Project Gaia

WHITE PAPER · V0.1

ROUND 1 PROTOCOL — THE METHOD, FIXED BEFORE TRAINING BEGAN; MODELS AND BENCHMARK RELEASED 12 NOV 2026

A small AI that knows when to say "ask your vet".

Building a small, specialised model that can be trusted on UK pet-care rules — and knowing where it must not be. Round 1 of an ongoing programme.

4
SYSTEMS COMPARED ON ONE BENCHMARK

3
MEASURES — ACCURACY, HEDGING, REFERRAL

2
ROUND 1 MODELS TRAINED — ONE AND THREE BILLION PARAMETERS

4,000
GPU HOURS AWARDED FOR ROUND 1

A research project of MyPetSitter Ltd · registered in England & Wales, company no. 17063676 · Harrogate, UK
AI research supported by UK Research and Innovation — a research award of computing time, not an endorsement of any product.
projectgaia.co.uk · research@mypetsitter.org · September 2026, updated 12 November 2026

SUMMARY

What this paper is

Round 1 protocol — the method, fixed before training began; models and benchmark released 12 Nov 2026

The UK pet-care industry is tens of thousands of micro-businesses with nobody to check an answer. General AI models answer their questions fluently and are wrong in ways that harm animals and breach licences. Project Gaia is building a small model that is right about this subject and honest about its limits, with the people who do the work reviewing what it learns. This paper describes ROUND 1: train small language models, evaluate them against a large commercial model on an open benchmark, and publish where the small model is good enough and where it is not. It stated the method before any training began, so that the result cannot be shaped to fit. Round 1 runs on awarded computing time — the acknowledgement on the last page names the funder, and it is named there once. The rounds after it will be supported by whatever programme fits them, and the project does not end with any of them.

What is already true

Three small preliminary models were trained before the award; in Round 1, a one-billion- and a three-billion-parameter model since (17 and 29 October 2026). A production assistant and an evaluation harness exist, and one measured finding — that instructing a model does not make it reliable — shaped the whole design.

What is a plan

Still to come: teaching the models to hold a conversation, the evaluation of their answers, the benchmark and the comparison. Their results go into version 1.0; until then this paper claims none.

Contents

1 The problem
2 What we measured before asking for compute
3 The research question
4 The design under test
5 Four systems, three measures
6 Preliminary work, complete
7 Data, and where it comes from
8 Round 1: the work programme
9 People: the contributor programme
10 Safety and ethics
11 What Round 1 publishes
12 How we would know we were wrong
13 Round 2: what continues after this
14 Who is doing this, and how to take part
15 Acknowledgements

1 · PROJECT GAIA

The problem

Dog walkers, sitters, groomers, trainers, daycare and boarding operators are mostly sole traders. They need dependable answers across three areas at once: animal welfare and husbandry practice; regulatory obligations — the Animal Welfare Act, DEFRA licensing under the 2018 Regulations, council permits, the law on livestock worrying; and the daily running of a business. They also need to know when a question is clinical and belongs with a veterinary surgeon.

General-purpose models answer all of it fluently. They import rules from other countries, state figures that were never in the guidance, and give confident advice about unwell animals. The rules themselves differ by activity, by species and by nation of the UK, and they are revised without notice. A model that memorised last year's guidance is wrong this year, and nothing in its tone says so.

> **The harm is not a vague answer. It is a confident, wrong one — given to somebody with no way of checking it.**

2 · PROJECT GAIA

What we measured before asking for compute

MyPetSitter Ltd operates a software platform used by these businesses, with an AI assistant already in production and an evaluation harness that scores it. That gave us something most proposals lack: evidence from real use.

One finding shaped the whole design. A behavioural contract — a rule about how an answer must be given — was stated explicitly in the system prompt, and was breached by both a small and a large commercial model on every trial, across repeated runs. That behaviour is now enforced in code rather than by prompting.

> **Instructing a model does not produce reliable behaviour. So this project separates what is TRAINED from what is ENFORCED, and measures each.**

3 · PROJECT GAIA

The research question

We call it the substitution frontier: at what scale can a small, domain-specialised, self-hosted model replace a metered frontier API in a domain where wrong answers harm animals — and where must it not?

It matters twice over. A small company pays a foreign provider per token on every call; and its customers are sole traders for whom metered pricing is itself the barrier to using AI at all. A model that runs on commodity hardware can also run where data cannot leave the premises. If the small model is good enough for part of the work, that part becomes affordable and local. If it is not, the honest conclusion is to keep renting — and the evidence for that is just as useful to other UK small businesses.

4 · PROJECT GAIA

The design under test

The design is deliberately unusual, and each part of it answers a specific way these systems fail.

Principle	What it means	The failure it prevents
The model learns the SHAPE of a good answer	Cite the source, give the date it was last reviewed, name the jurisdiction, defer to the authority	A fluent answer with nothing to check it against
Facts are retrieved, never memorised	Regulatory and welfare content is fetched from authoritative sources at answer time	A model that is silently out of date the day guidance changes
Clinical questions are referred by construction	Routing to a veterinary surgeon is built into the system, not left to the model's judgement	A diagnosis, a medicine or a dose from software
An unsafe action is absent, not forbidden	Sending a message, changing a record or cancelling a booking are not in the set the model can choose from	A prompt rule a differently-worded request gets around
The model drafts; a person sends	Drafting and sending are separate code paths and will remain so	An automated message nobody read

The system also says plainly that it is an AI whenever it is asked, and offers a human.

5 · PROJECT GAIA

Four systems, three measures

The same held-out question set is run against four systems, producing the frontier for this domain rather than a single score.

System	Why it is in the comparison
A large commercial model	The thing a small company rents today — the baseline to beat, or to keep
An open model, as released	Shows what specialisation adds, by its absence
The same open model, fine-tuned	The cheapest route to a specialised model
A model trained from scratch	Shows what is possible when tokenizer, data and size are all chosen for the domain

Measure	How it is scored
Retrieval accuracy	Does the answer rest on the right rule, for the right activity, species and nation?
Hedging compliance	An unhedged answer about a licensing obligation is scored a **FAILURE** even when it is factually correct
Veterinary referral	Does every clinical question end with a vet — with no diagnosis, medicine or dose on the way?

Scoring is repeated, because model output varies run to run. A rule counts as held only when it holds on every run.

The small model is enough

Questions it answers as reliably as the large one. This work can be done locally, at a fixed cost, where data cannot leave the premises.

Still needs the large model

Questions where the small model is measurably worse. Keep renting for these — knowingly, and only for these.

Belongs to a person

Clinical questions and irreversible actions. No model of any size answers these; the system refers or stops.

What the project produces is not a score. It is where these two lines fall — for this domain, with the evidence published.

6 · PROJECT GAIA

Preliminary work, complete

Three models have already been trained end to end on consumer and free-tier hardware. The transformer, the byte-level BPE tokenizer, the data pipeline and the training loop were written from scratch; the training code is device-agnostic, resumable, and instrumented to measure how many tokens each language costs.

Model	Trained on	Purpose
1.05 million parameters	A fanless laptop	Proved the pipeline end to end
12.6 million parameters	A free cloud GPU (T4)	Proved resumable training and instrumentation
53.9 million parameters	918 million tokens of FineWeb-Edu	Proved the data pipeline at scale

These are not useful models. They are the evidence that the awarded compute will be spent on research rather than on learning how to train.

7 · PROJECT GAIA

Data, and where it comes from

Data	Source and licence	Used for
Domain corpus — welfare, regulation, practice	Authored and reviewed by people, or published under an open licence (Open Government Licence, Creative Commons) or with written permission. Never scraped. Each record carries its source, jurisdiction and review date.	Retrieval; post-training examples
General text	FineWeb-Edu (English) and FineWeb-2 (Romanian), openly published datasets derived from Common Crawl, under ODC-By 1.0 — the pet-care pages of FineWeb-Edu read more often than the rest; for the three-billion model also openly licensed code (Common Pile's Stack v2 Edu, each file under its own permissive licence) and mathematics (FineMath, ODC-By 1.0)	Pre-training
Conversation	Openly licensed conversations written by people; and, for conversational style only, openly licensed AI-written conversations and conversations we generated with an open model, every pet-care topic screened out of them in English and Romanian. Each source is on the data card with its licence and who wrote it.	Teaching the model to hold a conversation
Evaluation phrasings	Real questions from the platform, pseudonymised before they leave it: names, addresses, contact details and booking identifiers stripped; only the wording kept	The held-out benchmark

- No customer data is uploaded to the research computer, and no model is trained on customer records.
- No Project Gaia research model is trained on pet-care answers written by another AI model.
- AI-written conversations teach only how to hold a conversation: every pet-care topic is screened out of them, and the data card names each source.
- Every dataset and its licence is recorded in a data card, published with the weights on 12 November 2026.

7 · DATA, AND WHERE IT COMES FROM — CONTINUED

One bias is measured rather than assumed. On a single identical sentence, an English-first tokenizer cost 3.32 times more tokens for Romanian than for English, and even a 151,643-token multilingual vocabulary cost 2.00 times more — early one-sentence figures, which overstate it. Measured properly in Round 1, on 20 MB of held-out text in each language, the cost was 1.80–1.86 times for an English-only vocabulary and 1.52–1.82 times for widely used published ones (see What we've learned). Either way it is a real disadvantage imposed on people who do not write in English. Reducing it is an explicit objective, and results are reported per language.

8 · PROJECT GAIA

Round 1: the work programme

The project has been awarded 4,000 GPU hours on Isambard-AI. The allocation was planned as follows, before any of it was used.

Stage	GPU hours	What it decides
Tokenizer ablation	60	Which vocabulary is fairest and cheapest across languages
Data-mix ablation	90	How much domain text the general text can carry
Main model — one billion parameters	450	The from-scratch system in the comparison
Post-training	200	Hedging and referral behaviour, taught from written examples
Evaluation and inference	300	All four systems, three measures, repeated runs
Stretch — three billion parameters	2,400	Attempted only if the week-five evidence justifies it
Contingency	500	Failed runs, restarts, a busy machine
Total	**4,000**	

The week-five decision, taken on 18 October 2026 — a week early. The one-billion model had trained cleanly; the larger model fitted its 2,400 hours with a wide margin at 150 billion tokens; more than six weeks remained; and the post-training data was on course. So the three-billion model was attempted. It trained from 18 to 29 October 2026, on 150 billion tokens.

Where the hours stand, 12 November 2026: 3,570.4 of the 4,000 GPU hours, on the machine's own record. The one-billion model's main run took 373.6 (13–17 October), the three-billion model's 1,706.8 (18–29 October) and its first longer-reading phase 48.7; the rest went on the ablations, two more longer-reading phases (4–6 November), the conversation stage (from 8 November) and the evaluations. All four systems were evaluated on the same questions — scored by machines and an openly licensed AI judge, never confirmed by people — and on 12 November 2026 Round 1's models and its benchmark were released. What remains is held for the results in version 1.0 of this paper and for anything that slips.

The compute is the cheaper half of the project. The work that makes it worth anything is done by hand: 450 hours of the Project Lead's time over the three months.

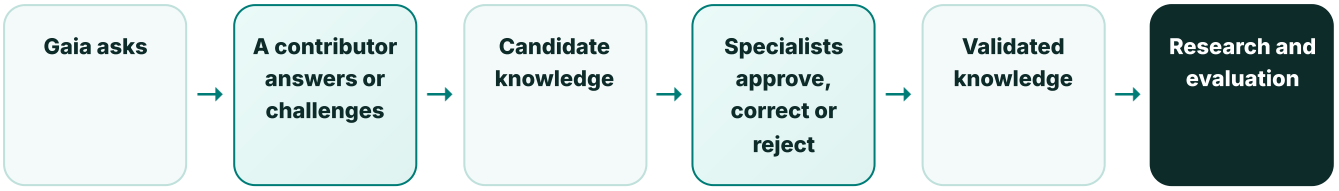
Hand-built work	Hours
Corpus curation and licence provenance	90
Ablation design and analysis — a GPU hour spent on an ablation nobody analyses is wasted	60
Training supervision: checkpointing, instrumentation, failure diagnosis, restarts	120
Post-training data: hedging and referral examples have to be written, not sampled	80
Evaluation harness and judging, across four systems and three measures	70
Publication: model cards, data card, benchmark, comparison	30
Total	**450**

The allocation was sized well below the maximum on purpose: a first award should be judged on delivery. A longer-term goal of a seven-billion-parameter model is something these ablations would justify, or rule out, before any larger request is made.

People: the contributor programme

This section describes a programme that runs alongside the awarded compute. It is a plan of MyPetSitter Ltd, not a commitment made in the award.

Gaia is being shaped by the industry she is for. People from every branch of pet care are invited to take part as volunteers — veterinary professionals, licensing officers, first-aiders, behaviourists, groomers, and the people who walk, board and care for animals every day. The unusual part is the direction of the conversation: instead of people asking the AI, Gaia asks them. What would you do here? What does the guidance leave out? What should an AI never get wrong in your field?



An answer does not become knowledge by being submitted, nor by being popular. Each item carries its state, so its history can always be traced:

State	Meaning
Submitted	One person has said it
Community consensus	Several contributors independently agree
Expert reviewed	A specialist in that field has looked at it
Validated	Approved — veterinary and legal items need two reviewers, or one member of the Project Gaia Board, and disagreement is recorded, not averaged
Eligible for the research dataset	Validated, correctly licensed by its author, and free of personal data
In use	Recorded against the exact evaluation or training run that used it

- An approval or a rejection of something Gaia said becomes an evaluation label. Only a person's own words become a knowledge record.
- Taking part is voluntary and unpaid. Recognition is by name, only for those who choose it, and records what a person did — questions answered, items validated — never a qualification they do not hold.
- Before anybody contributes, the way their contribution is used and credited is agreed in writing.
- Veterinary professionals are asked for advice and review, not endorsement.

10 · PROJECT GAIA

Safety and ethics

Question	Our answer
Personal data?	Limited and controlled. The corpus contains none. Evaluation phrasings are pseudonymised before leaving the platform. Nothing personal goes to the research computer.
Bias or harm from data?	Web-derived data carries the biases of the web; the data card names every dataset, and the models make no decision about a person. Tokenisation bias is measured and reported per language.
Impact on people?	The credible harm is a confident wrong answer about a legal or welfare duty. Retrieval instead of memory, mandatory hedging, and referral by construction address it.
Decisions about people?	It influences and is designed not to decide. Every output is a draft a human confirms, edits or discards.
Misuse of open weights?	Accepted as the cost of open publication, and limited by size and narrowness: under three billion parameters, no capable general reasoning, no autonomous action released. The model card states that it must not be relied on for legal or veterinary advice.

11 · PROJECT GAIA

What Round 1 publishes

- An open benchmark for pet-care professional question answering: held-out phrasings paired with authoritative answers, scored on the three measures, so anybody can run it against any model.
- Model weights under the Apache-2.0 licence, with model cards stating limitations and failure modes.
- A data card naming every dataset and its licence.
- The comparison across all four systems, whatever it shows — so that other small companies can judge self-hosting on evidence rather than on vendor claims.
- This paper, revised as version 1.0 with the results.

Released on 12 November 2026, earlier than planned, on Hugging Face (huggingface.co/ProjectGaiaUK): the weights of her chat model and of the base model it was made from, under Apache-2.0, with their model cards and the data card; and the benchmark, v0.1, with every system's scores. The full comparison, and this paper as version 1.0 with the results, are planned for the first week of December — inside the round, which ends on 21 December 2026.

Intellectual property in the work belongs to MyPetSitter Ltd. The method is intended to generalise to any regulated micro-business sector facing the same narrow, frequently revised guidance.

12 · PROJECT GAIA

How we would know we were wrong

- If the specialised models do not beat the un-tuned open model on retrieval accuracy, the corpus is not doing the work we think it is.
- If they beat it on accuracy but fail the veterinary rule more often, specialisation has made them bolder — which is worse, and we will say so.
- If the commercial model wins on all three measures at every size, the honest conclusion is that a small company should rent rather than build.
- If the fairer tokenizer costs more accuracy than it saves in tokens, that trade-off is published with its numbers rather than hidden.

Round 2: what continues after this

This paper is versioned on purpose. Round 1 ends with version 1.0 — the same document with the results in it. What follows is written the same way: stated before it starts, published whichever way it goes.

Three things continue no matter what Round 1 concludes, because none of them depends on the answer:

- The knowledge base — records written and checked by people who do the work, and which go stale without them. This is the asset that survives every model we train, and the one thing here that cannot be regenerated from anything else.
- The contributors. A professional who has answered questions for three months has not finished knowing things, and the rules they work to change every year.
- The benchmark. A public test is only useful if somebody keeps running it — against new models, including ones nobody has released yet.

Round 1 is the public demonstration of the method, on a national research computer, with the results published whichever way they land. The intake behind it — people who do the work contributing what they know, checked by specialists and kept with its provenance — belongs to MyPetSitter Ltd; it began before this round and carries on after it. The two feed each other: the round tests what the intake has produced so far, and what the round learns goes back into it. The benchmark and the weights are published. The knowledge base is the company's, as the Contributor Agreement says.

13 · ROUND 2: WHAT CONTINUES AFTER THIS — CONTINUED

What Round 2 is for depends on what Round 1 finds.

If Round 1 finds	Round 2 is
The small model is good enough for a defined part of the work	Hardening it for that part: more evaluation, more coverage, and the self-hosting comparison run in earnest
It reaches its limit sooner than hoped	Finding where the line actually falls, and what a larger or differently-trained model would need to cross it
The retrieval design is doing most of the work	Improving the knowledge base and the retrieval, which helps every model we might use — including one we rent

Support for Round 2 will be sought from whichever programme fits the work. None of it changes what has already been published, what contributors were promised, or the rule that a result is published whichever way it lands.

14 · PROJECT GAIA

Who is doing this, and how to take part

Project Gaia is carried out by MyPetSitter Ltd (company no. 17063676), Harrogate, UK — a micro-company whose software is used by licensed pet-care businesses. The work is led within the company. It will be conducted in the UK and aims to deliver its benefits in the UK.

If you know pet care from the inside, the project needs you: projectgaia.co.uk/contribute. The project is not crowdfunded and asks nothing of individuals; projectgaia.co.uk/supporters sets out exactly where the computing time, the credit and the hours come from, and how an organisation can sponsor the work. Contact: research@mypetsitter.org.

15 · ACKNOWLEDGEMENTS

Supported by

AI research supported by

UK Research and Innovation

UKRI AI Research Resource (AIRR) Gateway · project 0261-1548-0489-1 · 4,000 GPU hours of computing time.

This project will use resources provided by Isambard-AI. Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

The award supports the research. It is not an endorsement of any product.

With the support of Teds.dog. Sponsorship buys this mention and nothing else — never a result, a question, or a say in what is published.