

Comparison of model parameters for equine pain face detection using freely available Google Gemini models

Dan Børge Jensen
Sofie Hougård Knudsen
Casper Lindegaard
Karina Gleerup
Marta Massa Gyldenkerne

KØBENHAVNS UNIVERSITET



Project: AI-based feed optimisation – real-time processing of large amounts of data in pig production



Ministeriet for Fødevarer,
Landbrug og Fiskeri

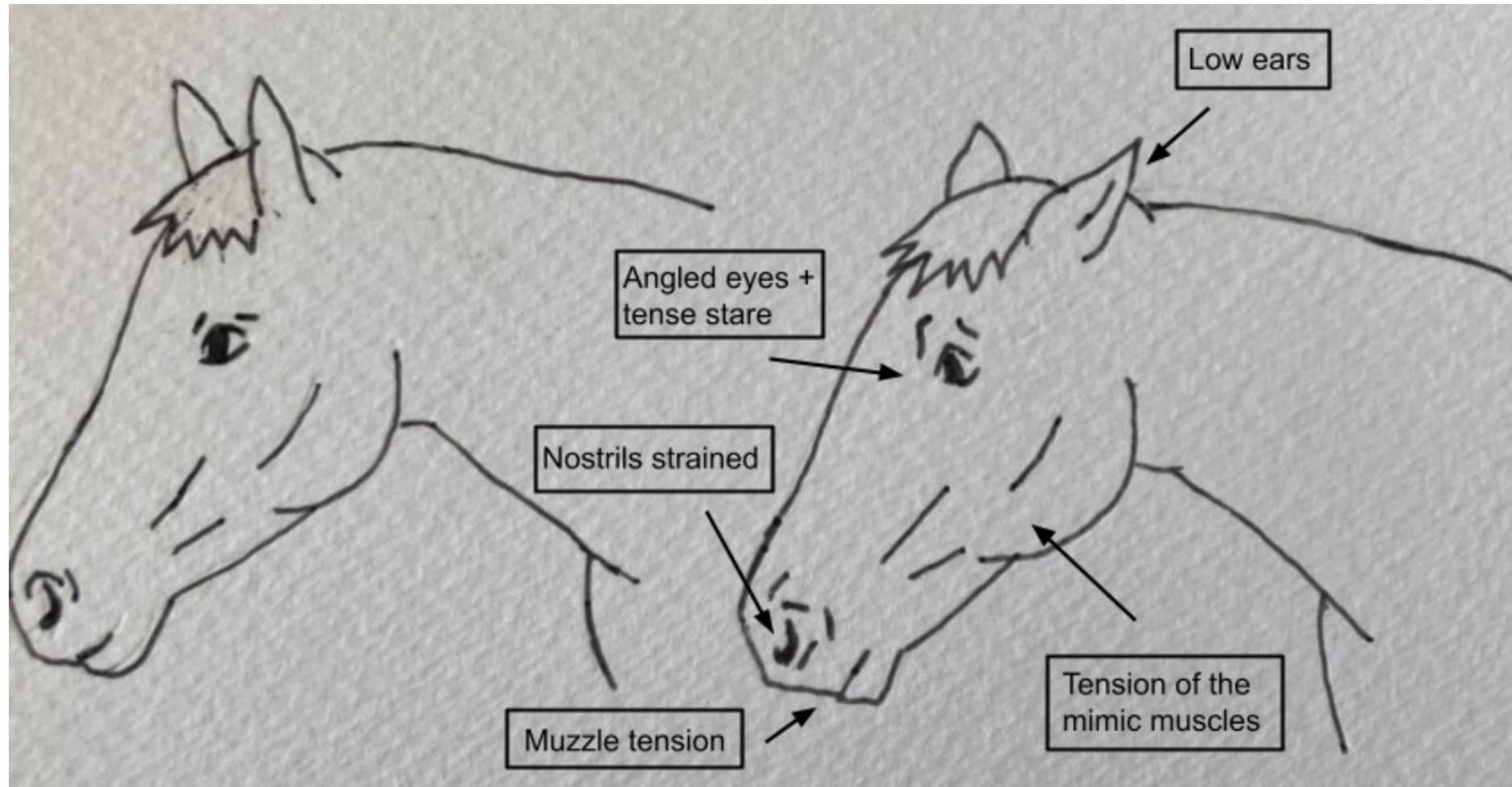
gudp

IQinAbox
SCIENTIFICITY BASED MONITORING

Motivation

- Studies have shown that 50 – 80 % of ridden horses suffer from pain
- Owners often mis-interpret pain behaviour as conflict behaviour
- Automatic pain assessment is expected to help horses to get earlier treatment

Visual identification of the equine pain face (EPF)



Pain or no pain?



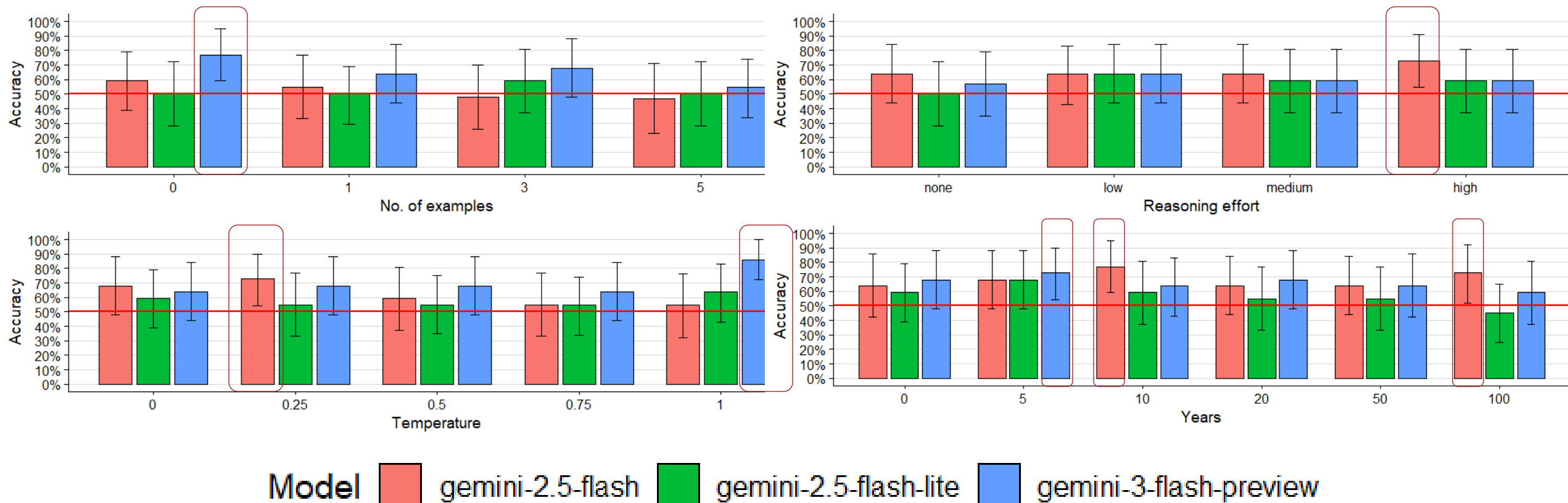
Materials and Methods

- 11 horses (4 different breeds, 16+ years old)
 - All suffering from *equine odontoclastic tooth resorption and hypercementosis* (EOTRH)
- 22 videos, recorded after eating a carrot
 - 1-7 days before surgery
 - 6-8 weeks after surgery
 - Annotated by blinded expert (Karina Glerup)
- 3 free-tier Google Gemini models
 - gemini-3-flash-preview
 - gemini-2.5-flash
 - gemini-2.5-flash-lite
- Assessment of settings' effect on Accuracy (see results)

Results

Effects of settings on EPF classification accuracy

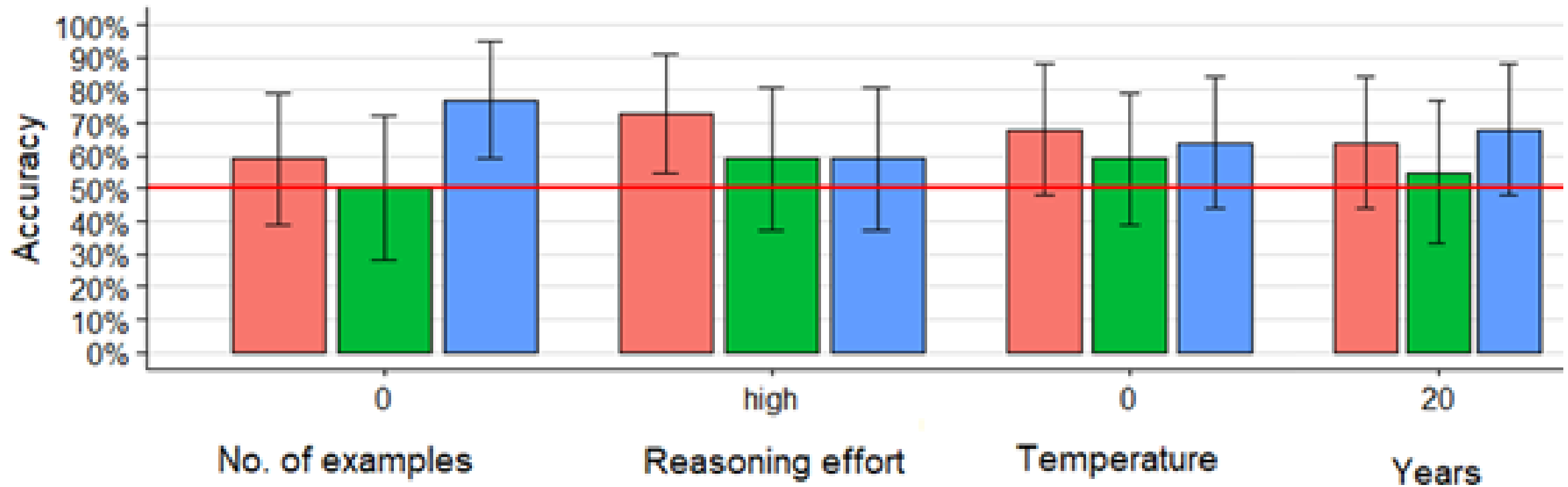
Best:
~85 % accuracy



Default settings: Examples = 0, Reasoning effort = high, Temperature = 0, Years = 20

Results

Repeated runs with same settings are not consistent!



Default settings: Examples = 0, Reasoning effort = high, Temperature = 0, Years = 20

Results

Token costs

	Million tokens		
	Input	Cache	Output
gemini-2.5-flash	0.6	0.4	0.04
gemini-2.5-flash-lite	0.5	0.5	0.05
gemini-3-flash-preview	0.2	0.1	0.03

Company	Model	Price per million tokens (Euros)			Estimated cost of experiment (Euros)	
		Input	Cache	Output	Min	Max
Anthropic	Opus4.8	0.83	1.04	4.17	0.41	1.18
	Sonnet 4.6	0.50	0.63	2.50	0.25	0.71
	Haiku4.5	0.17	0.21	0.83	0.08	0.24
OpenAi	GPT-5.5	0.67	0.07	4.00	0.25	0.62
	GPT-5.4	0.33	0.03	2.00	0.12	0.31
	GPT-5.4 mini	0.10	0.01	0.60	0.04	0.09
Deepseek	Deepseek chat	0.01	0.04	0.15	0.01	0.03
	Deepseek reasoners	0.02	0.07	0.29	0.02	0.06
Google	Gemini 3.5 Flash	0.20	0.13	1.20	0.09	0.24
	Gemini 3.1 Pro Preview	0.53	0.05	2.40	0.17	0.46
	Gemini 3.1 Flash-Lite	0.03	0.00	0.20	0.01	0.03
Total					1.46	3.97

Conclusions

- Existing large language models **may** be useful for assessing signs of pain in horses
 - Not gemini-2.5-flash-lite!
 - Free-tier Google Gemini models are not reliably consistent
- Specific settings have little to no impact on performance
- Videos work better than still images!