01_Models_and_Benchmarks

May 9, 2025

1 Medical Model Optimization + Mixture of Experts

1.1 Notebook 1: Model Selection & Benchmarking

Authors:
- Dan Harvey
- Xinzhuo Jiang

Affiliation:
High-Performance Machine Learning (HPML)
Columbia University

1.1.1 Project Overview

This project investigates how to optimize large language models (LLMs) for medical applications by leveraging modern efficiency techniques and modular model design.

1.1.2 Objectives

1. **Benchmark** multiple medical and general-purpose LLMs to establish quantitative performance baselines.

2. **Optimize** selected models through quantization, pruning, and architectural tuning.

3. **Design and evaluate** a Mixture-of-Experts (MoE) architecture with specialized, task-specific experts.

1.1.3 This notebook focuses on Step 1: loading and benchmarking key candidate models.

[1]:
```
##  Environment Setup: Dependencies and Imports
import os
import sys
import importlib
import subprocess
import torch
import platform
```

```
import time
```

[2]:
```
# Add project root to path
project_root = os.path.abspath(os.path.join(os.getcwd(), '..'))
if project_root not in sys.path:
    sys.path.append(project_root)

# Required packages
required_packages = [
    'torch', 'transformers', 'datasets', 'accelerate', 'flash_attn',
    'evaluate', 'lm_eval', 'sklearn', 'matplotlib', 'wandb',
    'tqdm', 'sentencepiece', 'scipy', 'einops'
]

# Check and install missing packages
for package in required_packages:
    try:
        module = importlib.import_module(package)
        print(f" {package} installed successfully")
        if package == 'torch':
            print(f"   Version: {torch.__version__}")
            print(f"   CUDA available: {torch.cuda.is_available()}")
            if torch.cuda.is_available():
                print(f"   CUDA version: {torch.version.cuda}")
                print(f"   GPU: {torch.cuda.get_device_name(0)}")
        elif hasattr(module, '__version__'):
            print(f"   Version: {module.__version__}")
    except ImportError:
        print(f" {package} not found. Installing...")
        subprocess.check_call([sys.executable, "-m", "pip", "install", package])
        module = importlib.import_module(package)
        print(f" {package} installed successfully (post-install)")
        if hasattr(module, '__version__'):
            print(f"   Version: {module.__version__}")

# You may need to restart the Kernel to use these
```

```
torch installed successfully
  Version: 2.6.0+cu124
  CUDA available: True
  CUDA version: 12.4
  GPU: NVIDIA A100-SXM4-40GB
transformers installed successfully
  Version: 4.51.3
datasets installed successfully
  Version: 3.6.0
accelerate installed successfully
```

```
  Version: 1.6.0
flash_attn installed successfully
  Version: 2.7.4.post1
evaluate installed successfully
  Version: 0.4.3
lm_eval installed successfully
sklearn installed successfully
  Version: 1.6.1
matplotlib installed successfully
  Version: 3.10.0
wandb installed successfully
  Version: 0.19.10
tqdm installed successfully
  Version: 4.67.1
sentencepiece installed successfully
  Version: 0.2.0
scipy installed successfully
  Version: 1.15.2
einops installed successfully
  Version: 0.8.1
```

[3]:
```python
# Load section dependencies
from transformers import AutoTokenizer, AutoModelForCausalLM
import gc
import lm_eval
```

2 Model Selection: Baseline Models

We will work with the following Hugging Face models:

Model Name	Size	Notes
`TsinghuaC3I/Llama-3-8B-UltraMedical`	8B	Medical domain-specific, fine-tuned, ideal teacher & benchmark
`meta-llama/Llama-3.2-3B`	3B	Same architecture, smaller, ideal as an expert or student
`Qwen/Qwen3-4B`	4B	Non-LLaMA expert for diversity in MoE

These models will serve as the baseline in our pipeline and will be evaluated for:

- Performance on medical QA and reasoning tasks
- Suitability for distillation and expert specialization
- Impact of downstream optimizations (quantization, pruning, MoE routing)

Note: All models are initially loaded in **full FP32 (float32) precision** to serve as accurate performance baselines before applying any quantization or memory optimization techniques.

[6]:
```
#  Hugging Face Access - Llama is Gated
!huggingface-cli login
```

```
    _|    _|  _|    _|    _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|
_|_|_|_|    _|_|      _|_|_|  _|_|_|_|
    _|    _|  _|    _|  _|        _|          _|    _|_|    _|  _|           _|
_|    _|  _|        _|
    _|_|_|_|  _|    _|  _|  _|_|  _|  _|_|    _|    _|  _|  _|  _|  _|_|
_|_|_|    _|_|_|_|  _|        _|_|_|
    _|    _|  _|    _|  _|    _|  _|    _|    _|    _|    _|_|  _|    _|      _|
_|    _|  _|        _|
    _|    _|    _|_|      _|_|_|    _|_|_|  _|_|_|  _|      _|    _|_|_|      _|
_|    _|    _|_|_|  _|_|_|_|

    To log in, `huggingface_hub` requires a token generated from
https://huggingface.co/settings/tokens .
Enter your token (input will not be visible):
Add token as git credential? (Y/n) n
Token is valid (permission: read).
The token `helm` has been saved to /root/.cache/huggingface/stored_tokens
Your token has been saved to /root/.cache/huggingface/token
Login successful.
The current active token is: `helm`
```

3 Load Baseline Models

3.1 Llama-3-8B-UltraMedical

Links

- Hugging Face Model Card - Paper / Source

Approximate GPU Memory Requirements: - **FP32**: ~32.4 GB
- **FP16**: ~48 GB
- **INT8**: ~24 GB
- **INT4**: ~12 GB

> These values are estimates and may vary based on sequence length, attention optimizations, and tokenizer overhead.

[ ]:
```
#Llama-3-8B-UltraMedical

tokenizer_llama8b_med = AutoTokenizer.from_pretrained(
    "TsinghuaC3I/Llama-3-8B-UltraMedical",
    trust_remote_code=True,
    use_auth_token=True
)
```

```
model_llama8b_med = AutoModelForCausalLM.from_pretrained(
    "TsinghuaC3I/Llama-3-8B-UltraMedical",
    trust_remote_code=True,
  device_map="auto",
    torch_dtype=torch.float32,
    use_auth_token=True
)

print(" Loaded Llama-3-8B-UltraMedical (FP32, device-mapped)")
```

```
/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/tokenization_auto.py:898: FutureWarning: The
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/auto_factory.py:476: FutureWarning: The
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
```

```
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
```

```
WARNING:accelerate.big_modeling:Some parameters are on the meta device because
they were offloaded to the cpu.
```

```
 Loaded Llama-3-8B-UltraMedical (FP32, device-mapped)
```

[ ]:
```
# Inspect your GPU's memory usage
!nvidia-smi
```

```
Wed May  7 02:54:17 2025
+-----------------------------------------------------------------------------
----------+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version:
12.4     |
|-----------------------------------------+------------------------+------------
----------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util
Compute M. |
|                                         |                        |
MIG M. |
|=========================================+========================+============
==========|
|   0  NVIDIA A100-SXM4-40GB          Off |   00000000:00:04.0 Off |
0 |
| N/A   35C    P0             50W /  400W |   37641MiB /  40960MiB |      0%
```

```
Default |
|                                         |                        |
Disabled |
+-----------------------------------------+------------------------+-----------
----------+

+-----------------------------------------------------------------------------
----------+
| Processes:
|
|  GPU   GI   CI        PID   Type   Process name
GPU Memory |
|        ID   ID
Usage      |
|=============================================================================
==========|
+-----------------------------------------------------------------------------
----------+
```

This model took 33139MiB / 40960MiB, or 32.4GB of GPU Memory.

[ ]:
```python
# Offload Model for GPU Space
del model_llama8b_med
del tokenizer_llama8b_med

gc.collect()
torch.cuda.empty_cache()
time.sleep(5)
gc.collect()
torch.cuda.empty_cache()
```

3.2 Llama-3.2-3B

Links

- Hugging Face Model Card
- Paper / Source

Approximate GPU Memory Requirements: - **FP32**: ~14.9 GB

> These are rough estimates. Actual usage depends on sequence length, architecture-specific memory optimizations, and tokenizer overhead.

[ ]:
```python
#Llama-3.2-3B

tokenizer_llama3b = AutoTokenizer.from_pretrained(
    "meta-llama/Llama-3.2-3B",
    trust_remote_code=True,
    use_auth_token=True
)
```

```
model_llama3b = AutoModelForCausalLM.from_pretrained(
    "meta-llama/Llama-3.2-3B",
    trust_remote_code=True,
    device_map="auto",
    torch_dtype=torch.float32,
    use_auth_token=True
)

print(" Loaded Llama-3.2-3B (FP32, device-mapped)")
```

```
/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/tokenization_auto.py:898: FutureWarning: The
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
tokenizer_config.json:   0%|          | 0.00/50.5k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/9.09M [00:00<?, ?B/s]
special_tokens_map.json:   0%|          | 0.00/301 [00:00<?, ?B/s]
/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/auto_factory.py:476: FutureWarning: The
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
config.json:   0%|          | 0.00/844 [00:00<?, ?B/s]
model.safetensors.index.json:   0%|          | 0.00/20.9k [00:00<?, ?B/s]
Fetching 2 files:   0%|          | 0/2 [00:00<?, ?it/s]
model-00002-of-00002.safetensors:   0%|          | 0.00/1.46G [00:00<?, ?B/s]
model-00001-of-00002.safetensors:   0%|          | 0.00/4.97G [00:00<?, ?B/s]
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
generation_config.json:   0%|          | 0.00/185 [00:00<?, ?B/s]
 Loaded Llama-3.2-3B (FP32, device-mapped)
```

[ ]:
```
# Inspect your GPU's memory usage
print("\n--- NVIDIA-SMI Snapshot ---")
print(subprocess.getoutput("nvidia-smi"))
```

```
--- NVIDIA-SMI Snapshot ---
Wed May  7 02:56:29 2025
+-----------------------------------------------------------------------------
----------+
```

```
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version:
12.4     |
|-----------------------------------------+------------------------+------------
----------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util
Compute M. |
|                                         |                        |
MIG M. |
|=========================================+========================+============
==========|
|   0  NVIDIA A100-SXM4-40GB          Off |   00000000:00:04.0 Off |
0 |
| N/A   35C    P0             50W /  400W |   14261MiB /  40960MiB |      0%
Default |
|                                         |                        |
Disabled |
+-----------------------------------------+------------------------+------------
----------+

+---------------------------------------------------------------------------------
----------+
| Processes:
|
|  GPU   GI   CI        PID   Type   Process name
GPU Memory |
|        ID   ID
Usage      |
|=================================================================================
==========|
+---------------------------------------------------------------------------------
----------+
```

This model took 14261MiB / 40960MiB, or 14.9GB of GPU Memory.

[ ]:
```python
# Offload Model for GPU Space
del model_llama3b
del tokenizer_llama3b
gc.collect()
torch.cuda.empty_cache()
time.sleep(5)
gc.collect()
torch.cuda.empty_cache()
```

3.3 Qwen3-4B

Links

- Hugging Face Model Card
- Paper / Source *(Qwen2 paper for reference — Qwen3 paper may be pending)*

Approximate GPU Memory Requirements: - FP32: ~16.9 GB

> Qwen models typically require `trust_remote_code=True` due to custom model implementations.

[ ]:
```
# Qwen3-4B

tokenizer_qwen4b = AutoTokenizer.from_pretrained(
    "Qwen/Qwen3-4B",
    trust_remote_code=True,
    use_auth_token=True
)

model_qwen4b = AutoModelForCausalLM.from_pretrained(
    "Qwen/Qwen3-4B",
    trust_remote_code=True,
    device_map="auto",
    torch_dtype=torch.float32,
    use_auth_token=True
)

print("  Loaded Qwen3-4B (FP32, device-mapped)")
```

```
tokenizer_config.json:   0%|          | 0.00/9.68k [00:00<?, ?B/s]
vocab.json:   0%|          | 0.00/2.78M [00:00<?, ?B/s]
merges.txt:   0%|          | 0.00/1.67M [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/11.4M [00:00<?, ?B/s]
config.json:   0%|          | 0.00/726 [00:00<?, ?B/s]
model.safetensors.index.json:   0%|          | 0.00/32.8k [00:00<?, ?B/s]
Fetching 3 files:   0%|          | 0/3 [00:00<?, ?it/s]
model-00001-of-00003.safetensors:   0%|          | 0.00/3.96G [00:00<?, ?B/s]
model-00002-of-00003.safetensors:   0%|          | 0.00/3.99G [00:00<?, ?B/s]
model-00003-of-00003.safetensors:   0%|          | 0.00/99.6M [00:00<?, ?B/s]
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
generation_config.json:   0%|          | 0.00/239 [00:00<?, ?B/s]
  Loaded Qwen3-4B (FP32, device-mapped)
```

[ ]:
```
# Inspect your GPU's memory usage
print("\n--- NVIDIA-SMI Snapshot ---")
print(subprocess.getoutput("nvidia-smi"))
```

```
--- NVIDIA-SMI Snapshot ---
Wed May  7 02:57:30 2025
+-----------------------------------------------------------------------------
----------+
| NVIDIA-SMI 550.54.15              Driver Version: 550.54.15      CUDA Version:
12.4     |
|-----------------------------------------+------------------------+-----------
----------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile
Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util
Compute M. |
|                                         |                        |
MIG M. |
|=========================================+========================+===========
==========|
|   0  NVIDIA A100-SXM4-40GB          Off |   00000000:00:04.0 Off |
0 |
| N/A   35C    P0             50W /  400W |   17331MiB /  40960MiB |      0%
Default |
|                                         |                        |
Disabled |
+-----------------------------------------+------------------------+-----------
----------+

+-----------------------------------------------------------------------------
----------+
| Processes:
|
|  GPU   GI   CI        PID   Type   Process name
GPU Memory |
|        ID   ID
Usage      |
|=============================================================================
==========|
+-----------------------------------------------------------------------------
----------+
```

** This model took 17331MiB / 40960MiB or ~16.9GB of GPU Memory **

[ ]:
```
# Offload Model for GPU Space
del model_qwen4b
del tokenizer_qwen4b
```

```
gc.collect()
torch.cuda.empty_cache()
time.sleep(5)
gc.collect()
torch.cuda.empty_cache()
```

4 Benchmarking

To establish performance baselines, we will:

- Load eeach model in full float32 (Already implemented above)
- Run each model through standard medical QA tasks (e.gPubMedQA).
- Repeat each benchmark 3 times and average results.

[ ]:
```
# Import section dependencies
import platform
import psutil
import distro
import numpy as np
```

[ ]:
```
!wandb login
```

```
wandb: WARNING Using legacy-service, which is deprecated.
If this is unintentional, you can fix it by ensuring you do not call
`wandb.require('legacy-service')` and do not set the
WANDB_X_REQUIRE_LEGACY_SERVICE environment variable.
wandb: Logging into wandb.ai. (Learn how to deploy a W&B server
locally: https://wandb.me/wandb-server)
wandb: You can find your API key in your browser here:
https://wandb.ai/authorize?ref=models
wandb: Paste an API key from your profile and hit enter, or press
ctrl+c to quit:
wandb: No netrc file found, creating one.
wandb: Appending key for api.wandb.ai to your netrc file:
/root/.netrc
wandb: Currently logged in as: dyh2111 (med-moe)
to https://api.wandb.ai. Use `wandb login --relogin` to force
relogin
```

[ ]:
```
# ==========================
#   System & OS Information
# ==========================
system_info = platform.uname()

print("  System Information")
print("-" * 40)
print(f"Node Name      : {system_info.node}")
```

```
print(f"System          : {platform.system()}")
print(f"OS Flavor       : {distro.name()}")
print(f"OS Version      : {distro.version()}")
print(f"Release         : {system_info.release}")
print(f"Architecture    : {platform.machine()}")
print(f"Python Version  : {platform.python_version()}")

# =====================
#   CPU Information
# =====================
cpu_count = psutil.cpu_count(logical=False)
logical_cpu_count = psutil.cpu_count(logical=True)

print("\n CPU Information")
print("-" * 40)
print(f"Processor       : {system_info.processor or platform.processor()}")
print(f"Physical Cores  : {cpu_count}")
print(f"Logical Cores   : {logical_cpu_count}")

# =====================
#   Memory Information
# =====================
memory_info = psutil.virtual_memory()

print("\n Memory Information")
print("-" * 40)
print(f"Total RAM       : {memory_info.total / 1024 ** 3:.2f} GB")
print(f"Available RAM   : {memory_info.available / 1024 ** 3:.2f} GB")
print(f"Used RAM        : {memory_info.used / 1024 ** 3:.2f} GB")

# =====================
#   Disk Information
# =====================
disk_info = psutil.disk_usage('/')

print("\n Disk Information")
print("-" * 40)
print(f"Total Space     : {disk_info.total / 1024 ** 3:.2f} GB")
print(f"Used Space      : {disk_info.used / 1024 ** 3:.2f} GB")
print(f"Free Space      : {disk_info.free / 1024 ** 3:.2f} GB")

# =====================
#   GPU Information
# =====================

print("\n GPU Info")
print("GPU:", torch.cuda.get_device_name(0))
```

```
print("CUDA Available:", True)
```

```
 System Information
----------------------------------------
Node Name      : 5d6c33d010a6
System         : Linux
OS Flavor      : Ubuntu
OS Version     : 22.04
Release        : 6.1.123+
Architecture   : x86_64
Python Version : 3.11.12

 CPU Information
----------------------------------------
Processor      : x86_64
Physical Cores : 6
Logical Cores  : 12

 Memory Information
----------------------------------------
Total RAM      : 83.48 GB
Available RAM  : 79.70 GB
Used RAM       : 2.89 GB

 Disk Information
----------------------------------------
Total Space    : 235.68 GB
Used Space     : 70.45 GB
Free Space     : 165.21 GB

 GPU Info
GPU: NVIDIA A100-SXM4-40GB
CUDA Available: True
```

4.1 Llama-3-8B-UltraMedical

4.2 Measure baseline performance

[ ]:
```
#Load Llama-3-8B-UltraMedical
model_name = "TsinghuaC3I/Llama-3-8B-UltraMedical"

# Load tokenizer once (doesn't affect model loading time)
tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True,
    use_auth_token=True
)
```

```
load_times = []

trials = 5

print(f"  Starting timed model loads ({trials} repetitions)...\n")

for i in range(trials):
    start_time = time.monotonic()

    model = AutoModelForCausalLM.from_pretrained(
        model_name,
        trust_remote_code=True,
        device_map="auto",
        torch_dtype=torch.float32,
        use_auth_token=True
    )

    elapsed = time.monotonic() - start_time
    load_times.append(elapsed)
    print(f"  Run {i + 1}: Loaded in {elapsed:.2f} seconds")

    # Clean up between runs (free GPU memory)
    del model
    gc.collect()
    torch.cuda.empty_cache()
    time.sleep(5)
    gc.collect()
    torch.cuda.empty_cache()

# Summary stats
mean_time = np.mean(load_times)
std_dev_time = np.std(load_times)

print(f"\n {model_name} Load Time Summary (FP32)")
print(f"- Average Load Time: {mean_time:.2f} seconds")
print(f"- Std Dev:           {std_dev_time:.2f} seconds")
```

```
/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/tokenization_auto.py:898: FutureWarning: The
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(

 Starting timed model loads (5 repetitions)…


/usr/local/lib/python3.11/dist-
packages/transformers/models/auto/auto_factory.py:476: FutureWarning: The
```

```
`use_auth_token` argument is deprecated and will be removed in v5 of
Transformers. Please use `token` instead.
  warnings.warn(
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
 Run 1: Loaded in 5.21 seconds
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
 Run 2: Loaded in 5.11 seconds
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
 Run 3: Loaded in 5.11 seconds
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
 Run 4: Loaded in 5.10 seconds
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
 Run 5: Loaded in 5.10 seconds

 TsinghuaC3I/Llama-3-8B-UltraMedical Load Time Summary (FP32)
- Average Load Time: 5.13 seconds
- Std Dev:           0.04 seconds
```

4.3 Llama-3.2-3B

[ ]:
```python
#Load Llama-3.2 3B

model_name = "meta-llama/Llama-3.2-3B"

# Load tokenizer once (doesn't affect model loading time)
tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True,
    use_auth_token=True
)

load_times = []

trials = 5

print(f" Starting timed model loads ({trials} repetitions)...\n")

for i in range(trials):
    start_time = time.monotonic()

    model = AutoModelForCausalLM.from_pretrained(
        model_name,
```

```
        trust_remote_code=True,
        device_map="auto",
        torch_dtype=torch.float32,
        use_auth_token=True
    )

    elapsed = time.monotonic() - start_time
    load_times.append(elapsed)
    print(f"  Run {i + 1}: Loaded in {elapsed:.2f} seconds")

    # Clean up between runs (free GPU memory)
    del model
    gc.collect()
    torch.cuda.empty_cache()
    time.sleep(5)
    gc.collect()
    torch.cuda.empty_cache()

# Summary stats
mean_time = np.mean(load_times)
std_dev_time = np.std(load_times)

print(f"\n {model_name} Load Time Summary (FP32)")
print(f"- Average Load Time: {mean_time:.2f} seconds")
print(f"- Std Dev:           {std_dev_time:.2f} seconds")
```

```
 Starting timed model loads (5 repetitions)…

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
 Run 1: Loaded in 2.68 seconds
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
 Run 2: Loaded in 2.67 seconds
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
 Run 3: Loaded in 2.45 seconds
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
 Run 4: Loaded in 2.45 seconds
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
 Run 5: Loaded in 2.46 seconds

 meta-llama/Llama-3.2-3B Load Time Summary (FP32)
- Average Load Time: 2.54 seconds
- Std Dev:           0.11 seconds
```

[ ]:
```python
import random
import json
import wandb
import subprocess
import time
import os
from datetime import datetime

# ----------------------------
#  Model and Task Config
# ----------------------------
model_name = "meta-llama/Llama-3.2-3B"
task_name = "pubmedqa"
output_base = "./results"

# ----------------------------
#  Start W&B run
# ----------------------------
run_name = f"{model_name.replace('/', '_')}_{task_name}_5x"
wandb_run = wandb.init(
    project="med-moe-baseline-evals",
    name=run_name,
    config={
        "model": model_name,
        "task": task_name,
        "batch_size": 8,
        "precision": "fp32",
        "eval_method": "lm_eval",
        "repeats": 5
    }
)

# ----------------------------
#  Run 5x Evaluation Loop
# ----------------------------
for i in range(5):
    print(f"\n Run {i + 1}/5")

    # Create timestamped output folder
    timestamp = datetime.now().strftime("%Y-%m-%dT%H-%M-%S")
    run_output_dir = os.path.join(output_base, f"run_{i+1}_{timestamp}")
    os.makedirs(run_output_dir, exist_ok=True)

    # Define lm_eval command
    command = [
        "lm_eval",
        "--model", "hf",
```

```
        "--tasks", task_name,
        "--model_args", f"pretrained={model_name},parallelize=True",
        "--device", "cuda:0",
        "--batch_size", "8",
        "--write_out",
        "--output_path", run_output_dir,
        "--trust_remote_code",
        "--confirm_run_unsafe_code"
    ]

    # Start timing
    start_time = time.monotonic()
    result = subprocess.run(command, capture_output=True, text=True)
    elapsed = time.monotonic() - start_time

    print(f"  Run {i + 1} completed in {elapsed:.2f} seconds")
    print("STDOUT:\n", result.stdout)

    # ----------------------------
    #   Find and parse result file
    # ----------------------------
    result_file = None
    for fname in os.listdir(run_output_dir):
        if fname.startswith("eval_results") and fname.endswith(".json"):
            result_file = os.path.join(run_output_dir, fname)
            break

    if result_file is None:
        print(f"  No eval_results_*.json found in {run_output_dir}")
        continue

    try:
        with open(result_file) as f:
            data = json.load(f)
        task_data = data["results"][task_name]

        acc = task_data.get("acc,none")
        stderr = task_data.get("acc_stderr,none")

        if acc is not None and stderr is not None:
            wandb_run.log({
                f"{task_name}/accuracy": acc,
                f"{task_name}/stddev": stderr,
                f"{task_name}/eval_time_sec": elapsed,
                "run_index": i + 1
            })
            print(f"  Logged to W&B: acc={acc:.3f}, stderr={stderr:.4f}")
```

```
        else:
            print(f"  Missing keys in result: {task_data.keys()}")

    except Exception as e:
        print(f"  Failed to parse results from {result_file}: {e}")

# ------------------------------
#   Finish W&B run
# ------------------------------
wandb_run.finish()
```

```
<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>


  Run 1/5
  Run 1 completed in 30.88 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


  No eval_results_*.json found in ./results/run_1_2025-05-07T03-52-27

  Run 2/5
  Run 2 completed in 30.71 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


  No eval_results_*.json found in ./results/run_2_2025-05-07T03-52-58

  Run 3/5
  Run 3 completed in 30.65 seconds
```

```
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


  No eval_results_*.json found in ./results/run_3_2025-05-07T03-53-28

  Run 4/5
  Run 4 completed in 30.57 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


  No eval_results_*.json found in ./results/run_4_2025-05-07T03-53-59

  Run 5/5
  Run 5 completed in 30.52 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


  No eval_results_*.json found in ./results/run_5_2025-05-07T03-54-29
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>
```

4.4 Qwen3-4B

[ ]:
```
#Load Quen3 4B

model_name = "Qwen/Qwen3-4B"
```

```python
# Load tokenizer once (doesn't affect model loading time)
tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True,
    use_auth_token=True
)

load_times = []

trials = 5

print(f" Starting timed model loads ({trials} repetitions)...\n")

for i in range(trials):
    start_time = time.monotonic()

    model = AutoModelForCausalLM.from_pretrained(
        model_name,
        trust_remote_code=True,
        device_map="auto",
        torch_dtype=torch.float32,
        use_auth_token=True
    )

    elapsed = time.monotonic() - start_time
    load_times.append(elapsed)
    print(f" Run {i + 1}: Loaded in {elapsed:.2f} seconds")

    # Clean up between runs (free GPU memory)
    del model
    gc.collect()
    torch.cuda.empty_cache()
    time.sleep(5)
    gc.collect()
    torch.cuda.empty_cache()

# Summary stats
mean_time = np.mean(load_times)
std_dev_time = np.std(load_times)

print(f"\n {model_name} Load Time Summary (FP32)")
print(f"- Average Load Time: {mean_time:.2f} seconds")
print(f"- Std Dev:           {std_dev_time:.2f} seconds")
```

```
 Starting timed model loads (5 repetitions)…

Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
```

```
  Run 1: Loaded in 3.27 seconds
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
  Run 2: Loaded in 3.02 seconds
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
  Run 3: Loaded in 3.01 seconds
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
  Run 4: Loaded in 3.02 seconds
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
  Run 5: Loaded in 3.12 seconds

  Qwen/Qwen3-4B Load Time Summary (FP32)
- Average Load Time: 3.09 seconds
- Std Dev:           0.10 seconds
```

[8]:
```python
import random
import json
import wandb
import subprocess
import time
import os
from datetime import datetime

# ------------------------------
#   Model and Task Config
# ------------------------------
model_name = "Qwen/Qwen3-4B"
task_name = "pubmedqa"
output_base = "./results"

# ------------------------------
#   Start W&B run
# ------------------------------
run_name = f"{model_name.replace('/', '_')}_{task_name}_5x"
wandb_run = wandb.init(
    project="med-moe-baseline-evals",
    name=run_name,
    config={
        "model": model_name,
        "task": task_name,
        "batch_size": 8,
        "precision": "fp32",
        "eval_method": "lm_eval",
        "repeats": 5
```

```
    }
)

# ----------------------------
#  Run 5x Evaluation Loop
# ----------------------------
for i in range(5):
    print(f"\n Run {i + 1}/5")

    # Create timestamped output folder
    timestamp = datetime.now().strftime("%Y-%m-%dT%H-%M-%S")
    run_output_dir = os.path.join(output_base, f"run_{i+1}_{timestamp}")
    os.makedirs(run_output_dir, exist_ok=True)

    # Define lm_eval command
    command = [
        "lm_eval",
        "--model", "hf",
        "--tasks", task_name,
        "--model_args", f"pretrained={model_name},parallelize=True",
        "--device", "cuda:0",
        "--batch_size", "8",
        "--write_out",
        "--output_path", run_output_dir,
        "--trust_remote_code", "--confirm_run_unsafe_code"
    ]

    # Start timing
    start_time = time.monotonic()
    result = subprocess.run(command, capture_output=True, text=True)
    elapsed = time.monotonic() - start_time

    print(f" Run {i + 1} completed in {elapsed:.2f} seconds")
    print("STDOUT:\n", result.stdout)
    print("STDERR:\n", result.stderr)

    # ----------------------------
    #  Find and parse result file
    # ----------------------------
    result_file = None
    for fname in os.listdir(run_output_dir):
        if fname.startswith("eval_results") and fname.endswith(".json"):
            result_file = os.path.join(run_output_dir, fname)
            break

    if result_file is None:
        print(f" No eval_results_*.json found in {run_output_dir}")
```

```
        continue

    try:
        with open(result_file) as f:
            data = json.load(f)
        task_data = data["results"][task_name]

        acc = task_data.get("acc,none")
        stderr = task_data.get("acc_stderr,none")

        if acc is not None and stderr is not None:
            wandb_run.log({
                f"{task_name}/accuracy": acc,
                f"{task_name}/stddev": stderr,
                f"{task_name}/eval_time_sec": elapsed,
                "run_index": i + 1
            })
            print(f"  Logged to W&B: acc={acc:.3f}, stderr={stderr:.4f}")
        else:
            print(f"  Missing keys in result: {task_data.keys()}")

    except Exception as e:
        print(f"  Failed to parse results from {result_file}: {e}")

# ----------------------------
#   Finish W&B run
# ----------------------------
wandb_run.finish()
```

```
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>
<IPython.core.display.HTML object>


  Run 1/5
  Run 1 completed in 43.17 seconds
STDOUT:
 hf (pretrained=Qwen/Qwen3-4B,parallelize=True,trust_remote_code=True),
gen_kwargs: (None), limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.768|±  |0.0189|
```

```
STDERR:
 2025-05-09 01:54:35.138132: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746755675.159982   13896 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746755675.166597   13896 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:01:54:51,923 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:01:54:51,923 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:01:54:51,924 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:01:54:51,924 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'Qwen/Qwen3-4B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:01:54:51,965 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:01:54:52,762 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
Loading checkpoint shards:  33%|        | 1/3 [00:01<00:02,  1.31s/it]
Loading checkpoint shards:  67%|       | 2/3 [00:02<00:01,  1.33s/it]
Loading checkpoint shards: 100%|      | 3/3 [00:02<00:00,  1.12it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:01:54:56,411 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
```

```
100%|      | 500/500 [00:00<00:00, 75126.35it/s]
2025-05-09:01:54:56,472 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:54:56,472 INFO     [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
demonstrate dysfunction of the anal sphincter and/or the musculus (m.)
puborectalis in patients with dyschesia using anorectal endosonography.',
'Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.', 'The anal sphincter became paradoxically shorter and/or
thicker during straining (versus the resting state) in 85% of patients but in
only 35% of control subjects. Changes in sphincter length were statistically
significantly different (p<0.01, chi(2) test) in patients compared with control
subjects. The m. puborectalis became paradoxically shorter and/or thicker during
straining in 80% of patients but in only 30% of controls. Both the changes in
length and thickness of the m. puborectalis were significantly different
(p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS',
'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and
```

over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:54:56,472 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and

thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:54:56,472 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of

control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:54:56,472 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:54:56,472 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or

thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:54:56,472 INFO   [lm_eval.evaluator:517] Running loglikelihood requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<19:30, 1.28it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:01<00:48, 30.31it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:29, 49.91it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:21, 64.93it/s]

```
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:18,
75.70it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:02<00:16,
84.40it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:02<00:14,
90.99it/s]
Running loglikelihood requests:  11%|         | 169/1500 [00:02<00:13,
96.05it/s]
Running loglikelihood requests:  13%|         | 193/1500 [00:02<00:13,
99.86it/s]
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:12,
103.59it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:03<00:11,
106.51it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:03<00:11,
109.63it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:03<00:10,
112.10it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:03<00:10,
114.54it/s]
Running loglikelihood requests:  22%|         | 337/1500 [00:03<00:09,
116.59it/s]
Running loglikelihood requests:  24%|         | 361/1500 [00:04<00:09,
118.47it/s]
Running loglikelihood requests:  26%|         | 385/1500 [00:04<00:09,
120.08it/s]
Running loglikelihood requests:  27%|         | 409/1500 [00:04<00:08,
122.02it/s]
Running loglikelihood requests:  29%|         | 433/1500 [00:04<00:08,
123.83it/s]
Running loglikelihood requests:  30%|         | 457/1500 [00:04<00:08,
125.21it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:05<00:08,
126.60it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:05<00:07,
127.78it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:05<00:07,
128.96it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:05<00:07,
129.95it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:05<00:07,
130.95it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:05<00:06,
131.75it/s]
Running loglikelihood requests:  42%|        | 625/1500 [00:06<00:06,
132.52it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:06<00:06,
132.96it/s]
```

```
Running loglikelihood requests:  45%|       | 673/1500 [00:06<00:06,
133.41it/s]
Running loglikelihood requests:  46%|       | 697/1500 [00:06<00:05,
135.02it/s]
Running loglikelihood requests:  48%|       | 721/1500 [00:06<00:05,
136.74it/s]
Running loglikelihood requests:  50%|       | 745/1500 [00:06<00:05,
137.95it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:07<00:05,
141.48it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:07<00:04,
144.24it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:07<00:04,
146.16it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:07<00:04,
147.76it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:07<00:04,
150.10it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:07<00:04,
152.37it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:08<00:03,
153.67it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:08<00:03,
154.78it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:08<00:03,
157.92it/s]
Running loglikelihood requests:  66%|       | 985/1500 [00:08<00:03,
160.33it/s]
Running loglikelihood requests:  67%|       | 1009/1500 [00:08<00:03,
162.33it/s]
Running loglikelihood requests:  69%|       | 1033/1500 [00:08<00:02,
163.99it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:08<00:02,
165.20it/s]
Running loglikelihood requests:  72%|      | 1081/1500 [00:09<00:02,
166.44it/s]
Running loglikelihood requests:  74%|      | 1105/1500 [00:09<00:02,
167.81it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:09<00:02,
170.95it/s]
Running loglikelihood requests:  77%|      | 1153/1500 [00:09<00:01,
173.79it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:09<00:01,
176.22it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:09<00:01,
178.37it/s]
Running loglikelihood requests:  82%|     | 1225/1500 [00:09<00:01,
182.42it/s]
```

```
Running loglikelihood requests:  83%|      | 1249/1500 [00:10<00:01,
185.84it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:10<00:01,
188.81it/s]
Running loglikelihood requests:  86%|      | 1297/1500 [00:10<00:01,
190.62it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:10<00:00,
193.70it/s]
Running loglikelihood requests:  90%|      | 1345/1500 [00:10<00:00,
196.33it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:10<00:00,
201.41it/s]
Running loglikelihood requests:  93%|     | 1393/1500 [00:10<00:00,
206.01it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:10<00:00,
214.34it/s]
Running loglikelihood requests:  96%|     | 1446/1500 [00:10<00:00,
235.22it/s]
Running loglikelihood requests:  99%|     | 1489/1500 [00:11<00:00,
251.91it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:11<00:00,
135.42it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:01:55:12,285 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_1_2025-05-09T01-54-30

  Run 2/5
  Run 2 completed in 46.50 seconds
STDOUT:
 hf (pretrained=Qwen/Qwen3-4B,parallelize=True,trust_remote_code=True),
gen_kwargs: (None), limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.768|±  |0.0189|


STDERR:
 2025-05-09 01:55:18.260852: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746755718.282057   14152 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
```

```
E0000 00:00:1746755718.288523   14152 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:01:55:35,028 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:01:55:35,028 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:01:55:35,029 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:01:55:35,030 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'Qwen/Qwen3-4B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:01:55:35,070 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:01:55:35,815 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
Loading checkpoint shards:  33%|        | 1/3 [00:01<00:02,  1.17s/it]
Loading checkpoint shards:  67%|       | 2/3 [00:02<00:01,  1.18s/it]
Loading checkpoint shards: 100%|     | 3/3 [00:02<00:00,  1.25it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:01:55:39,137 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|     | 500/500 [00:00<00:00, 99598.78it/s]
2025-05-09:01:55:39,197 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
```

models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:55:39,197 INFO   [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation

movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:55:39,197 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:55:39,197 INFO   [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal

endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:55:39,197 INFO   [lm_eval.evaluator_utils:206] Task: Configurabl

eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:55:39,197 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male',

'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002',
'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes',
'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography
demonstrated incomplete or even absent relaxation of the anal sphincter and the
m. puborectalis during a defecation movement in the majority of our patients
with dyschesia. This study highlights the value of this elegant ultrasonographic
technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'},
arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation
movements. The aim of this prospective study was to demonstrate dysfunction of
the anal sphincter and/or the musculus (m.) puborectalis in patients with
dyschesia using anorectal endosonography.\nTwenty consecutive patients with a
medical history of dyschesia and a control group of 20 healthy subjects
underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625
RT). In both groups, the dimensions of the anal sphincter and the m.
puborectalis were measured at rest, and during voluntary squeezing and
straining. Statistical analysis was performed within and between the two
groups.\nThe anal sphincter became paradoxically shorter and/or thicker during
straining (versus the resting state) in 85% of patients but in only 35% of
control subjects. Changes in sphincter length were statistically significantly
different (p<0.01, chi(2) test) in patients compared with control subjects. The
m. puborectalis became paradoxically shorter and/or thicker during straining in
80% of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.\nQuestion: Is anorectal
endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:01:55:39,197 INFO   [lm_eval.evaluator:517] Running loglikelihood
requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<19:50, 1.26it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:01<00:49,
29.97it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:29,
49.55it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:22,
64.59it/s]
Running loglikelihood requests: 6%| | 97/1500 [00:01<00:18,
75.51it/s]
Running loglikelihood requests: 8%| | 121/1500 [00:02<00:16,
84.24it/s]
Running loglikelihood requests: 10%| | 145/1500 [00:02<00:14,
90.87it/s]
Running loglikelihood requests: 11%| | 169/1500 [00:02<00:13,
96.04it/s]
Running loglikelihood requests: 13%| | 193/1500 [00:02<00:13,
99.83it/s]

```
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:12,
103.60it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:03<00:11,
106.47it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:03<00:11,
109.66it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:03<00:10,
112.12it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:03<00:10,
114.58it/s]
Running loglikelihood requests:  22%|        | 337/1500 [00:03<00:09,
116.69it/s]
Running loglikelihood requests:  24%|        | 361/1500 [00:04<00:09,
118.56it/s]
Running loglikelihood requests:  26%|        | 385/1500 [00:04<00:09,
119.93it/s]
Running loglikelihood requests:  27%|        | 409/1500 [00:04<00:08,
121.87it/s]
Running loglikelihood requests:  29%|        | 433/1500 [00:04<00:08,
123.52it/s]
Running loglikelihood requests:  30%|        | 457/1500 [00:04<00:08,
125.16it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:05<00:08,
126.74it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:05<00:07,
127.74it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:05<00:07,
128.36it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:05<00:07,
129.50it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:05<00:07,
130.69it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:05<00:06,
131.43it/s]
Running loglikelihood requests:  42%|       | 625/1500 [00:06<00:06,
132.18it/s]
Running loglikelihood requests:  43%|       | 649/1500 [00:06<00:06,
132.75it/s]
Running loglikelihood requests:  45%|       | 673/1500 [00:06<00:06,
133.24it/s]
Running loglikelihood requests:  46%|       | 697/1500 [00:06<00:05,
135.03it/s]
Running loglikelihood requests:  48%|       | 721/1500 [00:06<00:05,
136.73it/s]
Running loglikelihood requests:  50%|       | 745/1500 [00:07<00:05,
138.02it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:07<00:05,
141.57it/s]
```

```
Running loglikelihood requests:  53%|       | 793/1500 [00:07<00:04,
144.18it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:07<00:04,
146.10it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:07<00:04,
147.68it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:07<00:04,
150.15it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:07<00:04,
152.45it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:08<00:03,
153.72it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:08<00:03,
155.38it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:08<00:03,
158.05it/s]
Running loglikelihood requests:  66%|       | 985/1500 [00:08<00:03,
160.52it/s]
Running loglikelihood requests:  67%|       | 1009/1500 [00:08<00:03,
162.43it/s]
Running loglikelihood requests:  69%|       | 1033/1500 [00:08<00:02,
164.07it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:08<00:02,
165.52it/s]
Running loglikelihood requests:  72%|      | 1081/1500 [00:09<00:02,
166.85it/s]
Running loglikelihood requests:  74%|      | 1105/1500 [00:09<00:02,
168.26it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:09<00:02,
171.60it/s]
Running loglikelihood requests:  77%|      | 1153/1500 [00:09<00:01,
174.24it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:09<00:01,
176.37it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:09<00:01,
178.51it/s]
Running loglikelihood requests:  82%|     | 1225/1500 [00:09<00:01,
182.33it/s]
Running loglikelihood requests:  83%|     | 1249/1500 [00:10<00:01,
185.78it/s]
Running loglikelihood requests:  85%|     | 1273/1500 [00:10<00:01,
188.68it/s]
Running loglikelihood requests:  86%|     | 1297/1500 [00:10<00:01,
191.44it/s]
Running loglikelihood requests:  88%|     | 1321/1500 [00:10<00:00,
194.12it/s]
Running loglikelihood requests:  90%|     | 1345/1500 [00:10<00:00,
196.39it/s]
```

```
Running loglikelihood requests:  91%|      | 1369/1500 [00:10<00:00,
201.21it/s]
Running loglikelihood requests:  93%|      | 1393/1500 [00:10<00:00,
205.24it/s]
Running loglikelihood requests:  94%|      | 1417/1500 [00:10<00:00,
213.66it/s]
Running loglikelihood requests:  96%|      | 1447/1500 [00:10<00:00,
237.36it/s]
Running loglikelihood requests:  99%|      | 1489/1500 [00:11<00:00,
249.55it/s]
Running loglikelihood requests: 100%|      | 1500/1500 [00:11<00:00,
135.23it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:01:55:58,785 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_2_2025-05-09T01-55-14

  Run 3/5
  Run 3 completed in 42.78 seconds
STDOUT:
 hf (pretrained=Qwen/Qwen3-4B,parallelize=True,trust_remote_code=True),
gen_kwargs: (None), limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.768|±  |0.0189|


STDERR:
 2025-05-09 01:56:04.757336: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746755764.778788   14412 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746755764.785360   14412 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:01:56:21,443 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:01:56:21,443 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:01:56:21,444 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
```

```
fewshot manual seed to 1234
2025-05-09:01:56:21,444 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'Qwen/Qwen3-4B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:01:56:21,484 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:01:56:22,297 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
Loading checkpoint shards:  33%|        | 1/3 [00:01<00:02,  1.15s/it]
Loading checkpoint shards:  67%|       | 2/3 [00:02<00:01,  1.17s/it]
Loading checkpoint shards: 100%|      | 3/3 [00:02<00:00,  1.27it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:01:56:25,568 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|      | 500/500 [00:00<00:00, 77522.99it/s]
2025-05-09:01:56:25,627 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
```

thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:56:25,627 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of

control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:56:25,627 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:56:25,627 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or

thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:56:25,627 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary

squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:56:25,627 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with

dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:56:25,627 INFO [lm_eval.evaluator:517] Running loglikelihood requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<19:57, 1.25it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:01<00:49, 29.85it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:29, 49.38it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:22, 64.54it/s]
Running loglikelihood requests: 6%| | 97/1500 [00:01<00:18, 75.49it/s]
Running loglikelihood requests: 8%| | 121/1500 [00:02<00:16, 84.15it/s]
Running loglikelihood requests: 10%| | 145/1500 [00:02<00:14, 90.67it/s]
Running loglikelihood requests: 11%| | 169/1500 [00:02<00:13, 95.91it/s]
Running loglikelihood requests: 13%| | 193/1500 [00:02<00:13, 99.73it/s]
Running loglikelihood requests: 14%| | 217/1500 [00:02<00:12, 103.34it/s]
Running loglikelihood requests: 16%| | 241/1500 [00:03<00:11, 106.32it/s]
Running loglikelihood requests: 18%| | 265/1500 [00:03<00:11, 109.59it/s]
Running loglikelihood requests: 19%| | 289/1500 [00:03<00:10, 112.14it/s]
Running loglikelihood requests: 21%| | 313/1500 [00:03<00:10, 114.66it/s]

```
Running loglikelihood requests:  22%|         | 337/1500 [00:03<00:09,
116.52it/s]
Running loglikelihood requests:  24%|         | 361/1500 [00:04<00:09,
118.30it/s]
Running loglikelihood requests:  26%|         | 385/1500 [00:04<00:09,
119.94it/s]
Running loglikelihood requests:  27%|         | 409/1500 [00:04<00:08,
121.95it/s]
Running loglikelihood requests:  29%|         | 433/1500 [00:04<00:08,
123.77it/s]
Running loglikelihood requests:  30%|         | 457/1500 [00:04<00:08,
125.17it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:05<00:08,
126.62it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:05<00:07,
127.81it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:05<00:07,
128.85it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:05<00:07,
129.70it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:05<00:07,
130.80it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:05<00:06,
131.58it/s]
Running loglikelihood requests:  42%|       | 625/1500 [00:06<00:06,
132.23it/s]
Running loglikelihood requests:  43%|       | 649/1500 [00:06<00:06,
132.74it/s]
Running loglikelihood requests:  45%|       | 673/1500 [00:06<00:06,
133.44it/s]
Running loglikelihood requests:  46%|       | 697/1500 [00:06<00:05,
135.34it/s]
Running loglikelihood requests:  48%|       | 721/1500 [00:06<00:05,
136.81it/s]
Running loglikelihood requests:  50%|       | 745/1500 [00:07<00:05,
138.26it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:07<00:05,
141.39it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:07<00:04,
143.83it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:07<00:04,
146.02it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:07<00:04,
148.08it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:07<00:04,
150.73it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:07<00:04,
152.49it/s]
```

```
Running loglikelihood requests:  61%|       | 913/1500 [00:08<00:03,
154.13it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:08<00:03,
155.68it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:08<00:03,
158.67it/s]
Running loglikelihood requests:  66%|       | 985/1500 [00:08<00:03,
160.64it/s]
Running loglikelihood requests:  67%|       | 1009/1500 [00:08<00:03,
162.24it/s]
Running loglikelihood requests:  69%|       | 1033/1500 [00:08<00:02,
163.83it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:08<00:02,
165.31it/s]
Running loglikelihood requests:  72%|      | 1081/1500 [00:09<00:02,
167.00it/s]
Running loglikelihood requests:  74%|      | 1105/1500 [00:09<00:02,
168.64it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:09<00:02,
171.87it/s]
Running loglikelihood requests:  77%|      | 1153/1500 [00:09<00:01,
174.16it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:09<00:01,
176.68it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:09<00:01,
178.47it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:09<00:01,
182.45it/s]
Running loglikelihood requests:  83%|      | 1249/1500 [00:10<00:01,
185.42it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:10<00:01,
188.16it/s]
Running loglikelihood requests:  86%|      | 1297/1500 [00:10<00:01,
190.89it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:10<00:00,
193.90it/s]
Running loglikelihood requests:  90%|      | 1345/1500 [00:10<00:00,
196.06it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:10<00:00,
200.67it/s]
Running loglikelihood requests:  93%|     | 1393/1500 [00:10<00:00,
205.24it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:10<00:00,
213.25it/s]
Running loglikelihood requests:  96%|     | 1447/1500 [00:10<00:00,
237.01it/s]
Running loglikelihood requests:  99%|     | 1489/1500 [00:11<00:00,
251.09it/s]
```

```
Running loglikelihood requests: 100%|      | 1500/1500 [00:11<00:00,
135.22it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:01:56:41,552 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_3_2025-05-09T01-56-00

  Run 4/5
  Run 4 completed in 42.54 seconds
STDOUT:
 hf (pretrained=Qwen/Qwen3-4B,parallelize=True,trust_remote_code=True),
gen_kwargs: (None), limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.768|±  |0.0189|


STDERR:
 2025-05-09 01:56:47.533722: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746755807.554747   14658 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746755807.561046   14658 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:01:57:04,191 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:01:57:04,191 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:01:57:04,192 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:01:57:04,192 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'Qwen/Qwen3-4B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:01:57:04,232 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:01:57:05,034 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto
```

```
Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
Loading checkpoint shards:  33%|        | 1/3 [00:01<00:02,  1.15s/it]
Loading checkpoint shards:  67%|       | 2/3 [00:02<00:01,  1.17s/it]
Loading checkpoint shards: 100%|     | 3/3 [00:02<00:00,  1.27it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:01:57:08,309 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|     | 500/500 [00:00<00:00, 81043.09it/s]
2025-05-09:01:57:08,367 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:08,367 INFO     [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
```

endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:57:08,367 INFO   [lm_eval.evaluator_utils:206] Task: Configurabl

eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:08,367 INFO   [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male',

'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002',
'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes',
'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography
demonstrated incomplete or even absent relaxation of the anal sphincter and the
m. puborectalis during a defecation movement in the majority of our patients
with dyschesia. This study highlights the value of this elegant ultrasonographic
technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'},
arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation
movements. The aim of this prospective study was to demonstrate dysfunction of
the anal sphincter and/or the musculus (m.) puborectalis in patients with
dyschesia using anorectal endosonography.\nTwenty consecutive patients with a
medical history of dyschesia and a control group of 20 healthy subjects
underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625
RT). In both groups, the dimensions of the anal sphincter and the m.
puborectalis were measured at rest, and during voluntary squeezing and
straining. Statistical analysis was performed within and between the two
groups.\nThe anal sphincter became paradoxically shorter and/or thicker during
straining (versus the resting state) in 85% of patients but in only 35% of
control subjects. Changes in sphincter length were statistically significantly
different (p<0.01, chi(2) test) in patients compared with control subjects. The
m. puborectalis became paradoxically shorter and/or thicker during straining in
80% of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.\nQuestion: Is anorectal
endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:01:57:08,367 INFO   [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.

Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:08,367 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The

```
m. puborectalis became paradoxically shorter and/or thicker during straining in
80% of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.\nQuestion: Is anorectal
endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:01:57:08,368 INFO     [lm_eval.evaluator:517] Running loglikelihood
requests

Running loglikelihood requests:   0%|          | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests:   0%|          | 1/1500 [00:00<19:27,  1.28it/s]
Running loglikelihood requests:   2%|         | 25/1500 [00:01<00:48,
30.35it/s]
Running loglikelihood requests:   3%|         | 49/1500 [00:01<00:29,
49.91it/s]
Running loglikelihood requests:   5%|         | 73/1500 [00:01<00:21,
64.95it/s]
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:18,
75.93it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:02<00:16,
84.60it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:02<00:14,
91.18it/s]
Running loglikelihood requests:  11%|        | 169/1500 [00:02<00:13,
96.39it/s]
Running loglikelihood requests:  13%|        | 193/1500 [00:02<00:13,
100.24it/s]
Running loglikelihood requests:  14%|        | 217/1500 [00:02<00:12,
103.64it/s]
Running loglikelihood requests:  16%|        | 241/1500 [00:03<00:11,
106.36it/s]
Running loglikelihood requests:  18%|        | 265/1500 [00:03<00:11,
109.80it/s]
Running loglikelihood requests:  19%|        | 289/1500 [00:03<00:10,
112.50it/s]
Running loglikelihood requests:  21%|        | 313/1500 [00:03<00:10,
115.10it/s]
Running loglikelihood requests:  22%|        | 337/1500 [00:03<00:09,
116.96it/s]
Running loglikelihood requests:  24%|        | 361/1500 [00:04<00:09,
118.71it/s]
Running loglikelihood requests:  26%|        | 385/1500 [00:04<00:09,
119.94it/s]
Running loglikelihood requests:  27%|        | 409/1500 [00:04<00:08,
122.08it/s]
Running loglikelihood requests:  29%|        | 433/1500 [00:04<00:08,
123.58it/s]
```

```
Running loglikelihood requests:  30%|        | 457/1500 [00:04<00:08,
124.87it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:05<00:08,
126.70it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:05<00:07,
127.96it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:05<00:07,
128.81it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:05<00:07,
129.58it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:05<00:07,
130.47it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:05<00:06,
131.16it/s]
Running loglikelihood requests:  42%|       | 625/1500 [00:06<00:06,
131.84it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:06<00:06,
132.51it/s]
Running loglikelihood requests:  45%|        | 673/1500 [00:06<00:06,
133.43it/s]
Running loglikelihood requests:  46%|        | 697/1500 [00:06<00:05,
135.35it/s]
Running loglikelihood requests:  48%|        | 721/1500 [00:06<00:05,
136.52it/s]
Running loglikelihood requests:  50%|        | 745/1500 [00:06<00:05,
138.07it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:07<00:05,
141.32it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:07<00:04,
143.77it/s]
Running loglikelihood requests:  54%|        | 817/1500 [00:07<00:04,
145.88it/s]
Running loglikelihood requests:  56%|        | 841/1500 [00:07<00:04,
147.95it/s]
Running loglikelihood requests:  58%|        | 865/1500 [00:07<00:04,
150.60it/s]
Running loglikelihood requests:  59%|        | 889/1500 [00:07<00:04,
152.59it/s]
Running loglikelihood requests:  61%|        | 913/1500 [00:08<00:03,
154.27it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:08<00:03,
155.78it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:08<00:03,
158.61it/s]
Running loglikelihood requests:  66%|       | 985/1500 [00:08<00:03,
160.50it/s]
Running loglikelihood requests:  67%|       | 1009/1500 [00:08<00:03,
162.33it/s]
```

```
Running loglikelihood requests:  69%|       | 1033/1500 [00:08<00:02,
163.85it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:08<00:02,
165.41it/s]
Running loglikelihood requests:  72%|      | 1081/1500 [00:09<00:02,
167.16it/s]
Running loglikelihood requests:  74%|      | 1105/1500 [00:09<00:02,
168.76it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:09<00:02,
171.98it/s]
Running loglikelihood requests:  77%|      | 1153/1500 [00:09<00:01,
174.24it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:09<00:01,
176.73it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:09<00:01,
178.58it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:09<00:01,
182.72it/s]
Running loglikelihood requests:  83%|      | 1249/1500 [00:10<00:01,
186.09it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:10<00:01,
188.67it/s]
Running loglikelihood requests:  86%|      | 1297/1500 [00:10<00:01,
191.60it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:10<00:00,
194.16it/s]
Running loglikelihood requests:  90%|      | 1345/1500 [00:10<00:00,
196.45it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:10<00:00,
201.25it/s]
Running loglikelihood requests:  93%|     | 1393/1500 [00:10<00:00,
205.92it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:10<00:00,
213.72it/s]
Running loglikelihood requests:  96%|     | 1442/1500 [00:10<00:00,
223.66it/s]
Running loglikelihood requests:  99%|     | 1489/1500 [00:11<00:00,
253.18it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:11<00:00,
135.50it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:01:57:24,092 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

 No eval_results_*.json found in ./results/run_4_2025-05-09T01-56-43

 Run 5/5
 Run 5 completed in 43.07 seconds
```

```
STDOUT:
 hf (pretrained=Qwen/Qwen3-4B,parallelize=True,trust_remote_code=True),
gen_kwargs: (None), limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.768|±  |0.0189|


STDERR:
 2025-05-09 01:57:30.130966: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746755850.152146   14902 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746755850.158564   14902 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:01:57:46,860 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:01:57:46,860 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:01:57:46,861 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:01:57:46,862 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'Qwen/Qwen3-4B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:01:57:46,902 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:01:57:47,890 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/3 [00:00<?, ?it/s]
Loading checkpoint shards:  33%|        | 1/3 [00:01<00:02,  1.15s/it]
Loading checkpoint shards:  67%|       | 2/3 [00:02<00:01,  1.16s/it]
Loading checkpoint shards: 100%|     | 3/3 [00:02<00:00,  1.27it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
```

```
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:01:57:51,210 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|     | 500/500 [00:00<00:00, 100313.40it/s]
2025-05-09:01:57:51,270 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:51,270 INFO     [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
demonstrate dysfunction of the anal sphincter and/or the musculus (m.)
puborectalis in patients with dyschesia using anorectal endosonography.',
'Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.', 'The anal sphincter became paradoxically shorter and/or
```

thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:57:51,270 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary

squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:51,270 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with

dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:57:51,270 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:01:57:51,270 INFO   [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to

demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:01:57:51,271 INFO [lm_eval.evaluator:517] Running loglikelihood requests

```
Running loglikelihood requests:   0%|          | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests:   0%|          | 1/1500 [00:00<19:52,  1.26it/s]
Running loglikelihood requests:   2%|         | 25/1500 [00:01<00:49,
29.93it/s]
Running loglikelihood requests:   3%|         | 49/1500 [00:01<00:29,
49.48it/s]
Running loglikelihood requests:   5%|         | 73/1500 [00:01<00:22,
64.53it/s]
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:18,
75.62it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:02<00:16,
84.22it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:02<00:14,
90.84it/s]
Running loglikelihood requests:  11%|         | 169/1500 [00:02<00:13,
96.14it/s]
Running loglikelihood requests:  13%|         | 193/1500 [00:02<00:13,
100.07it/s]
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:12,
103.61it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:03<00:11,
106.35it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:03<00:11,
109.74it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:03<00:10,
112.46it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:03<00:10,
115.05it/s]
Running loglikelihood requests:  22%|        | 337/1500 [00:03<00:09,
116.90it/s]
Running loglikelihood requests:  24%|        | 361/1500 [00:04<00:09,
118.53it/s]
Running loglikelihood requests:  26%|        | 385/1500 [00:04<00:09,
119.90it/s]
Running loglikelihood requests:  27%|        | 409/1500 [00:04<00:08,
122.16it/s]
Running loglikelihood requests:  29%|        | 433/1500 [00:04<00:08,
123.70it/s]
Running loglikelihood requests:  30%|        | 457/1500 [00:04<00:08,
125.13it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:05<00:08,
126.88it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:05<00:07,
128.14it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:05<00:07,
129.10it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:05<00:07,
130.05it/s]
```

```
Running loglikelihood requests:  38%|        | 577/1500 [00:05<00:07,
131.11it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:05<00:06,
131.96it/s]
Running loglikelihood requests:  42%|        | 625/1500 [00:06<00:06,
132.70it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:06<00:06,
133.43it/s]
Running loglikelihood requests:  45%|        | 673/1500 [00:06<00:06,
134.08it/s]
Running loglikelihood requests:  46%|        | 697/1500 [00:06<00:05,
135.51it/s]
Running loglikelihood requests:  48%|        | 721/1500 [00:06<00:05,
137.05it/s]
Running loglikelihood requests:  50%|        | 745/1500 [00:06<00:05,
138.33it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:07<00:05,
141.82it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:07<00:04,
144.55it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:07<00:04,
146.62it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:07<00:04,
147.79it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:07<00:04,
149.68it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:07<00:04,
151.93it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:08<00:03,
153.73it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:08<00:03,
154.84it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:08<00:03,
157.59it/s]
Running loglikelihood requests:  66%|       | 985/1500 [00:08<00:03,
159.91it/s]
Running loglikelihood requests:  67%|       | 1009/1500 [00:08<00:03,
161.67it/s]
Running loglikelihood requests:  69%|       | 1033/1500 [00:08<00:02,
163.28it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:08<00:02,
164.98it/s]
Running loglikelihood requests:  72%|      | 1081/1500 [00:09<00:02,
166.87it/s]
Running loglikelihood requests:  74%|      | 1105/1500 [00:09<00:02,
168.71it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:09<00:02,
172.09it/s]
```

```
Running loglikelihood requests:  77%|      | 1153/1500 [00:09<00:01,
174.48it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:09<00:01,
177.03it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:09<00:01,
178.58it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:09<00:01,
182.58it/s]
Running loglikelihood requests:  83%|      | 1249/1500 [00:10<00:01,
185.95it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:10<00:01,
188.81it/s]
Running loglikelihood requests:  86%|      | 1297/1500 [00:10<00:01,
191.25it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:10<00:00,
193.68it/s]
Running loglikelihood requests:  90%|      | 1345/1500 [00:10<00:00,
195.82it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:10<00:00,
201.14it/s]
Running loglikelihood requests:  93%|     | 1393/1500 [00:10<00:00,
205.78it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:10<00:00,
212.96it/s]
Running loglikelihood requests:  96%|     | 1443/1500 [00:10<00:00,
225.84it/s]
Running loglikelihood requests:  99%|     | 1489/1500 [00:11<00:00,
252.27it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:11<00:00,
135.33it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:01:58:07,151 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

 No eval_results_*.json found in ./results/run_5_2025-05-09T01-57-25

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>
```

[9]:
```python
import random
import json
import wandb
import subprocess
import time
import os
from datetime import datetime
```

```
# ----------------------------
#  Model and Task Config
# ----------------------------
model_name = "meta-llama/Llama-3.2-3B"
task_name = "pubmedqa"
output_base = "./results"

# ----------------------------
#  Start W&B run
# ----------------------------
run_name = f"{model_name.replace('/', '_')}_{task_name}_5x"
wandb_run = wandb.init(
    project="med-moe-baseline-evals",
    name=run_name,
    config={
        "model": model_name,
        "task": task_name,
        "batch_size": 8,
        "precision": "fp32",
        "eval_method": "lm_eval",
        "repeats": 5
    }
)

# ----------------------------
#  Run 5x Evaluation Loop
# ----------------------------
for i in range(5):
    print(f"\n Run {i + 1}/5")

    # Create timestamped output folder
    timestamp = datetime.now().strftime("%Y-%m-%dT%H-%M-%S")
    run_output_dir = os.path.join(output_base, f"run_{i+1}_{timestamp}")
    os.makedirs(run_output_dir, exist_ok=True)

    # Define lm_eval command
    command = [
        "lm_eval",
        "--model", "hf",
        "--tasks", task_name,
        "--model_args", f"pretrained={model_name},parallelize=True",
        "--device", "cuda:0",
        "--batch_size", "8",
        "--write_out",
        "--output_path", run_output_dir,
        "--trust_remote_code", "--confirm_run_unsafe_code"
```

```
    ]

    # Start timing
    start_time = time.monotonic()
    result = subprocess.run(command, capture_output=True, text=True)
    elapsed = time.monotonic() - start_time

    print(f"  Run {i + 1} completed in {elapsed:.2f} seconds")
    print("STDOUT:\n", result.stdout)
    print("STDERR:\n", result.stderr)

    # ----------------------------
    #   Find and parse result file
    # ----------------------------
    result_file = None
    for fname in os.listdir(run_output_dir):
        if fname.startswith("eval_results") and fname.endswith(".json"):
            result_file = os.path.join(run_output_dir, fname)
            break

    if result_file is None:
        print(f"  No eval_results_*.json found in {run_output_dir}")
        continue

    try:
        with open(result_file) as f:
            data = json.load(f)
        task_data = data["results"][task_name]

        acc = task_data.get("acc,none")
        stderr = task_data.get("acc_stderr,none")

        if acc is not None and stderr is not None:
            wandb_run.log({
                f"{task_name}/accuracy": acc,
                f"{task_name}/stddev": stderr,
                f"{task_name}/eval_time_sec": elapsed,
                "run_index": i + 1
            })
            print(f"  Logged to W&B: acc={acc:.3f}, stderr={stderr:.4f}")
        else:
            print(f"  Missing keys in result: {task_data.keys()}")

    except Exception as e:
        print(f"  Failed to parse results from {result_file}: {e}")

# ----------------------------
```

```
#  Finish W&B run
# ----------------------------
wandb_run.finish()
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
 Run 1/5
 Run 1 completed in 60.62 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


STDERR:
 2025-05-09 02:00:51.229049: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746756051.250535   15878 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746756051.257001   15878 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:02:01:08,057 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:02:01:08,057 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:02:01:08,059 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:02:01:08,059 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'meta-llama/Llama-3.2-3B', 'parallelize': True,
'trust_remote_code': True}
```

```
2025-05-09:02:01:08,099 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:02:01:10,214 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Fetching 2 files:   0%|          | 0/2 [00:00<?, ?it/s]
Fetching 2 files:  50%|        | 1/2 [00:20<00:20, 20.37s/it]
Fetching 2 files: 100%|      | 2/2 [00:20<00:00, 10.18s/it]

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|        | 1/2 [00:01<00:01,  1.45s/it]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.18it/s]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.07it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:02:01:33,957 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|      | 500/500 [00:00<00:00, 79715.37it/s]
2025-05-09:02:01:34,016 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
```

thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:01:34,016 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of

control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:01:34,016 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:01:34,016 INFO   [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or

thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:01:34,016 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary

squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:01:34,016 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with

dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:01:34,016 INFO   [lm_eval.evaluator:517] Running loglikelihood requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<17:23, 1.44it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:00<00:39, 37.12it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:22, 65.20it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:16, 86.73it/s]
Running loglikelihood requests: 6%| | 97/1500 [00:01<00:13, 104.06it/s]
Running loglikelihood requests: 8%| | 121/1500 [00:01<00:11, 117.11it/s]
Running loglikelihood requests: 10%| | 145/1500 [00:01<00:10, 126.84it/s]
Running loglikelihood requests: 11%| | 169/1500 [00:01<00:09, 134.80it/s]
Running loglikelihood requests: 13%| | 193/1500 [00:01<00:08, 145.68it/s]
Running loglikelihood requests: 14%| | 217/1500 [00:02<00:08, 153.97it/s]
Running loglikelihood requests: 16%| | 241/1500 [00:02<00:07, 160.41it/s]
Running loglikelihood requests: 18%| | 265/1500 [00:02<00:07, 165.70it/s]
Running loglikelihood requests: 19%| | 289/1500 [00:02<00:07, 169.64it/s]
Running loglikelihood requests: 21%| | 313/1500 [00:02<00:06, 172.68it/s]

```
Running loglikelihood requests:  22%|         | 337/1500 [00:02<00:06,
175.43it/s]
Running loglikelihood requests:  24%|         | 361/1500 [00:02<00:06,
177.34it/s]
Running loglikelihood requests:  26%|         | 385/1500 [00:03<00:06,
180.06it/s]
Running loglikelihood requests:  27%|         | 409/1500 [00:03<00:05,
182.06it/s]
Running loglikelihood requests:  29%|         | 433/1500 [00:03<00:05,
183.73it/s]
Running loglikelihood requests:  30%|         | 457/1500 [00:03<00:05,
185.98it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:03<00:05,
187.75it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:03<00:05,
189.69it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:03<00:05,
191.19it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:03<00:04,
193.07it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:04<00:04,
194.74it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:04<00:04,
196.66it/s]
Running loglikelihood requests:  42%|        | 625/1500 [00:04<00:04,
198.46it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:04<00:04,
199.73it/s]
Running loglikelihood requests:  45%|        | 673/1500 [00:04<00:04,
200.37it/s]
Running loglikelihood requests:  46%|        | 697/1500 [00:04<00:03,
201.79it/s]
Running loglikelihood requests:  48%|        | 721/1500 [00:04<00:03,
203.48it/s]
Running loglikelihood requests:  50%|        | 745/1500 [00:04<00:03,
204.91it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:04<00:03,
205.79it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:05<00:03,
208.96it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:05<00:03,
212.14it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:05<00:03,
214.31it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:05<00:02,
216.76it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:05<00:02,
218.89it/s]
```

```
Running loglikelihood requests:  61%|       | 913/1500 [00:05<00:02,
219.63it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:05<00:02,
221.31it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:05<00:02,
222.65it/s]
Running loglikelihood requests:  66%|       | 994/1500 [00:05<00:01,
253.12it/s]
Running loglikelihood requests:  69%|       | 1033/1500 [00:06<00:02,
231.22it/s]
Running loglikelihood requests:  72%|      | 1076/1500 [00:06<00:01,
280.09it/s]
Running loglikelihood requests:  74%|      | 1106/1500 [00:06<00:01,
231.60it/s]
Running loglikelihood requests:  77%|      | 1153/1500 [00:06<00:01,
238.15it/s]
Running loglikelihood requests:  80%|      | 1201/1500 [00:06<00:01,
250.20it/s]
Running loglikelihood requests:  83%|      | 1249/1500 [00:06<00:00,
259.37it/s]
Running loglikelihood requests:  86%|      | 1297/1500 [00:07<00:00,
267.47it/s]
Running loglikelihood requests:  90%|      | 1345/1500 [00:07<00:00,
274.87it/s]
Running loglikelihood requests:  93%|     | 1393/1500 [00:07<00:00,
285.53it/s]
Running loglikelihood requests:  96%|     | 1441/1500 [00:07<00:00,
302.97it/s]
Running loglikelihood requests:  99%|     | 1489/1500 [00:07<00:00,
332.51it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:07<00:00,
194.96it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:02:01:45,857 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_1_2025-05-09T02-00-47

  Run 2/5
  Run 2 completed in 38.26 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|
```

```
STDERR:
 2025-05-09 02:01:51.911008: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746756111.932779   16207 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746756111.939192   16207 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:02:02:08,572 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:02:02:08,572 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:02:02:08,573 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:02:02:08,573 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'meta-llama/Llama-3.2-3B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:02:02:08,614 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:02:02:09,385 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|        | 1/2 [00:01<00:01,  1.58s/it]
Loading checkpoint shards: 100%|      | 2/2 [00:02<00:00,  1.08it/s]
Loading checkpoint shards: 100%|      | 2/2 [00:02<00:00,  1.02s/it]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:02:02:12,165 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…
```

```
  0%|          | 0/500 [00:00<?, ?it/s]
100%|     | 500/500 [00:00<00:00, 80163.30it/s]
2025-05-09:02:02:12,225 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:02:12,225 INFO     [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
demonstrate dysfunction of the anal sphincter and/or the musculus (m.)
puborectalis in patients with dyschesia using anorectal endosonography.',
'Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.', 'The anal sphincter became paradoxically shorter and/or
thicker during straining (versus the resting state) in 85% of patients but in
only 35% of control subjects. Changes in sphincter length were statistically
significantly different (p<0.01, chi(2) test) in patients compared with control
subjects. The m. puborectalis became paradoxically shorter and/or thicker during
straining in 80% of patients but in only 30% of controls. Both the changes in
length and thickness of the m. puborectalis were significantly different
(p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS',
```

'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:02:12,225 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80%

of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:02:12,225 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during

straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:02:12,225 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:02:12,225 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between

the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:02:12,225 INFO   [lm_eval.evaluator:517] Running loglikelihood requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<17:11, 1.45it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:00<00:39, 37.47it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:22, 65.70it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:16,

```
87.17it/s]
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:13,
104.40it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:01<00:11,
117.45it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:01<00:10,
127.22it/s]
Running loglikelihood requests:  11%|         | 169/1500 [00:01<00:09,
135.04it/s]
Running loglikelihood requests:  13%|         | 193/1500 [00:01<00:09,
145.07it/s]
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:08,
153.61it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:02<00:07,
160.21it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:02<00:07,
165.33it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:02<00:07,
169.35it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:02<00:06,
172.50it/s]
Running loglikelihood requests:  22%|         | 337/1500 [00:02<00:06,
175.35it/s]
Running loglikelihood requests:  24%|         | 361/1500 [00:02<00:06,
177.41it/s]
Running loglikelihood requests:  26%|         | 385/1500 [00:03<00:06,
179.75it/s]
Running loglikelihood requests:  27%|         | 409/1500 [00:03<00:06,
181.77it/s]
Running loglikelihood requests:  29%|         | 433/1500 [00:03<00:05,
183.54it/s]
Running loglikelihood requests:  30%|         | 457/1500 [00:03<00:05,
186.72it/s]
Running loglikelihood requests:  32%|         | 481/1500 [00:03<00:05,
188.69it/s]
Running loglikelihood requests:  34%|         | 505/1500 [00:03<00:05,
190.51it/s]
Running loglikelihood requests:  35%|         | 529/1500 [00:03<00:05,
192.02it/s]
Running loglikelihood requests:  37%|         | 553/1500 [00:03<00:04,
193.04it/s]
Running loglikelihood requests:  38%|         | 577/1500 [00:04<00:04,
194.45it/s]
Running loglikelihood requests:  40%|         | 601/1500 [00:04<00:04,
196.71it/s]
Running loglikelihood requests:  42%|         | 625/1500 [00:04<00:04,
198.34it/s]
Running loglikelihood requests:  43%|         | 649/1500 [00:04<00:04,
```

```
198.90it/s]
Running loglikelihood requests:  45%|       | 673/1500 [00:04<00:04,
200.23it/s]
Running loglikelihood requests:  46%|       | 697/1500 [00:04<00:03,
201.69it/s]
Running loglikelihood requests:  48%|       | 721/1500 [00:04<00:03,
203.38it/s]
Running loglikelihood requests:  50%|       | 745/1500 [00:04<00:03,
203.92it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:04<00:03,
205.10it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:05<00:03,
208.30it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:05<00:03,
211.60it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:05<00:03,
213.92it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:05<00:02,
216.27it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:05<00:02,
218.42it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:05<00:02,
219.32it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:05<00:02,
220.67it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:05<00:02,
221.93it/s]
Running loglikelihood requests:  66%|       | 988/1500 [00:05<00:02,
235.49it/s]
Running loglikelihood requests:  68%|       | 1023/1500 [00:06<00:01,
268.45it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:06<00:01,
228.68it/s]
Running loglikelihood requests:  73%|      | 1096/1500 [00:06<00:01,
268.55it/s]
Running loglikelihood requests:  75%|      | 1129/1500 [00:06<00:01,
230.97it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:06<00:01,
242.76it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:06<00:01,
253.89it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:07<00:00,
262.09it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:07<00:00,
269.62it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:07<00:00,
276.46it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:07<00:00,
```

```
293.16it/s]
Running loglikelihood requests:  98%|      | 1465/1500 [00:07<00:00,
310.26it/s]
Running loglikelihood requests: 100%|      | 1500/1500 [00:07<00:00,
194.80it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:02:02:24,082 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_2_2025-05-09T02-01-47

  Run 3/5
  Run 3 completed in 38.23 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


STDERR:
 2025-05-09 02:02:30.192398: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746756150.213914   16433 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746756150.220423   16433 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:02:02:46,800 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:02:02:46,800 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:02:02:46,801 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:02:02:46,802 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'meta-llama/Llama-3.2-3B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:02:02:46,842 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
```

```
when placing model.
2025-05-09:02:02:47,687 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|        | 1/2 [00:01<00:01,  1.43s/it]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.20it/s]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.08it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:02:02:50,427 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|      | 500/500 [00:00<00:00, 77726.99it/s]
2025-05-09:02:02:50,488 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
```

yes
(end of target on previous line)
2025-05-09:02:02:50,489 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal

endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:02:02:50,489 INFO [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:02:50,489 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
demonstrate dysfunction of the anal sphincter and/or the musculus (m.)
puborectalis in patients with dyschesia using anorectal endosonography.',
'Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.', 'The anal sphincter became paradoxically shorter and/or
thicker during straining (versus the resting state) in 85% of patients but in
only 35% of control subjects. Changes in sphincter length were statistically
significantly different (p<0.01, chi(2) test) in patients compared with control
subjects. The m. puborectalis became paradoxically shorter and/or thicker during
straining in 80% of patients but in only 30% of controls. Both the changes in
length and thickness of the m. puborectalis were significantly different

(p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:02:50,489 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m.

puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:02:50,489 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two

```
groups.\nThe anal sphincter became paradoxically shorter and/or thicker during
straining (versus the resting state) in 85% of patients but in only 35% of
control subjects. Changes in sphincter length were statistically significantly
different (p<0.01, chi(2) test) in patients compared with control subjects. The
m. puborectalis became paradoxically shorter and/or thicker during straining in
80% of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.\nQuestion: Is anorectal
endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:02:02:50,489 INFO     [lm_eval.evaluator:517] Running loglikelihood
requests

Running loglikelihood requests:   0%|          | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests:   0%|          | 1/1500 [00:00<17:11,  1.45it/s]
Running loglikelihood requests:   2%|         | 25/1500 [00:00<00:39,
37.37it/s]
Running loglikelihood requests:   3%|         | 49/1500 [00:01<00:22,
65.42it/s]
Running loglikelihood requests:   5%|         | 73/1500 [00:01<00:16,
86.98it/s]
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:13,
104.34it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:01<00:11,
117.66it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:01<00:10,
127.60it/s]
Running loglikelihood requests:  11%|         | 169/1500 [00:01<00:09,
135.01it/s]
Running loglikelihood requests:  13%|         | 193/1500 [00:01<00:08,
145.80it/s]
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:08,
154.14it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:02<00:07,
160.72it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:02<00:07,
165.74it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:02<00:07,
169.49it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:02<00:06,
172.68it/s]
Running loglikelihood requests:  22%|         | 337/1500 [00:02<00:06,
175.09it/s]
Running loglikelihood requests:  24%|         | 361/1500 [00:02<00:06,
177.16it/s]
Running loglikelihood requests:  26%|         | 385/1500 [00:03<00:06,
179.05it/s]
```

```
Running loglikelihood requests:  27%|         | 409/1500 [00:03<00:06,
181.47it/s]
Running loglikelihood requests:  29%|         | 433/1500 [00:03<00:05,
182.84it/s]
Running loglikelihood requests:  30%|         | 457/1500 [00:03<00:05,
186.22it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:03<00:05,
189.04it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:03<00:05,
191.38it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:03<00:05,
193.07it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:03<00:04,
194.53it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:04<00:04,
195.96it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:04<00:04,
198.00it/s]
Running loglikelihood requests:  42%|        | 625/1500 [00:04<00:04,
199.68it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:04<00:04,
200.87it/s]
Running loglikelihood requests:  45%|        | 673/1500 [00:04<00:04,
201.95it/s]
Running loglikelihood requests:  46%|        | 697/1500 [00:04<00:03,
202.58it/s]
Running loglikelihood requests:  48%|        | 721/1500 [00:04<00:03,
204.05it/s]
Running loglikelihood requests:  50%|        | 745/1500 [00:04<00:03,
205.34it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:04<00:03,
206.12it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:05<00:03,
209.22it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:05<00:03,
212.16it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:05<00:03,
214.01it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:05<00:02,
216.46it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:05<00:02,
218.52it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:05<00:02,
219.44it/s]
Running loglikelihood requests:  62%|      | 937/1500 [00:05<00:02,
220.80it/s]
Running loglikelihood requests:  64%|      | 961/1500 [00:05<00:02,
222.02it/s]
```

```
Running loglikelihood requests:  66%|       | 993/1500 [00:05<00:02,
249.75it/s]
Running loglikelihood requests:  68%|       | 1025/1500 [00:06<00:01,
270.00it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:06<00:01,
225.08it/s]
Running loglikelihood requests:  74%|      | 1103/1500 [00:06<00:01,
283.32it/s]
Running loglikelihood requests:  76%|      | 1134/1500 [00:06<00:01,
236.30it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:06<00:01,
237.16it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:06<00:01,
250.03it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:07<00:00,
260.53it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:07<00:00,
269.26it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:07<00:00,
276.02it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:07<00:00,
292.95it/s]
Running loglikelihood requests:  98%|     | 1465/1500 [00:07<00:00,
309.36it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:07<00:00,
195.15it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:02:03:02,353 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_3_2025-05-09T02-02-25

  Run 4/5
  Run 4 completed in 38.42 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


STDERR:
 2025-05-09 02:03:08.365706: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
```

```
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746756188.386815   16661 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746756188.393108   16661 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:02:03:25,131 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:02:03:25,132 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:02:03:25,133 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:02:03:25,133 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'meta-llama/Llama-3.2-3B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:02:03:25,173 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:02:03:26,005 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|        | 1/2 [00:01<00:01,  1.43s/it]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.20it/s]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.08it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:02:03:28,707 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|      | 500/500 [00:00<00:00, 80830.68it/s]
2025-05-09:02:03:28,768 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
```

aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography

demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)

target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2)

test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator_utils:210] Request: Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in

length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:03:28,768 INFO [lm_eval.evaluator:517] Running loglikelihood requests

Running loglikelihood requests: 0%| | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests: 0%| | 1/1500 [00:00<17:45, 1.41it/s]
Running loglikelihood requests: 2%| | 25/1500 [00:00<00:40, 36.54it/s]
Running loglikelihood requests: 3%| | 49/1500 [00:01<00:22, 64.46it/s]
Running loglikelihood requests: 5%| | 73/1500 [00:01<00:16, 85.89it/s]
Running loglikelihood requests: 6%| | 97/1500 [00:01<00:13, 103.22it/s]
Running loglikelihood requests: 8%| | 121/1500 [00:01<00:11, 116.45it/s]
Running loglikelihood requests: 10%| | 145/1500 [00:01<00:10,

```
126.39it/s]
Running loglikelihood requests:  11%|         | 169/1500 [00:01<00:09,
134.05it/s]
Running loglikelihood requests:  13%|         | 193/1500 [00:01<00:09,
145.00it/s]
Running loglikelihood requests:  14%|         | 217/1500 [00:02<00:08,
153.51it/s]
Running loglikelihood requests:  16%|         | 241/1500 [00:02<00:07,
160.14it/s]
Running loglikelihood requests:  18%|         | 265/1500 [00:02<00:07,
165.35it/s]
Running loglikelihood requests:  19%|         | 289/1500 [00:02<00:07,
169.33it/s]
Running loglikelihood requests:  21%|         | 313/1500 [00:02<00:06,
172.39it/s]
Running loglikelihood requests:  22%|        | 337/1500 [00:02<00:06,
175.45it/s]
Running loglikelihood requests:  24%|        | 361/1500 [00:02<00:06,
177.19it/s]
Running loglikelihood requests:  26%|        | 385/1500 [00:03<00:06,
179.32it/s]
Running loglikelihood requests:  27%|        | 409/1500 [00:03<00:06,
180.18it/s]
Running loglikelihood requests:  29%|        | 433/1500 [00:03<00:05,
181.85it/s]
Running loglikelihood requests:  30%|        | 457/1500 [00:03<00:05,
185.21it/s]
Running loglikelihood requests:  32%|        | 481/1500 [00:03<00:05,
187.83it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:03<00:05,
190.46it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:03<00:05,
192.41it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:03<00:04,
194.06it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:04<00:04,
195.57it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:04<00:04,
197.44it/s]
Running loglikelihood requests:  42%|       | 625/1500 [00:04<00:04,
199.14it/s]
Running loglikelihood requests:  43%|       | 649/1500 [00:04<00:04,
200.35it/s]
Running loglikelihood requests:  45%|       | 673/1500 [00:04<00:04,
201.61it/s]
Running loglikelihood requests:  46%|       | 697/1500 [00:04<00:03,
202.35it/s]
Running loglikelihood requests:  48%|       | 721/1500 [00:04<00:03,
```

```
203.60it/s]
Running loglikelihood requests:  50%|       | 745/1500 [00:04<00:03,
205.06it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:04<00:03,
206.14it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:05<00:03,
208.80it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:05<00:03,
211.01it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:05<00:03,
212.98it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:05<00:02,
215.37it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:05<00:02,
216.65it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:05<00:02,
218.21it/s]
Running loglikelihood requests:  62%|       | 937/1500 [00:05<00:02,
220.00it/s]
Running loglikelihood requests:  64%|       | 961/1500 [00:05<00:02,
220.73it/s]
Running loglikelihood requests:  66%|       | 995/1500 [00:05<00:01,
254.41it/s]
Running loglikelihood requests:  69%|       | 1028/1500 [00:06<00:01,
276.05it/s]
Running loglikelihood requests:  70%|       | 1057/1500 [00:06<00:02,
220.94it/s]
Running loglikelihood requests:  73%|       | 1098/1500 [00:06<00:01,
267.57it/s]
Running loglikelihood requests:  75%|       | 1129/1500 [00:06<00:01,
227.37it/s]
Running loglikelihood requests:  78%|       | 1177/1500 [00:06<00:01,
240.23it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:06<00:01,
251.87it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:07<00:00,
261.08it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:07<00:00,
269.37it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:07<00:00,
276.33it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:07<00:00,
293.13it/s]
Running loglikelihood requests:  98%|     | 1465/1500 [00:07<00:00,
310.31it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:07<00:00,
194.14it/s]
fatal: not a git repository (or any of the parent directories): .git
```

```
2025-05-09:02:03:40,642 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_4_2025-05-09T02-03-04

  Run 5/5
  Run 5 completed in 38.33 seconds
STDOUT:
 hf (pretrained=meta-
llama/Llama-3.2-3B,parallelize=True,trust_remote_code=True), gen_kwargs: (None),
limit: None, num_fewshot: None, batch_size: 8
| Tasks  |Version|Filter|n-shot|Metric|   |Value|   |Stderr|
|--------|------:|------|-----:|------|---|----:|---|-----:|
|pubmedqa|      1|none  |     0|acc   |↑  |0.732|±  |0.0198|


STDERR:
 2025-05-09 02:03:46.917891: E
external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:477] Unable to register
cuFFT factory: Attempting to register factory for plugin cuFFT when one has
already been registered
WARNING: All log messages before absl::InitializeLog() is called are written to
STDERR
E0000 00:00:1746756226.939653   16887 cuda_dnn.cc:8310] Unable to register cuDNN
factory: Attempting to register factory for plugin cuDNN when one has already
been registered
E0000 00:00:1746756226.946178   16887 cuda_blas.cc:1418] Unable to register
cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has
already been registered
2025-05-09:02:04:03,786 INFO     [lm_eval.__main__:368] Passed
`--trust_remote_code`, setting environment variable
`HF_DATASETS_TRUST_REMOTE_CODE=true`
2025-05-09:02:04:03,786 INFO     [lm_eval.__main__:379] Selected Tasks:
['pubmedqa']
2025-05-09:02:04:03,788 INFO     [lm_eval.evaluator:169] Setting random seed to
0 | Setting numpy seed to 1234 | Setting torch manual seed to 1234 | Setting
fewshot manual seed to 1234
2025-05-09:02:04:03,788 INFO     [lm_eval.evaluator:206] Initializing hf model,
with arguments: {'pretrained': 'meta-llama/Llama-3.2-3B', 'parallelize': True,
'trust_remote_code': True}
2025-05-09:02:04:03,828 INFO     [lm_eval.models.huggingface:153] Using
`accelerate launch` or `parallelize=True`, device 'cuda:0' will be overridden
when placing model.
2025-05-09:02:04:04,617 INFO     [lm_eval.models.huggingface:359] Model parallel
was set to True, setting max memory per GPU to {0: 42027843584} and device map
to auto

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
```

```
Loading checkpoint shards:  50%|        | 1/2 [00:01<00:01,  1.43s/it]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.20it/s]
Loading checkpoint shards: 100%|      | 2/2 [00:01<00:00,  1.08it/s]
/usr/local/lib/python3.11/dist-packages/datasets/load.py:1231: FutureWarning:
The repository for bigbio/pubmed_qa contains custom code which must be executed
to correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/bigbio/pubmed_qa
You can avoid this message in future by passing the argument
`trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the
next major release of `datasets`.
  warnings.warn(
2025-05-09:02:04:07,208 INFO     [lm_eval.api.task:420] Building contexts for
pubmedqa on rank 0…

  0%|          | 0/500 [00:00<?, ?it/s]
100%|     | 500/500 [00:00<00:00, 79446.60it/s]
2025-05-09:02:04:07,266 INFO     [lm_eval.evaluator_utils:206] Task: Configurabl
eTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=5
00); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The
aim of this prospective study was to demonstrate dysfunction of the anal
sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using
anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining
(versus the resting state) in 85% of patients but in only 35% of control
subjects. Changes in sphincter length were statistically significantly different
(p<0.01, chi(2) test) in patients compared with control subjects. The m.
puborectalis became paradoxically shorter and/or thicker during straining in 80%
of patients but in only 30% of controls. Both the changes in length and
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:04:07,266 INFO     [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
```

demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes', 'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' yes'), idx=0, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:04:07,266 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):

Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:
(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:04:07,266 INFO   [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.', 'Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.', 'The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS', 'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution', 'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male', 'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002', 'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes',

'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography demonstrated incomplete or even absent relaxation of the anal sphincter and the m. puborectalis during a defecation movement in the majority of our patients with dyschesia. This study highlights the value of this elegant ultrasonographic technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'}, arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.\nTwenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.\nThe anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.\nQuestion: Is anorectal endosonography valuable in dyschesia?\nAnswer:', ' no'), idx=1, metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa', doc_id=0, repeats=1)
2025-05-09:02:04:07,266 INFO   [lm_eval.evaluator_utils:206] Task: ConfigurableTask(task_name=pubmedqa,output_type=multiple_choice,num_fewshot=0,num_samples=500); document 0; context prompt (starting on next line):
Abstract: Dyschesia can be provoked by inappropriate defecation movements. The aim of this prospective study was to demonstrate dysfunction of the anal sphincter and/or the musculus (m.) puborectalis in patients with dyschesia using anorectal endosonography.
Twenty consecutive patients with a medical history of dyschesia and a control group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal sphincter and the m. puborectalis were measured at rest, and during voluntary squeezing and straining. Statistical analysis was performed within and between the two groups.
The anal sphincter became paradoxically shorter and/or thicker during straining (versus the resting state) in 85% of patients but in only 35% of control subjects. Changes in sphincter length were statistically significantly different (p<0.01, chi(2) test) in patients compared with control subjects. The m. puborectalis became paradoxically shorter and/or thicker during straining in 80% of patients but in only 30% of controls. Both the changes in length and thickness of the m. puborectalis were significantly different (p<0.01, chi(2) test) in patients versus control subjects.
Question: Is anorectal endosonography valuable in dyschesia?
Answer:

(end of prompt on previous line)
target string or answer choice index (starting on next line):
yes
(end of target on previous line)
2025-05-09:02:04:07,266 INFO [lm_eval.evaluator_utils:210] Request:
Instance(request_type='loglikelihood', doc={'QUESTION': 'Is anorectal
endosonography valuable in dyschesia?', 'CONTEXTS': ['Dyschesia can be provoked
by inappropriate defecation movements. The aim of this prospective study was to
demonstrate dysfunction of the anal sphincter and/or the musculus (m.)
puborectalis in patients with dyschesia using anorectal endosonography.',
'Twenty consecutive patients with a medical history of dyschesia and a control
group of 20 healthy subjects underwent linear anorectal endosonography (Toshiba
models IUV 5060 and PVL-625 RT). In both groups, the dimensions of the anal
sphincter and the m. puborectalis were measured at rest, and during voluntary
squeezing and straining. Statistical analysis was performed within and between
the two groups.', 'The anal sphincter became paradoxically shorter and/or
thicker during straining (versus the resting state) in 85% of patients but in
only 35% of control subjects. Changes in sphincter length were statistically
significantly different (p<0.01, chi(2) test) in patients compared with control
subjects. The m. puborectalis became paradoxically shorter and/or thicker during
straining in 80% of patients but in only 30% of controls. Both the changes in
length and thickness of the m. puborectalis were significantly different
(p<0.01, chi(2) test) in patients versus control subjects.'], 'LABELS': ['AIMS',
'METHODS', 'RESULTS'], 'MESHES': ['Adolescent', 'Adult', 'Aged', 'Aged, 80 and
over', 'Anal Canal', 'Case-Control Studies', 'Chi-Square Distribution',
'Constipation', 'Defecation', 'Endosonography', 'Female', 'Humans', 'Male',
'Middle Aged', 'Pelvic Floor', 'Rectum'], 'YEAR': '2002',
'reasoning_required_pred': 'yes', 'reasoning_free_pred': 'yes',
'final_decision': 'yes', 'LONG_ANSWER': 'Linear anorectal endosonography
demonstrated incomplete or even absent relaxation of the anal sphincter and the
m. puborectalis during a defecation movement in the majority of our patients
with dyschesia. This study highlights the value of this elegant ultrasonographic
technique in the diagnosis of "pelvic floor dyssynergia" or "anismus".'},
arguments=('Abstract: Dyschesia can be provoked by inappropriate defecation
movements. The aim of this prospective study was to demonstrate dysfunction of
the anal sphincter and/or the musculus (m.) puborectalis in patients with
dyschesia using anorectal endosonography.\nTwenty consecutive patients with a
medical history of dyschesia and a control group of 20 healthy subjects
underwent linear anorectal endosonography (Toshiba models IUV 5060 and PVL-625
RT). In both groups, the dimensions of the anal sphincter and the m.
puborectalis were measured at rest, and during voluntary squeezing and
straining. Statistical analysis was performed within and between the two
groups.\nThe anal sphincter became paradoxically shorter and/or thicker during
straining (versus the resting state) in 85% of patients but in only 35% of
control subjects. Changes in sphincter length were statistically significantly
different (p<0.01, chi(2) test) in patients compared with control subjects. The
m. puborectalis became paradoxically shorter and/or thicker during straining in
80% of patients but in only 30% of controls. Both the changes in length and

```
thickness of the m. puborectalis were significantly different (p<0.01, chi(2)
test) in patients versus control subjects.\nQuestion: Is anorectal
endosonography valuable in dyschesia?\nAnswer:', ' maybe'), idx=2,
metadata=('pubmedqa', 0, 1), resps=[], filtered_resps={}, task_name='pubmedqa',
doc_id=0, repeats=1)
2025-05-09:02:04:07,266 INFO     [lm_eval.evaluator:517] Running loglikelihood
requests

Running loglikelihood requests:   0%|          | 0/1500 [00:00<?, ?it/s]
Running loglikelihood requests:   0%|          | 1/1500 [00:00<16:59,  1.47it/s]
Running loglikelihood requests:   2%|         | 25/1500 [00:00<00:39,
37.80it/s]
Running loglikelihood requests:   3%|         | 49/1500 [00:01<00:21,
65.99it/s]
Running loglikelihood requests:   5%|         | 73/1500 [00:01<00:16,
87.39it/s]
Running loglikelihood requests:   6%|         | 97/1500 [00:01<00:13,
104.49it/s]
Running loglikelihood requests:   8%|         | 121/1500 [00:01<00:11,
117.43it/s]
Running loglikelihood requests:  10%|         | 145/1500 [00:01<00:10,
126.98it/s]
Running loglikelihood requests:  11%|        | 169/1500 [00:01<00:09,
134.56it/s]
Running loglikelihood requests:  13%|        | 193/1500 [00:01<00:08,
145.34it/s]
Running loglikelihood requests:  14%|        | 217/1500 [00:02<00:08,
153.88it/s]
Running loglikelihood requests:  16%|        | 241/1500 [00:02<00:07,
160.59it/s]
Running loglikelihood requests:  18%|        | 265/1500 [00:02<00:07,
165.62it/s]
Running loglikelihood requests:  19%|        | 289/1500 [00:02<00:07,
169.54it/s]
Running loglikelihood requests:  21%|        | 313/1500 [00:02<00:06,
172.59it/s]
Running loglikelihood requests:  22%|        | 337/1500 [00:02<00:06,
175.00it/s]
Running loglikelihood requests:  24%|        | 361/1500 [00:02<00:06,
177.34it/s]
Running loglikelihood requests:  26%|        | 385/1500 [00:03<00:06,
179.09it/s]
Running loglikelihood requests:  27%|        | 409/1500 [00:03<00:06,
181.39it/s]
Running loglikelihood requests:  29%|        | 433/1500 [00:03<00:05,
183.16it/s]
Running loglikelihood requests:  30%|        | 457/1500 [00:03<00:05,
185.98it/s]
```

```
Running loglikelihood requests:  32%|        | 481/1500 [00:03<00:05,
188.69it/s]
Running loglikelihood requests:  34%|        | 505/1500 [00:03<00:05,
191.19it/s]
Running loglikelihood requests:  35%|        | 529/1500 [00:03<00:05,
193.04it/s]
Running loglikelihood requests:  37%|        | 553/1500 [00:03<00:04,
194.54it/s]
Running loglikelihood requests:  38%|        | 577/1500 [00:04<00:04,
195.96it/s]
Running loglikelihood requests:  40%|        | 601/1500 [00:04<00:04,
198.05it/s]
Running loglikelihood requests:  42%|       | 625/1500 [00:04<00:04,
199.73it/s]
Running loglikelihood requests:  43%|        | 649/1500 [00:04<00:04,
200.89it/s]
Running loglikelihood requests:  45%|        | 673/1500 [00:04<00:04,
201.73it/s]
Running loglikelihood requests:  46%|        | 697/1500 [00:04<00:03,
202.53it/s]
Running loglikelihood requests:  48%|        | 721/1500 [00:04<00:03,
203.95it/s]
Running loglikelihood requests:  50%|        | 745/1500 [00:04<00:03,
205.25it/s]
Running loglikelihood requests:  51%|       | 769/1500 [00:04<00:03,
206.13it/s]
Running loglikelihood requests:  53%|       | 793/1500 [00:05<00:03,
209.04it/s]
Running loglikelihood requests:  54%|       | 817/1500 [00:05<00:03,
212.12it/s]
Running loglikelihood requests:  56%|       | 841/1500 [00:05<00:03,
214.20it/s]
Running loglikelihood requests:  58%|       | 865/1500 [00:05<00:02,
215.75it/s]
Running loglikelihood requests:  59%|       | 889/1500 [00:05<00:02,
218.09it/s]
Running loglikelihood requests:  61%|       | 913/1500 [00:05<00:02,
219.04it/s]
Running loglikelihood requests:  62%|      | 937/1500 [00:05<00:02,
220.82it/s]
Running loglikelihood requests:  64%|      | 961/1500 [00:05<00:02,
222.20it/s]
Running loglikelihood requests:  66%|      | 991/1500 [00:05<00:02,
244.26it/s]
Running loglikelihood requests:  68%|      | 1026/1500 [00:06<00:01,
274.76it/s]
Running loglikelihood requests:  70%|      | 1057/1500 [00:06<00:01,
225.82it/s]
```

```
Running loglikelihood requests:  74%|      | 1104/1500 [00:06<00:01,
286.43it/s]
Running loglikelihood requests:  76%|      | 1136/1500 [00:06<00:01,
240.04it/s]
Running loglikelihood requests:  78%|      | 1177/1500 [00:06<00:01,
235.92it/s]
Running loglikelihood requests:  82%|      | 1225/1500 [00:06<00:01,
248.70it/s]
Running loglikelihood requests:  85%|      | 1273/1500 [00:07<00:00,
259.46it/s]
Running loglikelihood requests:  88%|      | 1321/1500 [00:07<00:00,
268.42it/s]
Running loglikelihood requests:  91%|     | 1369/1500 [00:07<00:00,
275.30it/s]
Running loglikelihood requests:  94%|     | 1417/1500 [00:07<00:00,
292.40it/s]
Running loglikelihood requests:  98%|     | 1465/1500 [00:07<00:00,
309.95it/s]
Running loglikelihood requests: 100%|     | 1500/1500 [00:07<00:00,
195.31it/s]
fatal: not a git repository (or any of the parent directories): .git
2025-05-09:02:04:19,067 INFO     [lm_eval.loggers.evaluation_tracker:209] Saving
results aggregated

  No eval_results_*.json found in ./results/run_5_2025-05-09T02-03-42

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>
```