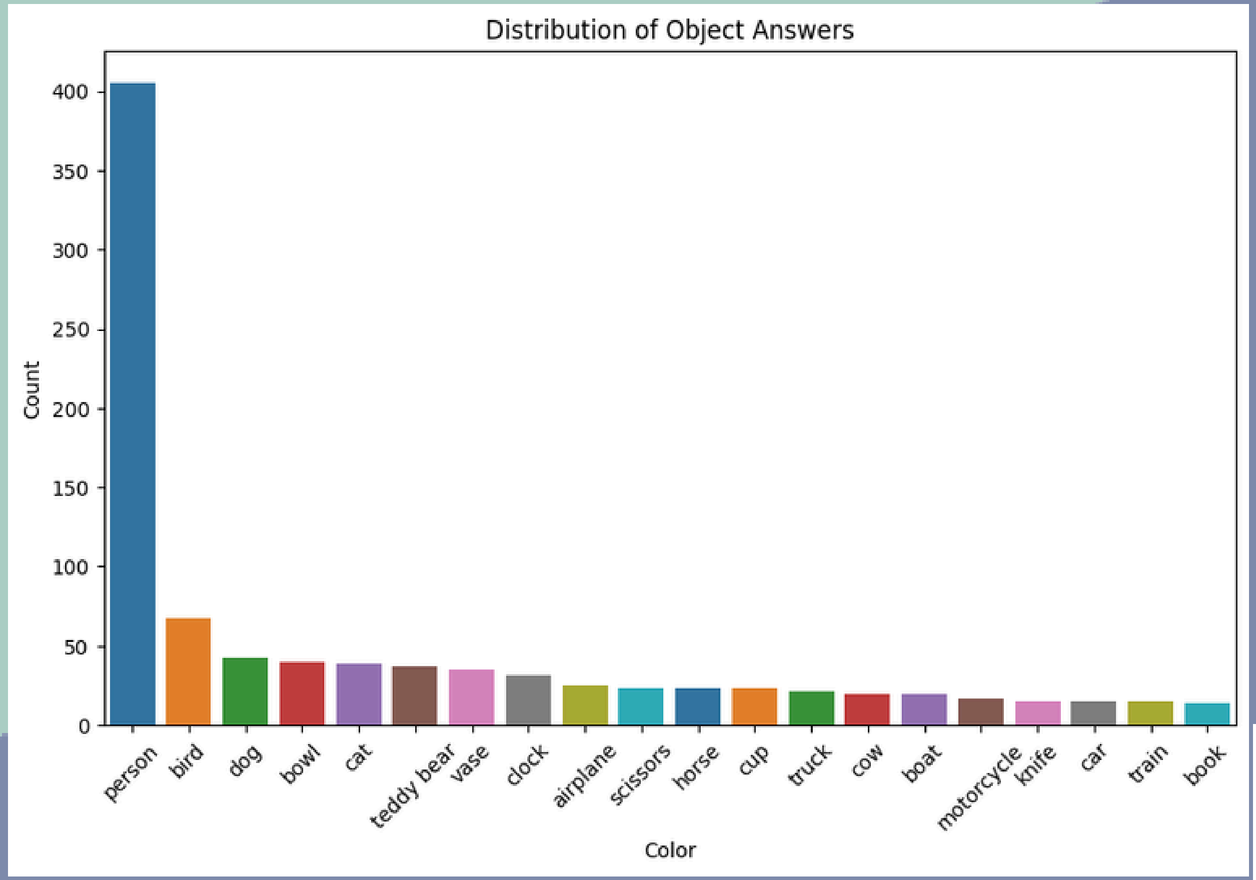
Test set



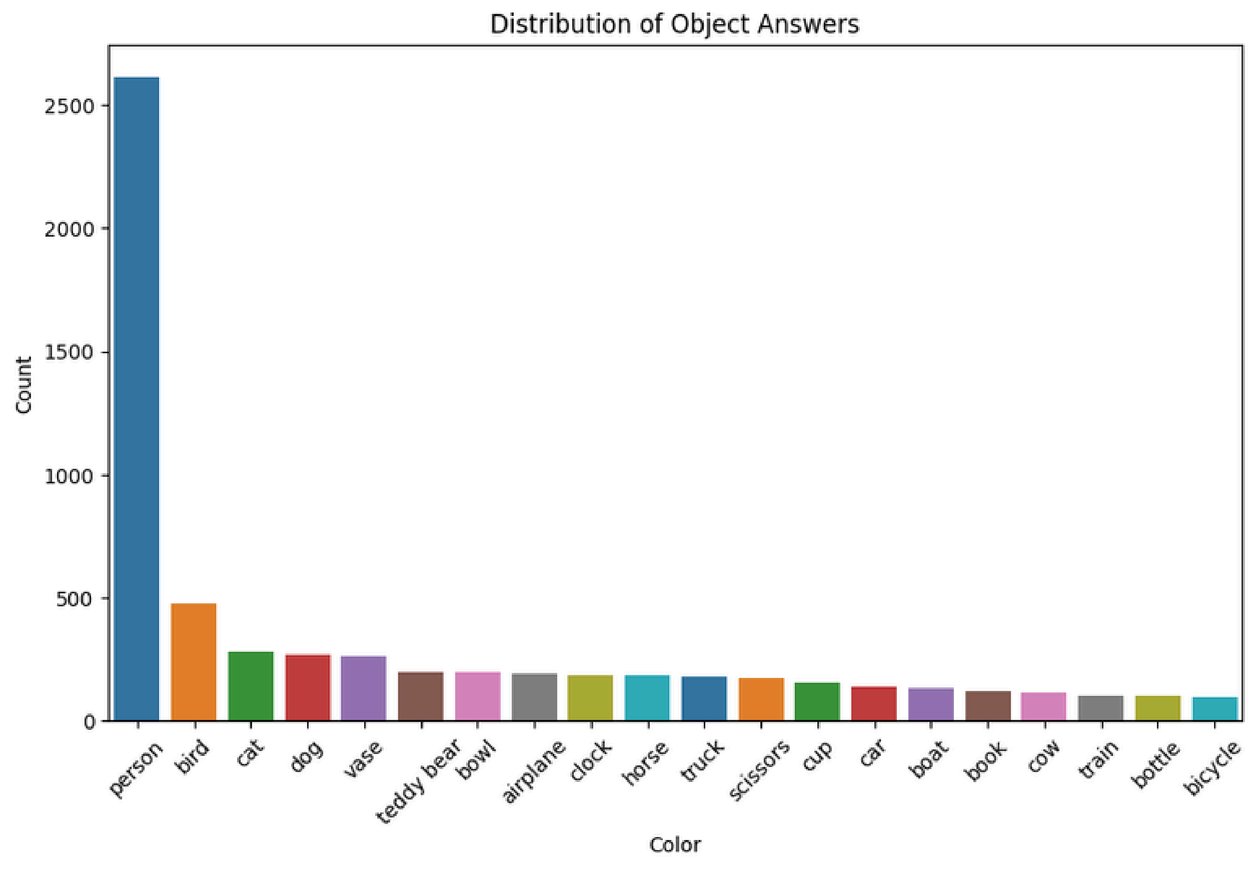
Eval set



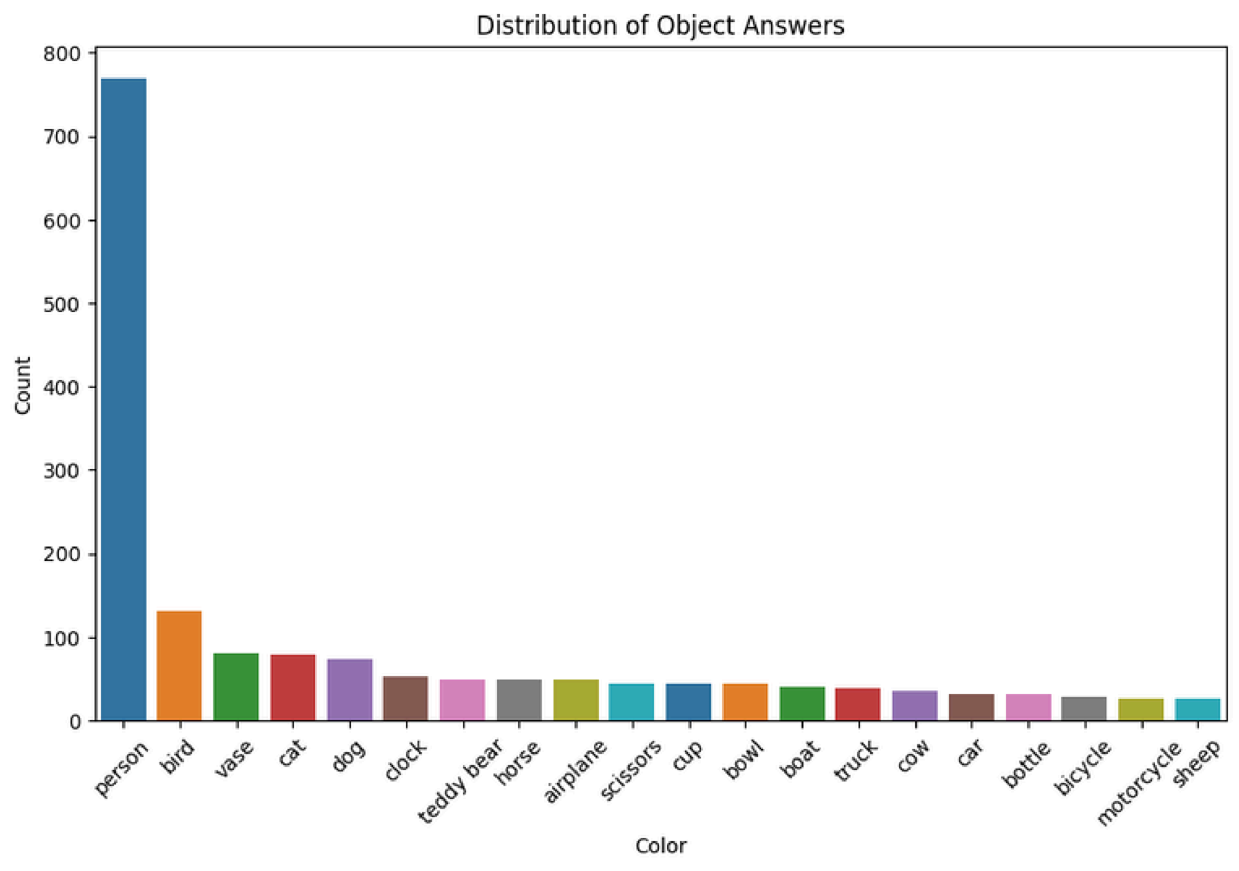
Train set



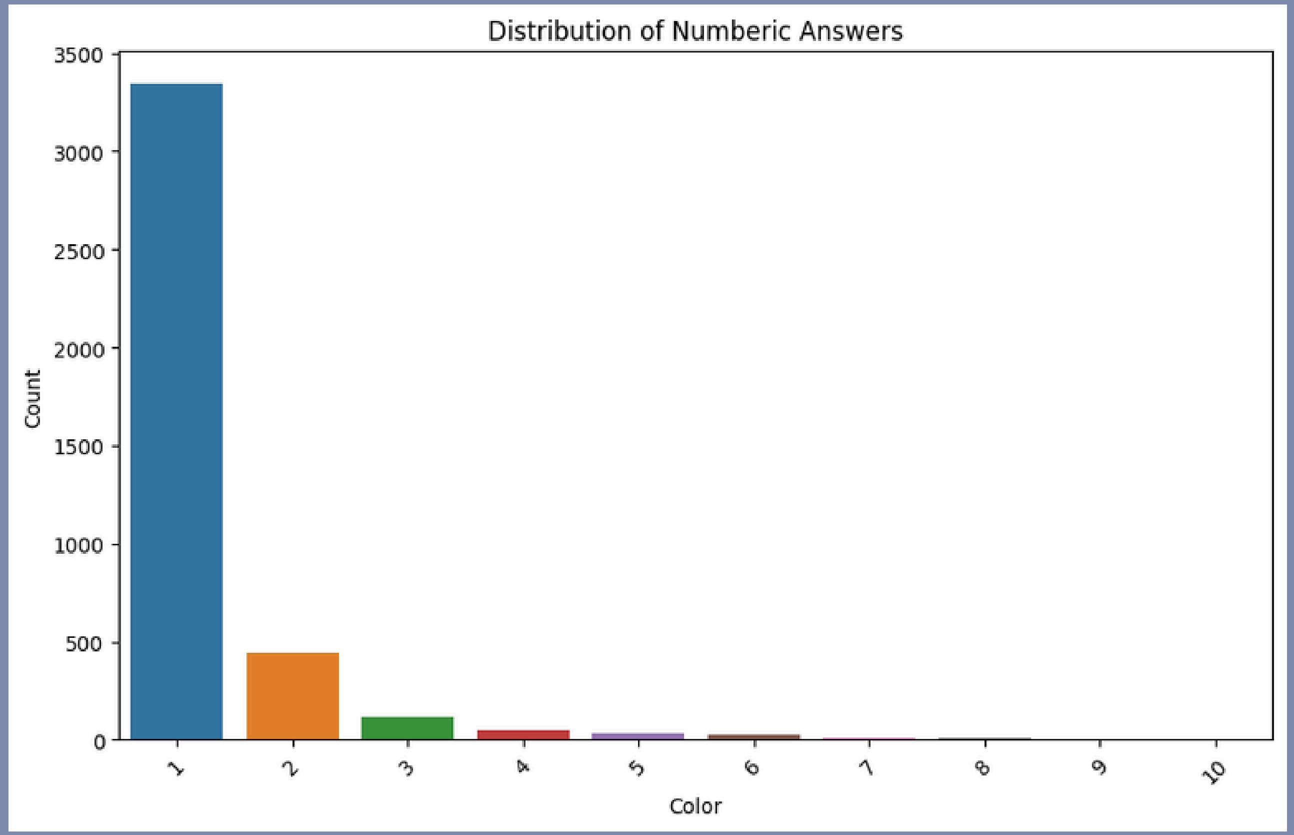
Test set



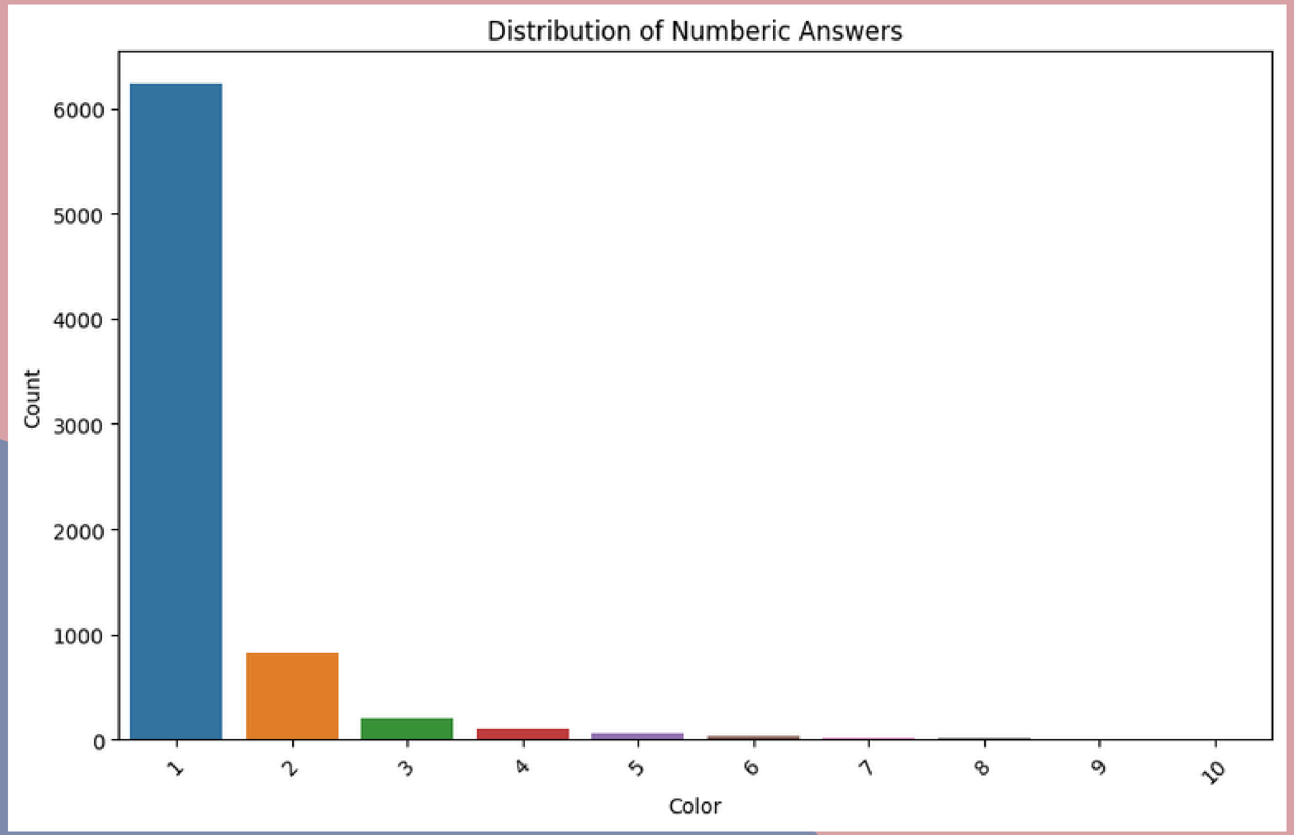
Train set



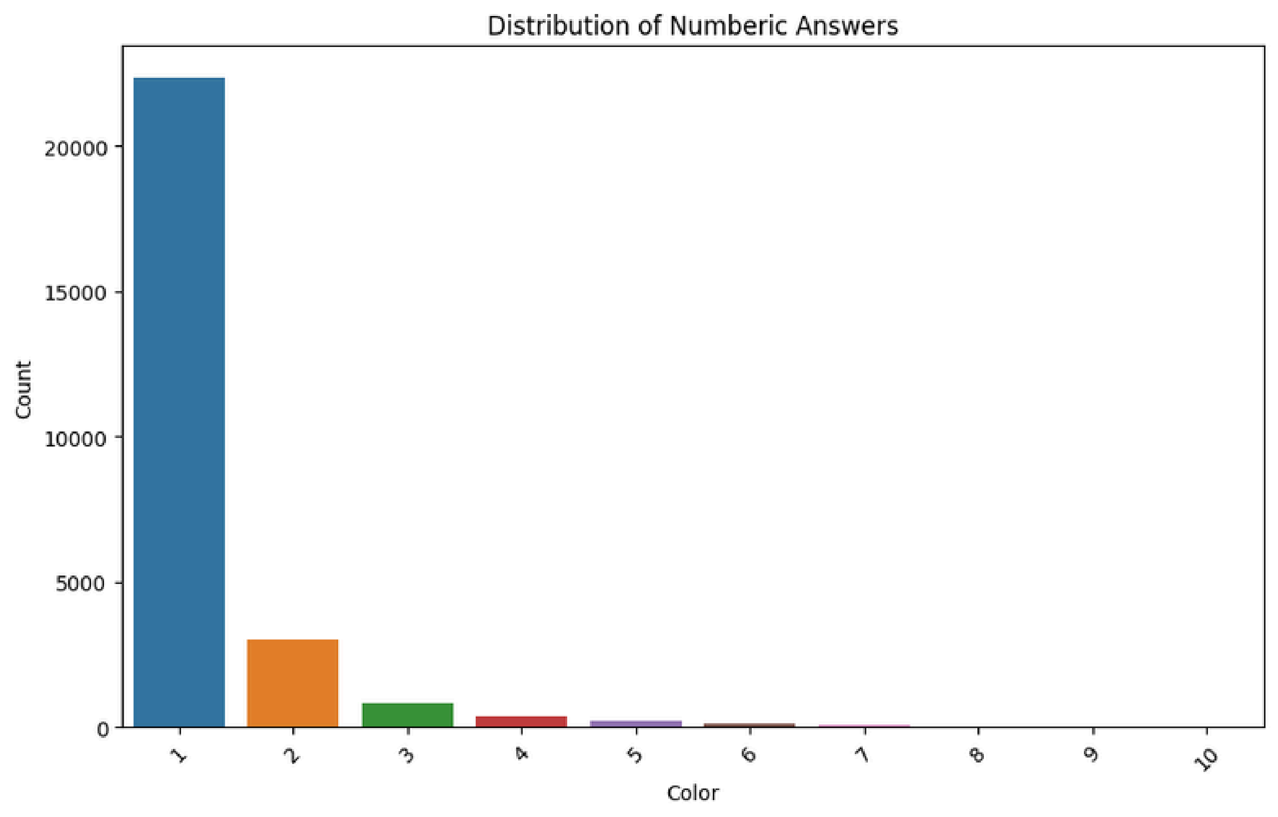
Eval set



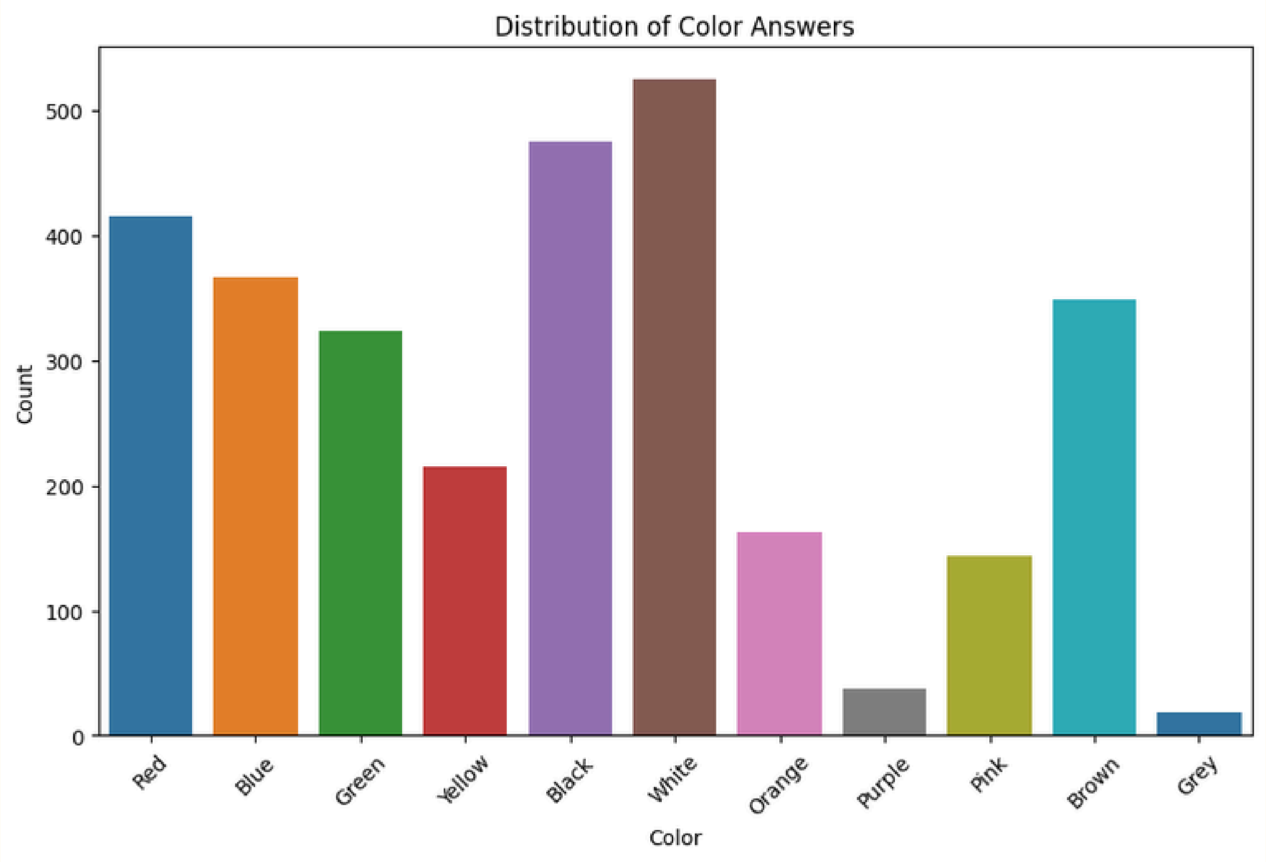
Test set



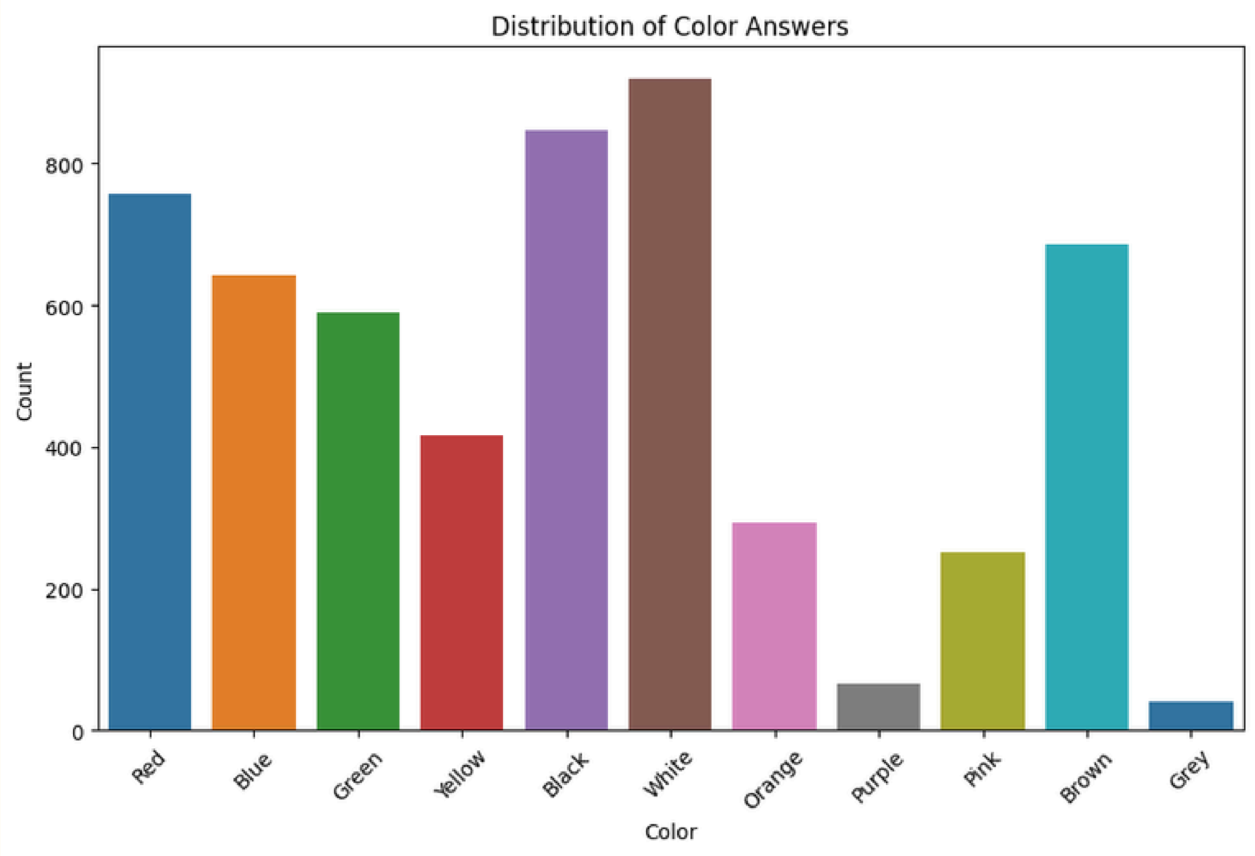
Eval set



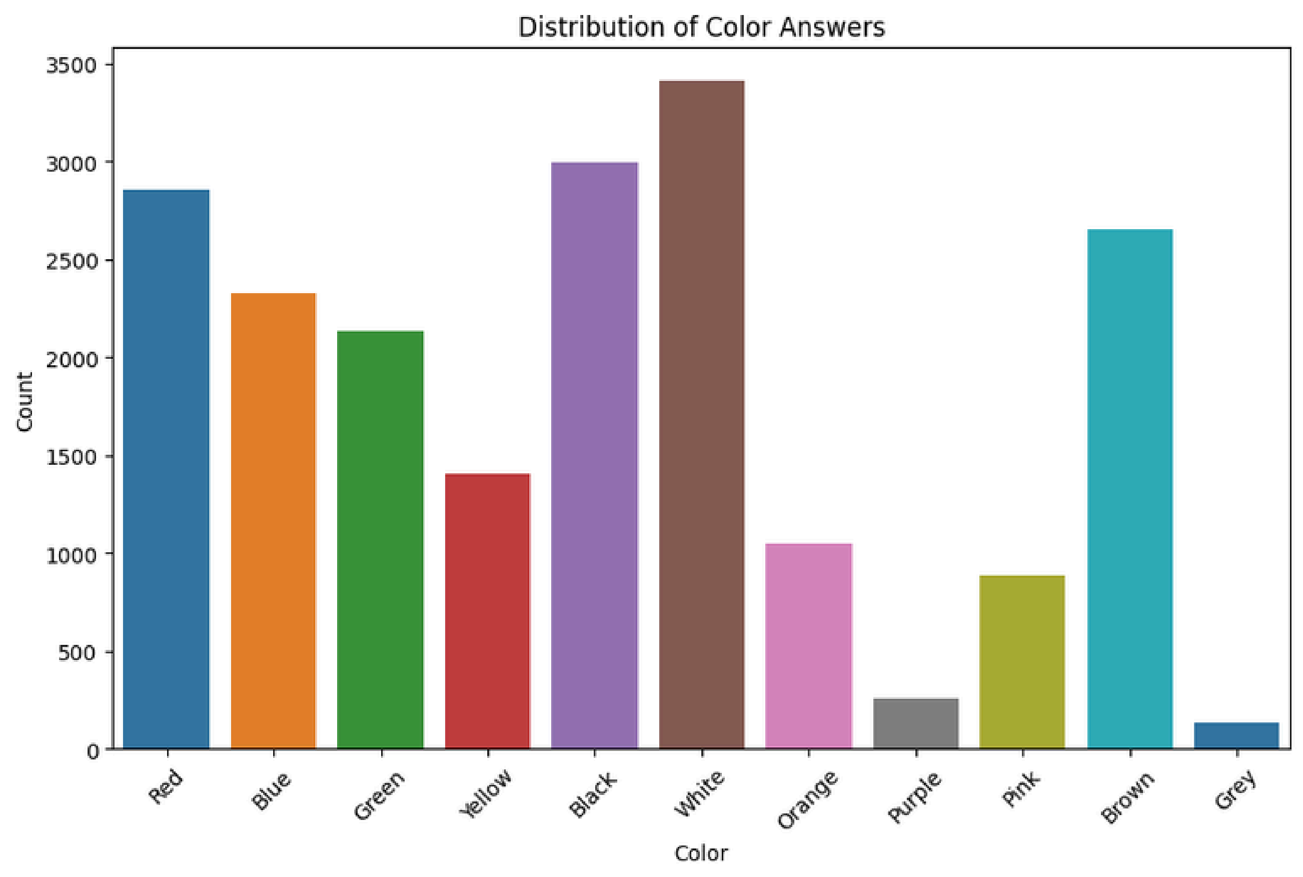
Train set



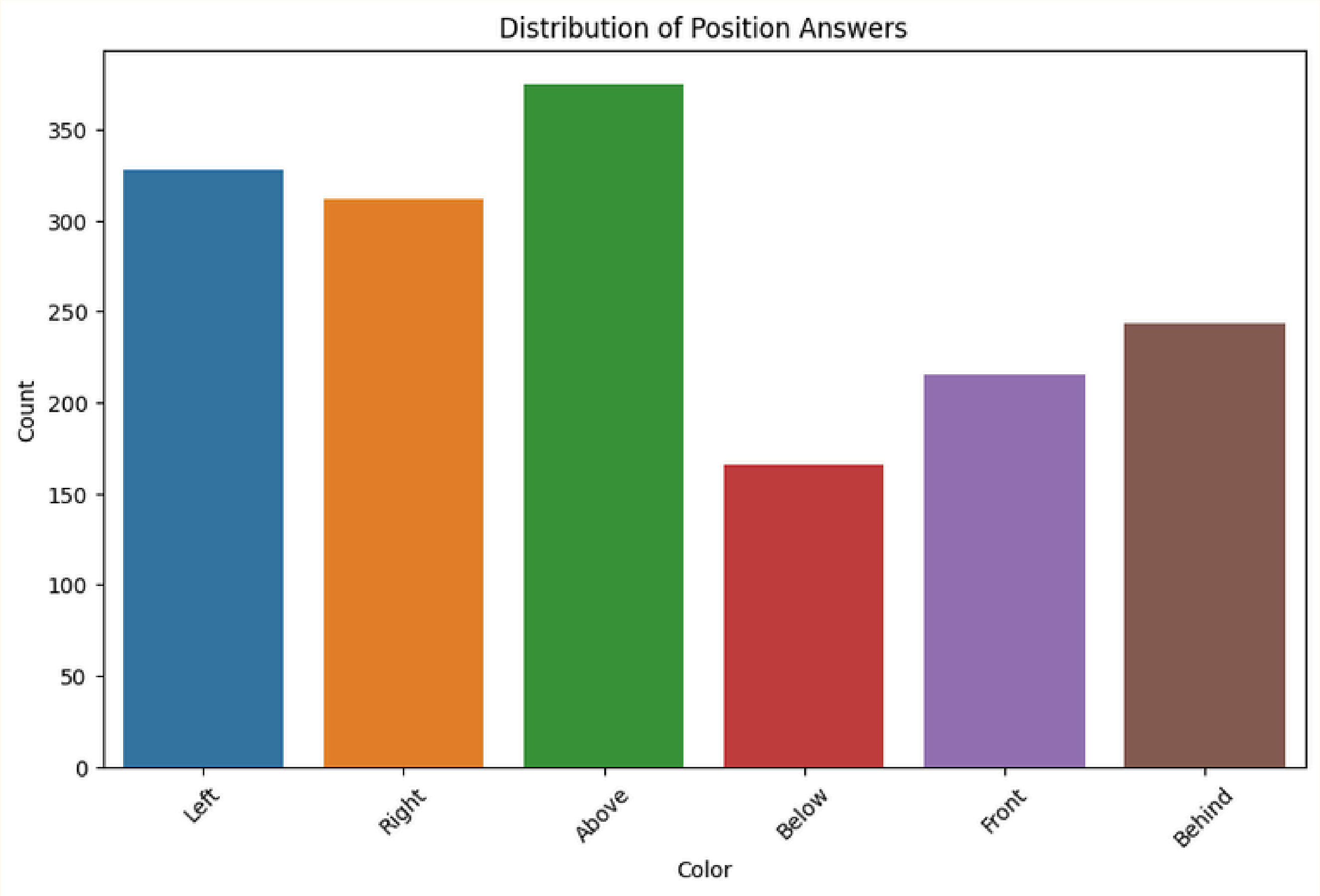
Test set



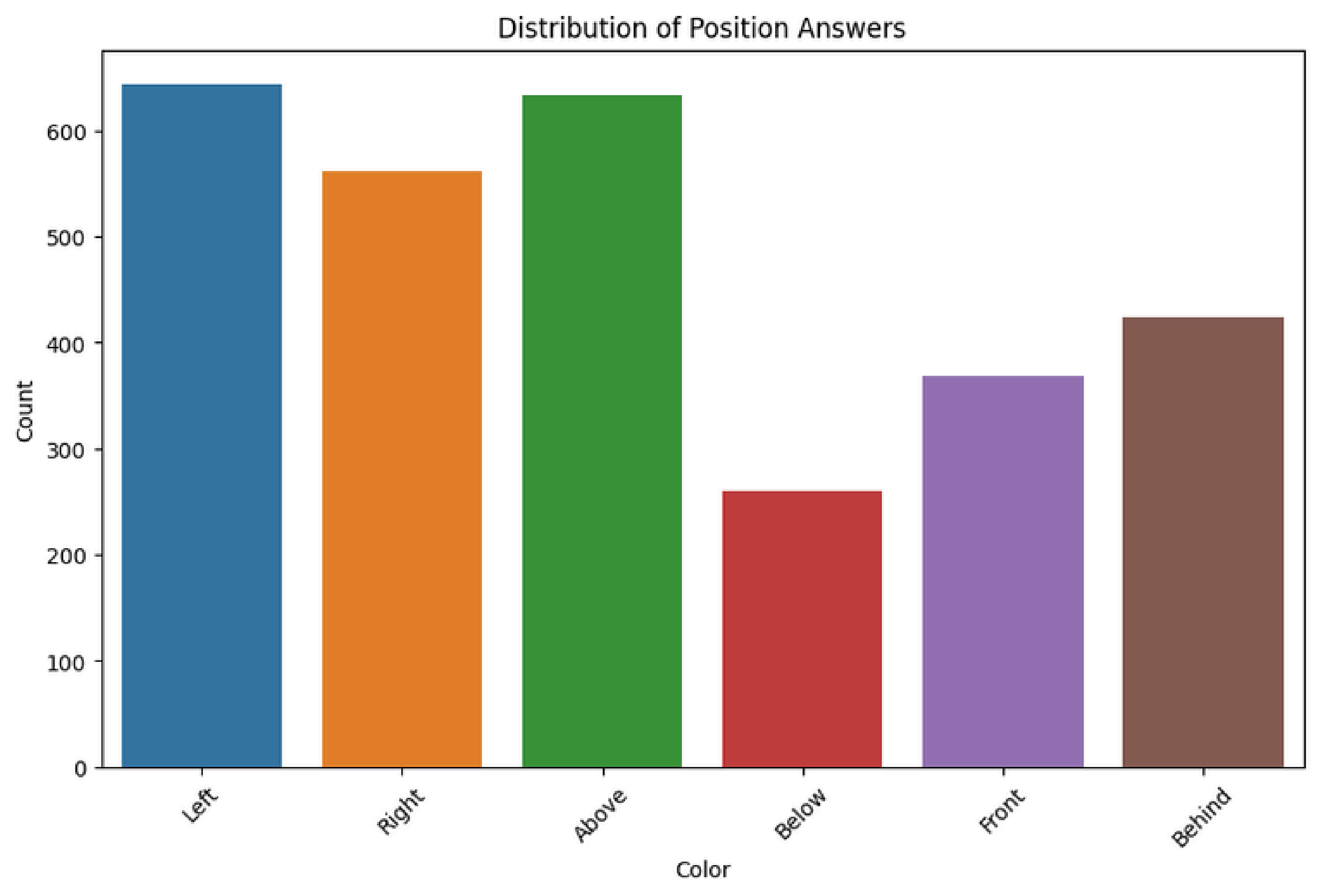
Eval set



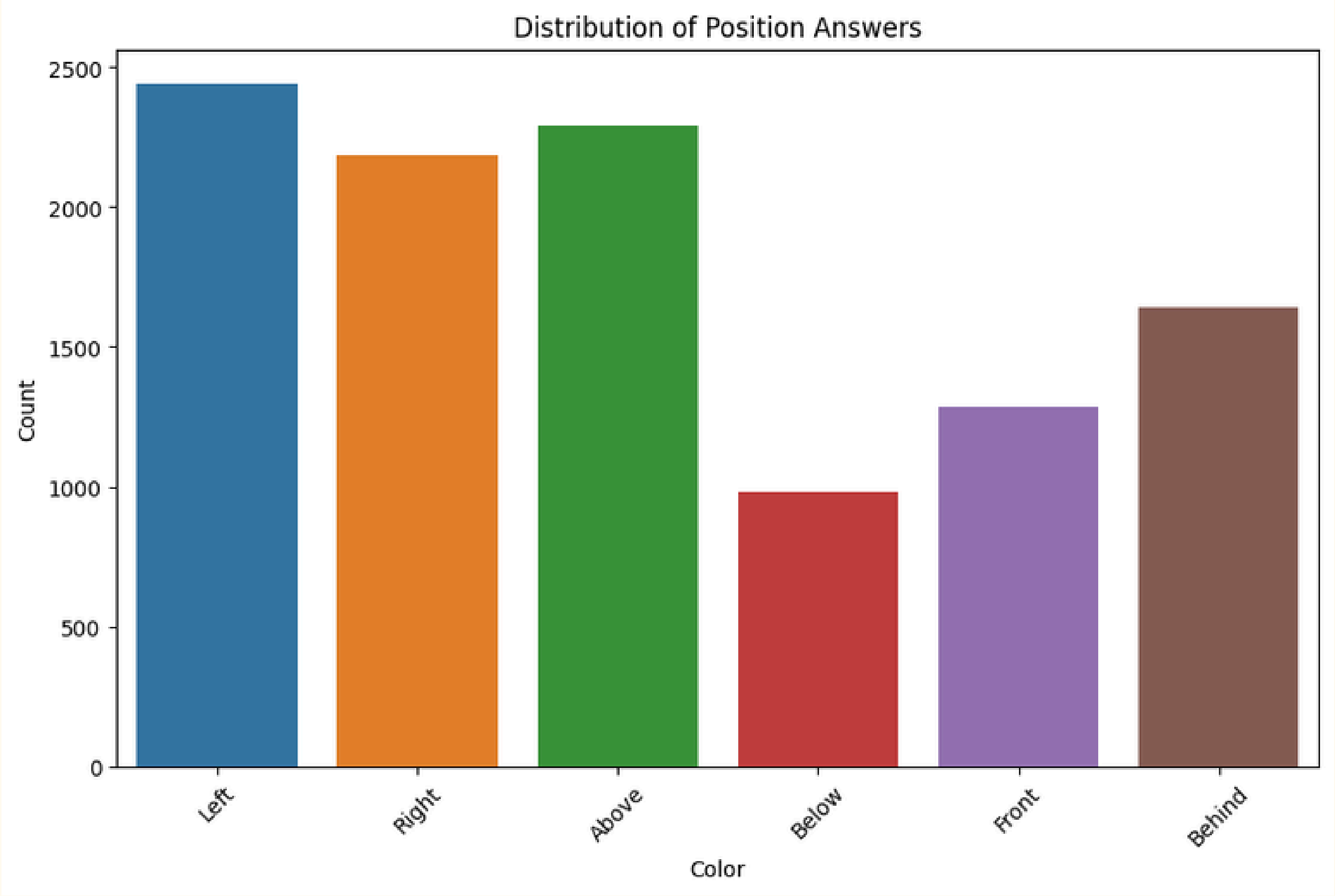
Train set



Test set



Eval set



Train set

Models

ViLt

Beit-3

ViLT

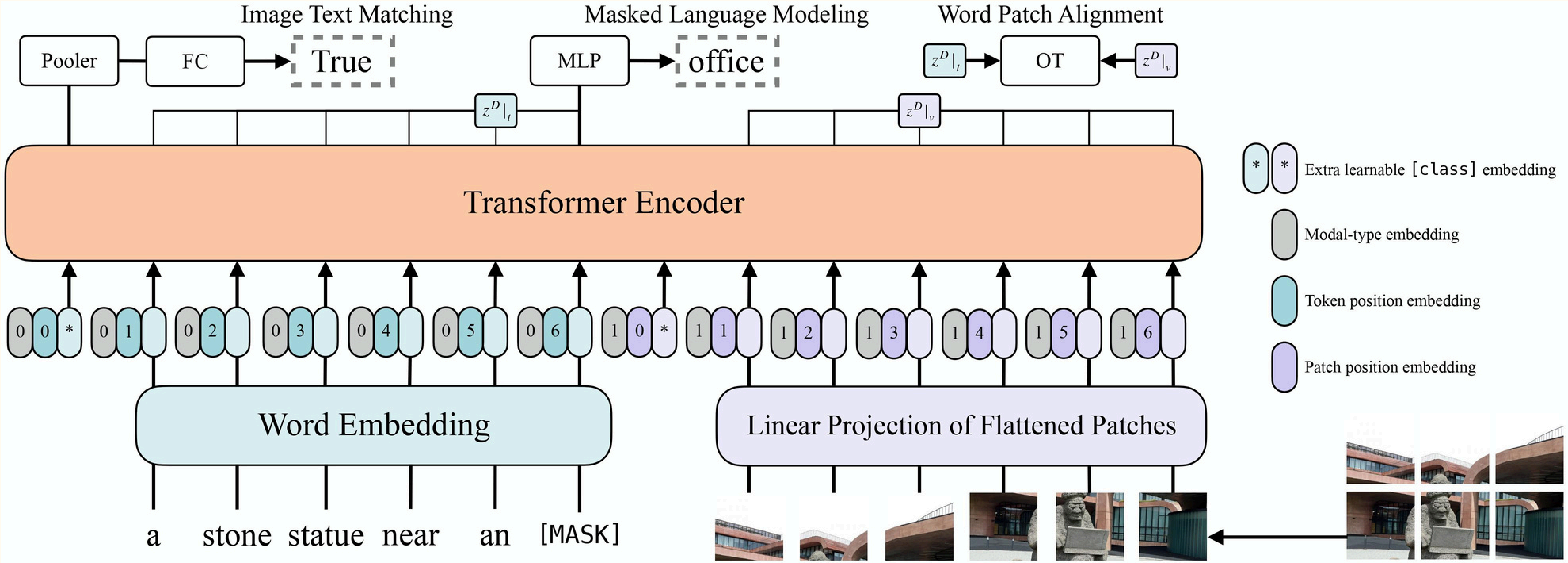
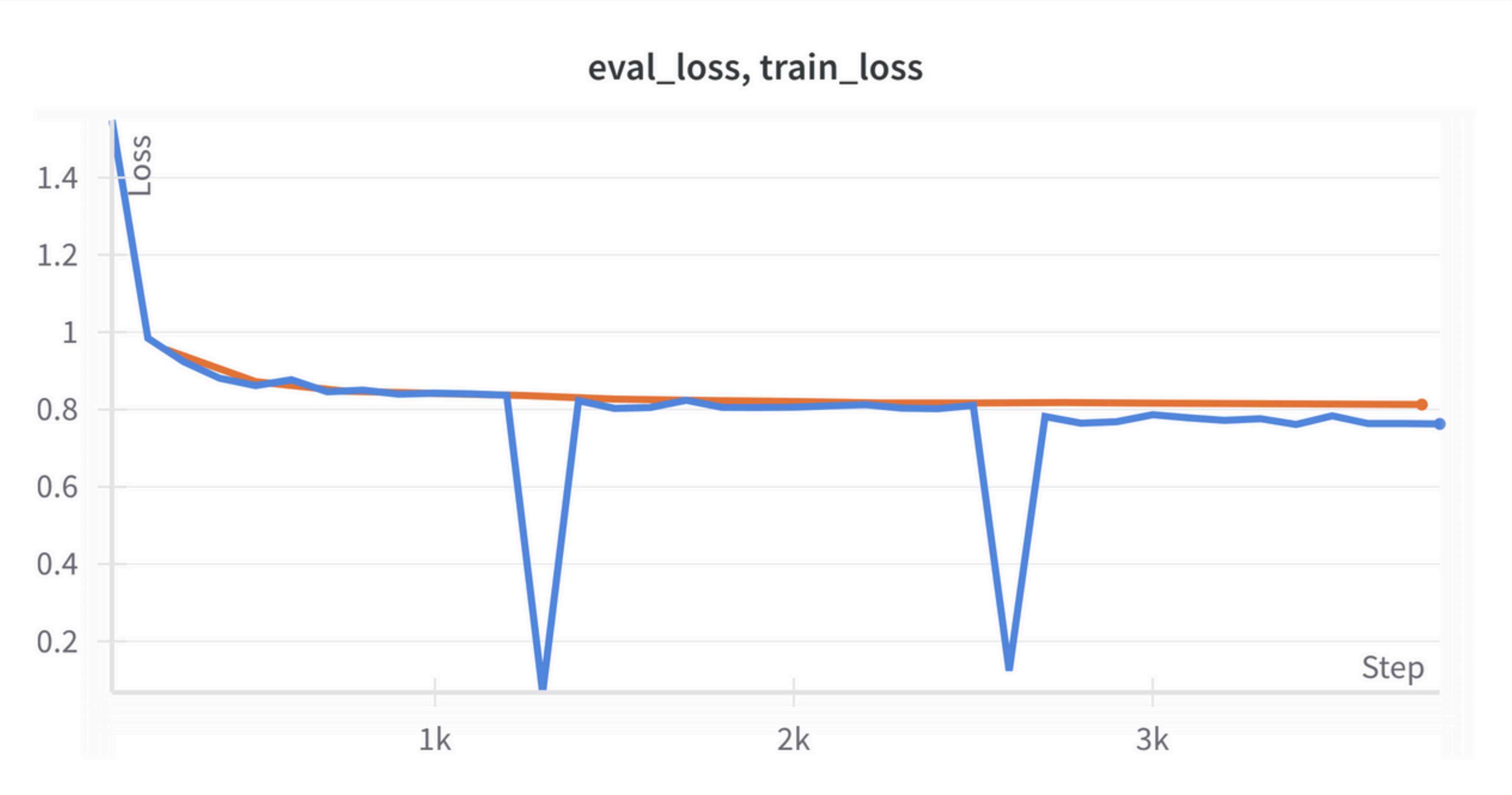
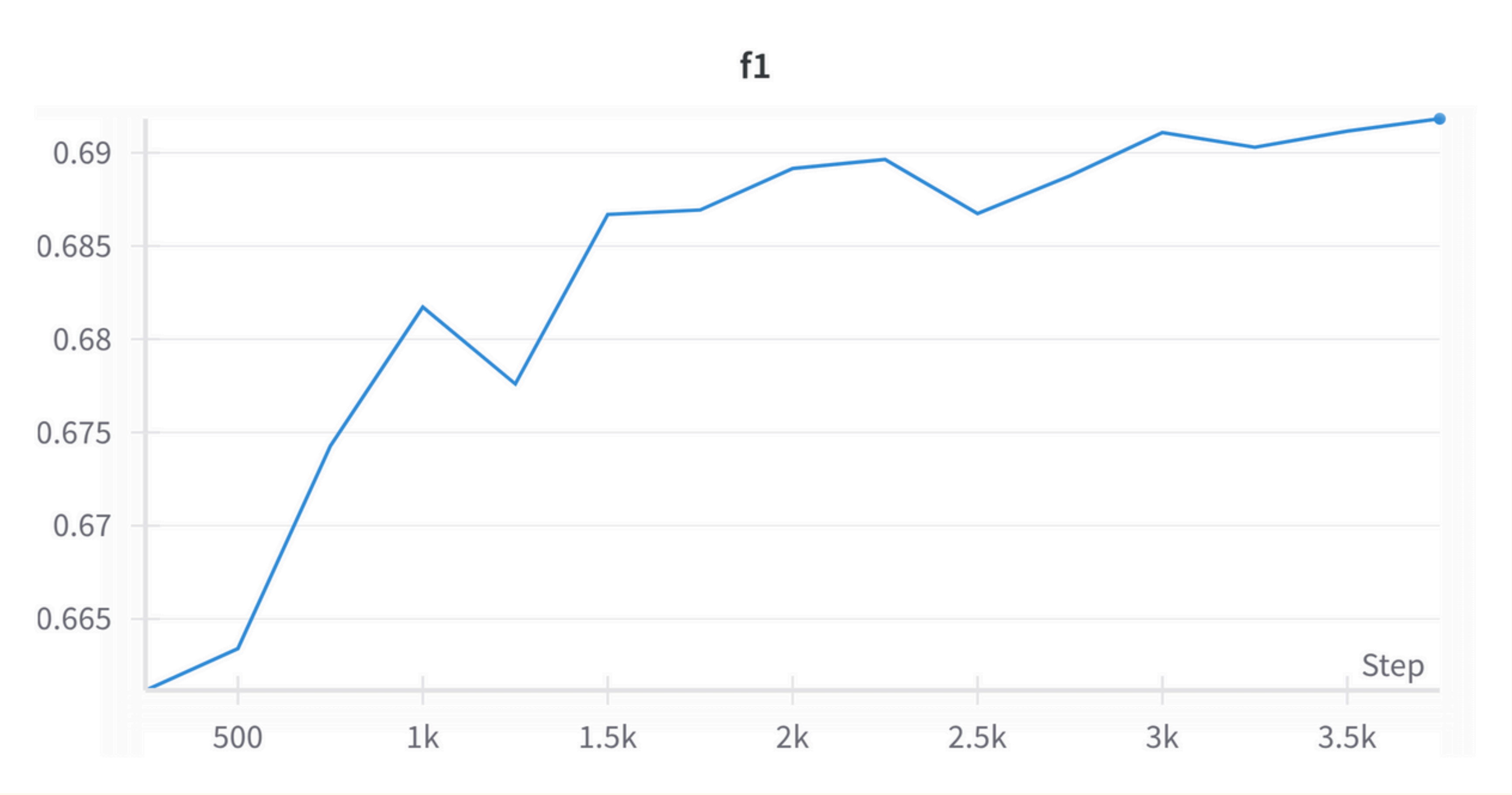
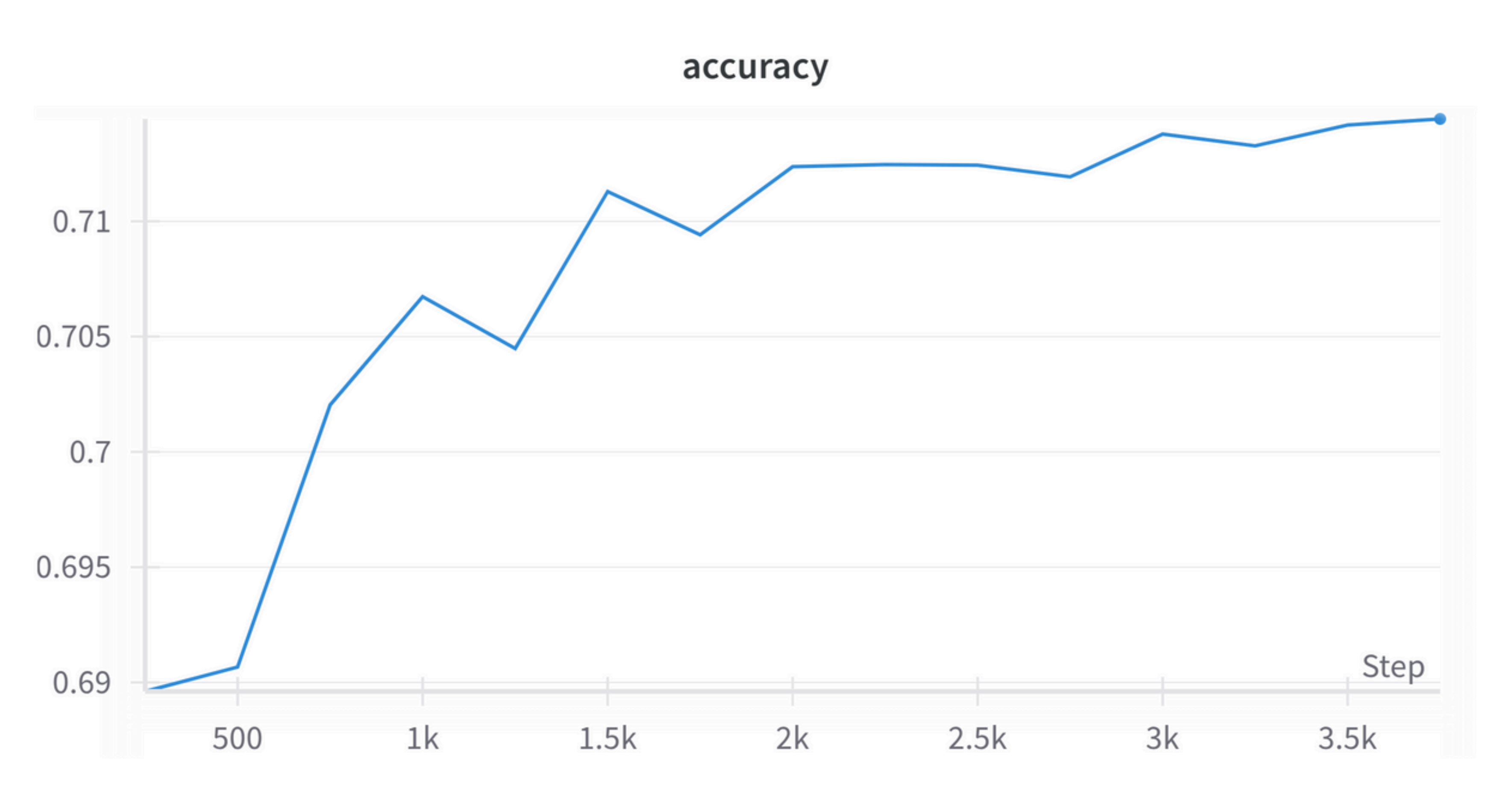


Image Text Matching
Masked Language Modeling
Word Patch Alignment
Pooler
FC
True
MLP
office
$z^D|_t$
OT
$z^D|_v$
Transformer Encoder
Extra learnable [class] embedding
Modal-type embedding
Token position embedding
Patch position embedding
0 0 *
0 1
0 2
0 3
0 4
0 5
0 6
1 0 *
1 1
1 2
1 3
1 4
1 5
1 6
Word Embedding
Linear Projection of Flattened Patches
a
stone
statue
near
an
[MASK]

Finetune from dandelin/vilt-b32-finetuned-vqa







Beit-3

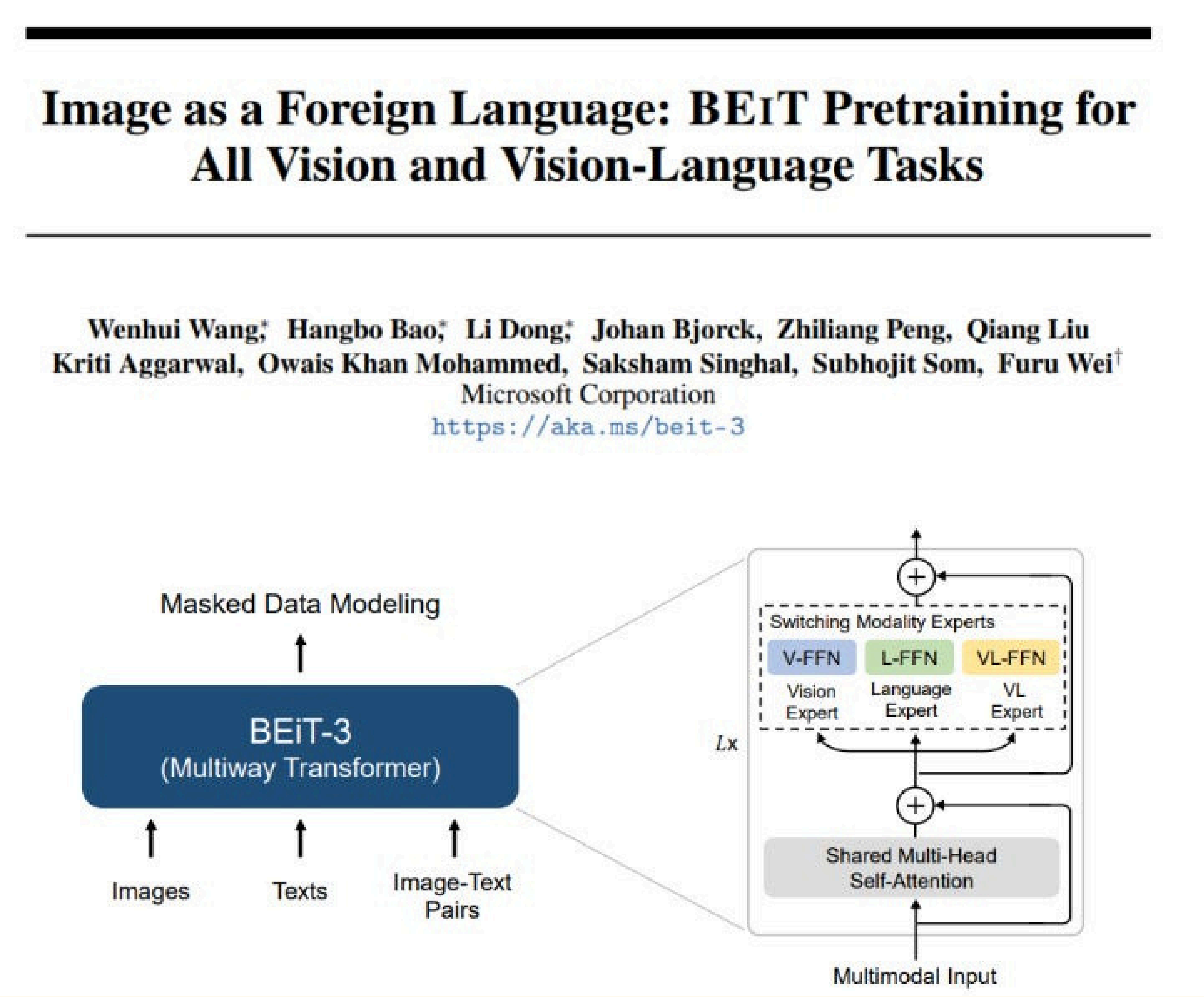
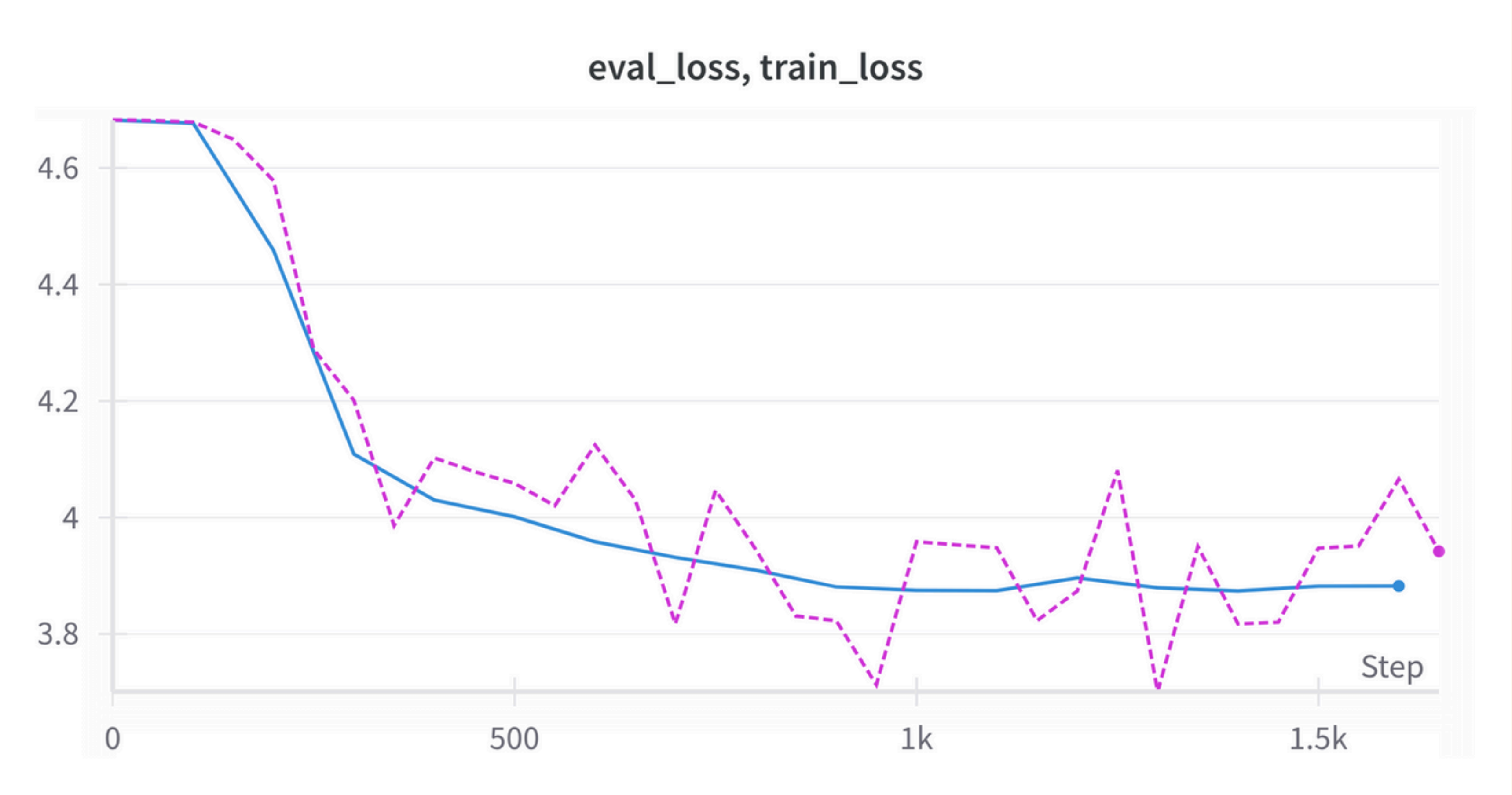
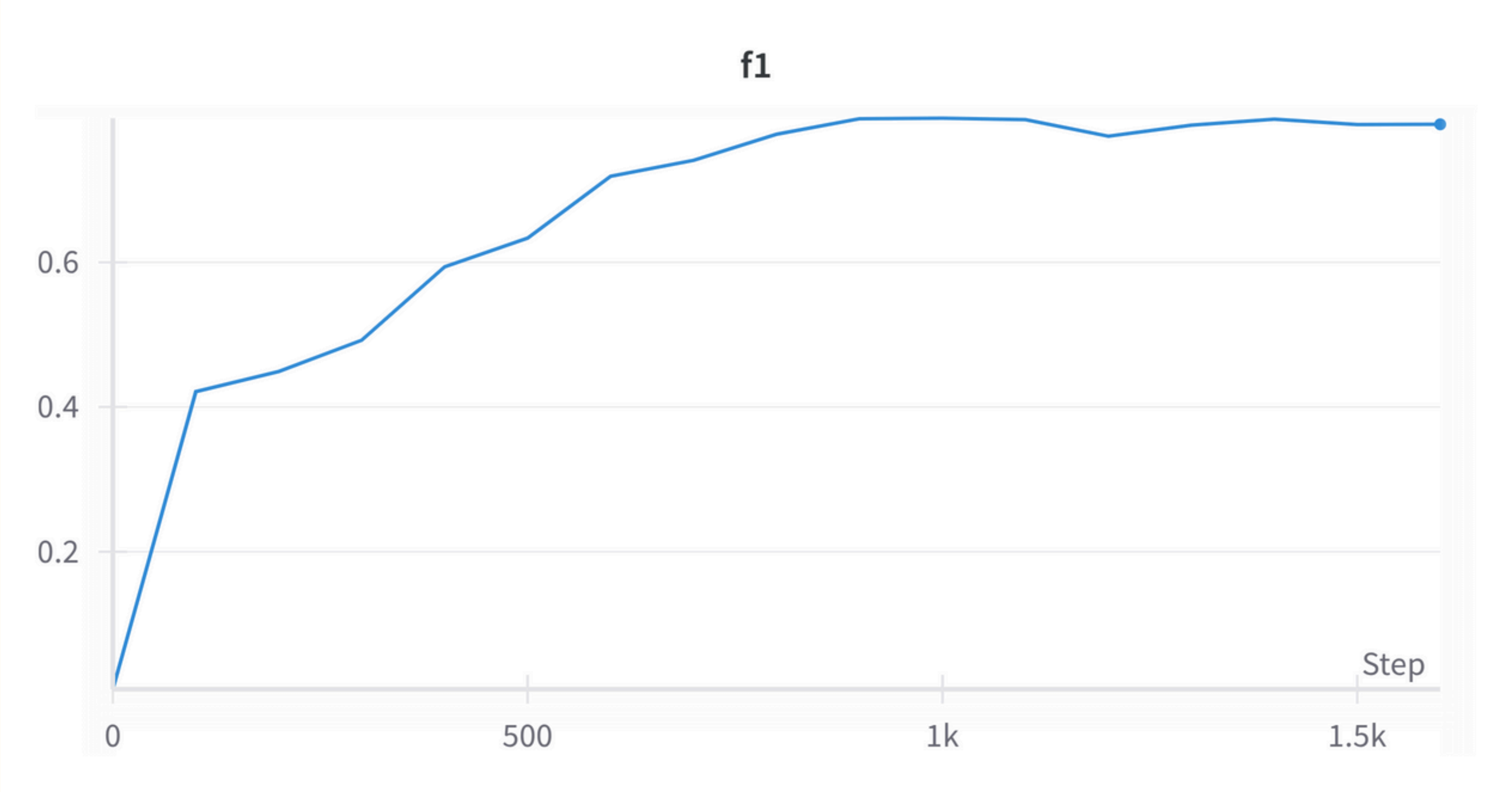


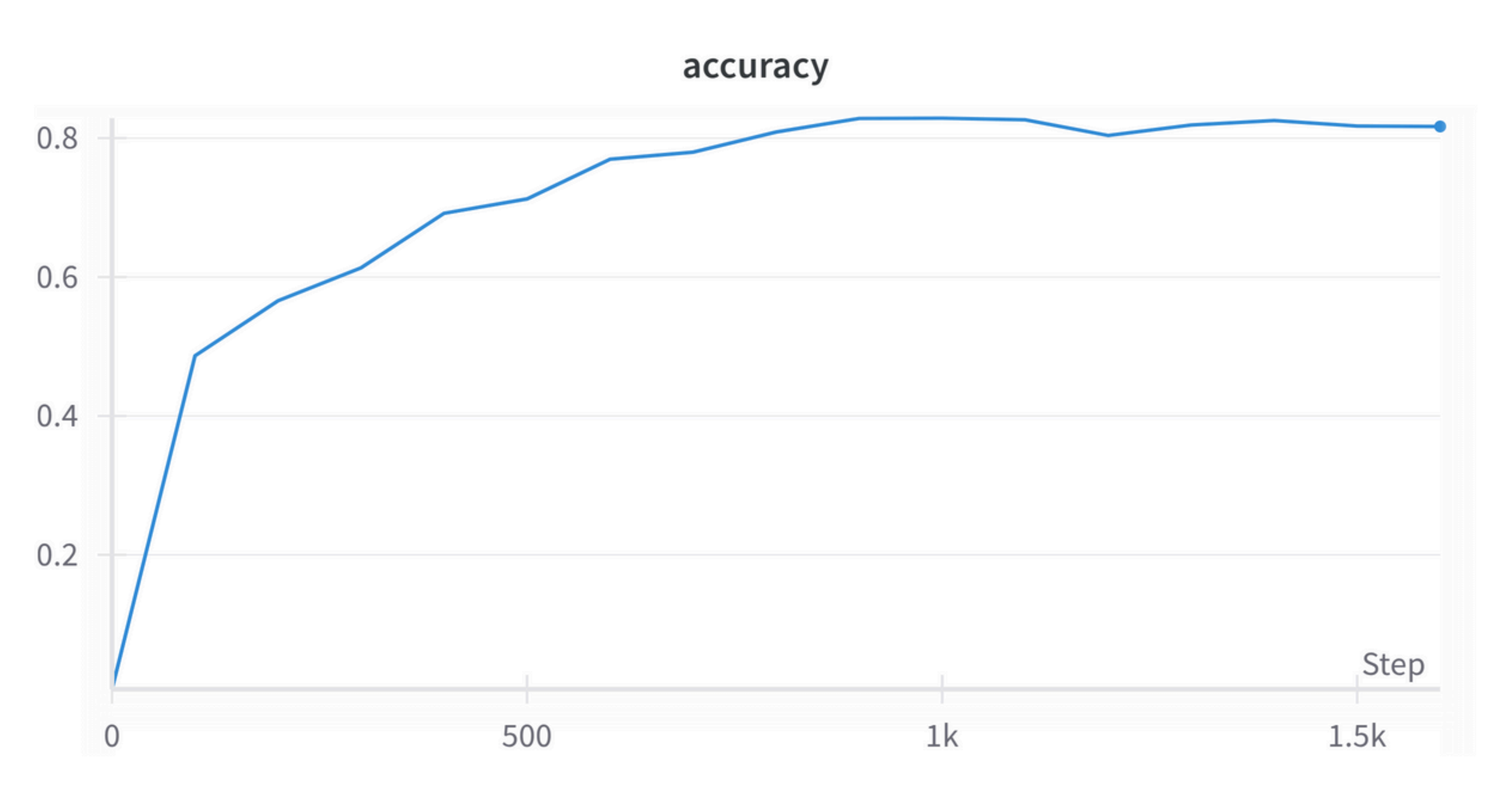
Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks

Wenhui Wang*, Hangbo Bao*, Li Dong*, Johan Bjorck, Zhiliang Peng, Qiang Liu
Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, Furu Wei†
Microsoft Corporation
https://aka.ms/beit-3



beit3_base_patch16_224 (VQAv2 finetuned)





Thanks for watching