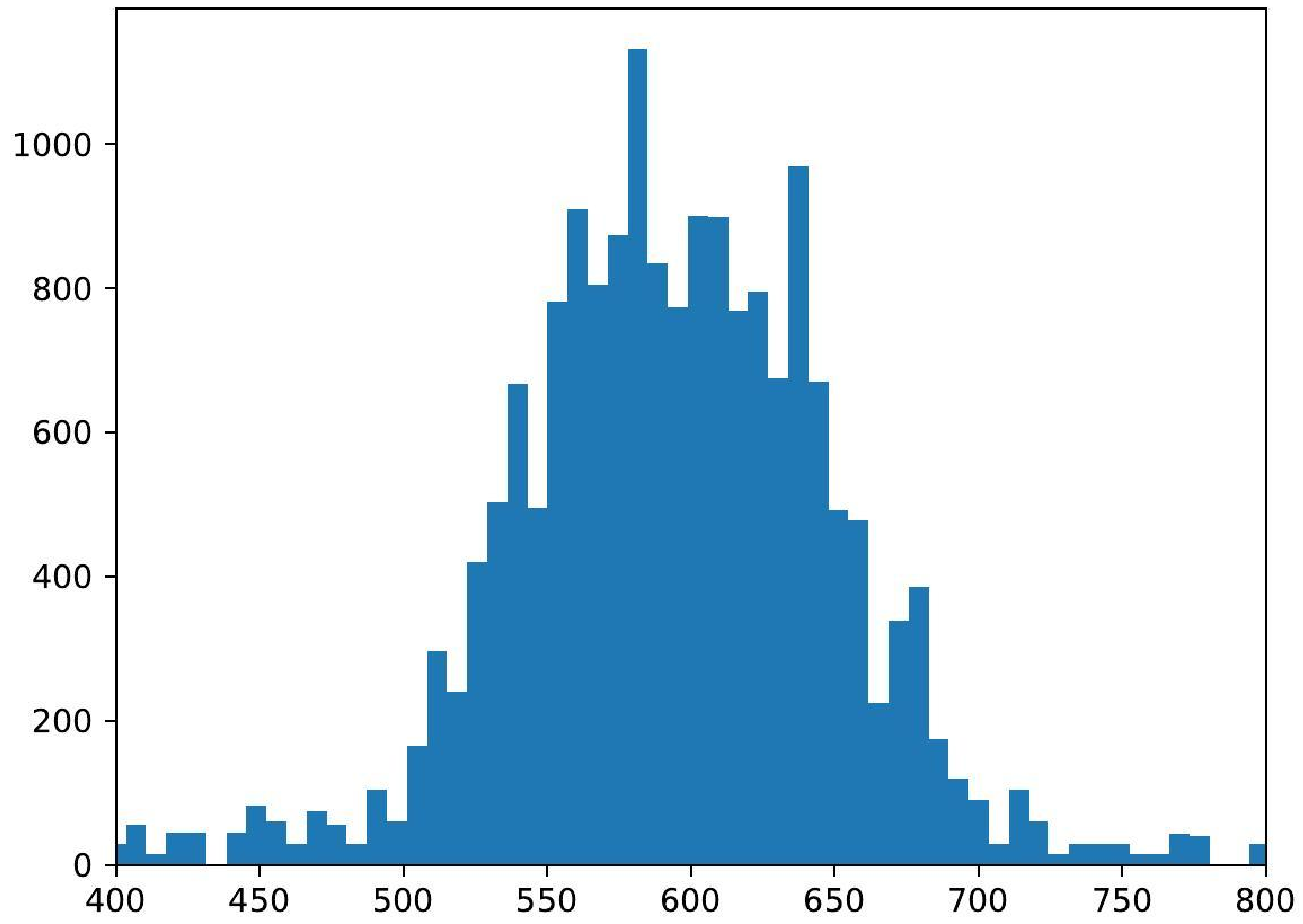

Figure 10: The context length distribution of test samples: Most samples are longer than 500 tokens.

~~~
whether the response (1) follows the given instruction and (2) is
correct. If both answers correctly respond to the prompt, you should
judge it as a tie.

Example 1:
```
Text: We report the development of GPT-4, a large-scale, multimodal
model which can accept image and text inputs and produce text outputs.
While less capable than humans in many real-world scenarios, GPT-4
exhibits human-level performance on various professional and academic
benchmarks, including passing a simulated bar exam with a score around
the top 10% of test takers. GPT-4 is a Transformerbased model
pre-trained to predict the next token in a document. The post-training
alignment process results in improved performance on measures of
factuality and adherence to desired behavior. A core component of this
project was developing infrastructure and optimization methods that
behave predictably across a wide range of scales. This allowed us to
accurately predict some aspects of GPT-4's performance based on models
trained with no more than 1/1,000th the compute of GPT-4.
Prompt: What is GPT4?
Assistant A: GPT4 is a large-scale language-trained transformer-based
model.
Assistant B: GPT4 can produce outputs.
```

Your output should be:
```
{"reason": "The instruction asks what GPT4 is, and from the original
text, we know that GPT4 is a multimodal, large-scale model that can
generate text. Therefore, Assistant A is the closer answer, while
Assistant B did not follow the instruction well in providing a
response.", "choice": "A"}
```

Example 2:
```
Text: Making language models bigger does not inherently make them better
at following a user's intent. For example, large language models can
generate outputs that are untruthful, toxic, or simply not helpful to
the user. In other words, these models are not aligned with their users.
~~~