Foundation Model Scope 3 Emissions Calculation Advisory

Guide for Corporate Carbon Accounting

1 Executive Summary

This advisory provides a practical framework for calculating Scope 3 emissions from Foundation Model services across text, audio, video, and music generation for corporate carbon accounting. As external services, foundation model providers' operational and embodied emissions become your organisation's Scope 3 emissions under the GHG Protocol.

1.1 Key Findings by Modality:

- Text: 0.05 to 0.8 gCO2e per 1,000 tokens
- Audio: 0.1 to 2.0 gCO2e per minute processed
- Images: 15-25 gCO2e per image generated
- Video: 800-2,000 gCO2e per minute generated
- Music: 200-500 gCO2e per song generated

1.2 Default Recommendation:

Use **0.4 gCO2e per 1,000 tokens** for text as baseline, with modality-specific multipliers for mixed usage.

2 Framework Overview

2.1 Core Components

The following 3 elements for Foundation Model Scope 3 emissions have been considered:

- Operational Emissions - Energy consumed during inference/generation
- Embodied Emissions - Infrastructure manufacturing and construction
- Training Attribution - Amortised training costs across model lifetime

2.2 Standard Formula by Modality

Text Models:

```
Total Scope 3 = (Tokens × 0.4 gCO2e/1K tokens × Location Multiplier)
+ Training Attribution
```

Multimodal Models:

```
Total Scope 3 = (Usage × Modality Factor × Location Multiplier) +
Training Attribution
```

Default Calculation Framework:

```
Total Scope 3 = (Usage Units × Emissions Factor × 1.0) + (Total ×
0.2)
```

3 Component 1: Foundation Model Type and Size Impact

3.1 Emissions Factors by Modality

Modality	Usage Unit	Energy Range	Emissions Factor	Default
Text (LLM)	1,000 tokens	0.2-2.5 Wh	0.05-0.8 gCO2e	**0.4 gCO2e**
Audio Processing	Per minute	0.5-8 Wh	0.1-2.0 gCO2e	**0.8 gCO2e**
Image Generation	Per image	50-100 Wh	15-25 gCO2e	**20 gCO2e**
Video Generation	Per minute	3,000-8,000 Wh	800-2,000 gCO2e	**1,200 gCO2e**
Music Generation	Per song (3-4 min)	800-2,000 Wh	200-500 gCO2e	**350 gCO2e**

3.2 Text Model Size Categories

Model Category	Parameters	Examples	Emissions Factor (gCO2e/1K tokens)	Default Usage
Large Models	70B+	GPT-4, Claude-3 Opus	0.6-0.8	0.8
Medium Models	7-50B	GPT-3.5, Claude-3 Sonnet	0.2-0.6	**0.4**
Small Models	1-7B	Gemma-2B, Quantised models	0.05-0.2	0.1

Default Assumption: Most enterprise usage involves medium-sized text models, with growing multimodal usage.

3.3 Audio Model Examples

- **Speech-to-Text (Whisper):** 0.1-0.3 gCO2e per minute
- **Text-to-Speech (Standard):** 0.4-0.8 gCO2e per minute
- **Text-to-Speech (HD):** 0.8-1.5 gCO2e per minute
- **Audio Generation:** 1.0-2.0 gCO2e per minute

3.4 Visual Model Examples

- **Image Generation (DALL-E):** 15-25 gCO2e per image

- **Image Generation (Midjourney):** 12-20 gCO2e per image
- **Video Generation (5-sec, 16fps):** 200-400 gCO2e per generation
- **Video Generation (1-min, 30fps):** 1,200+ gCO2e per generation

4 Component 2: Geographic Location Impact

4.1 Regional Carbon Intensity Multipliers

Region Type	Examples	Carbon Intensity	Multiplier
Low-Carbon	Nordic countries, Quebec, renewable-heavy	<200 gCO2/kWh	0.5x
Average	US/EU average	300-400 gCO2/kWh	**1.0x**
High-Carbon	Coal-heavy grids	>500 gCO2/kWh	2.0-3.0x

Default Assumption: Most commercial LLM providers operate in regions with average carbon intensity. **Recommended Multiplier: 1.0x (no adjustment)**

4.2 Provider-Specific Considerations

- Google Cloud: Offers region selection tools, generally lower carbon intensity
- Microsoft Azure: Mixed carbon intensity, provides sustainability calculators
- AWS: Limited regional transparency, use average multipliers

Note: Provider location data is often unavailable. Default to average regional factors unless specific provider data exists.

5 Component 3: Training Cost Attribution

5.1 Training Emissions Data

Model	Training Emissions	Amortisation Method
GPT-3	552 tonnes CO2e	Usage-based allocation
BLOOM	25 tonnes CO2e	Geographic efficiency
LLaMA-2	1,999 tonnes CO2e	Parameter scaling

5.2 Attribution Methodology

Three Approaches:

- Simple Percentage: Add 20% to operational emissions
- Usage-Based: Allocate based on proportion of total model usage
- Time-Based: Amortise over expected model lifetime (e.g. 3-5 years)

Default Assumption: Training costs represent approximately 20% of total lifecycle emissions for heavily-used models. **Recommended Addition: 20% of operational emissions**

5.3 Default Training Calculation

```
Training Attribution = Operational Emissions × 0.2
Total Emissions = Operational + Training = Operational × 1.2
```

6 Implementation Guide

6.1 Step 1: Data Collection

Required Data:

- Total token usage across all LLM services
- Service provider and model types used
- Geographic regions (if known)
- Annual service costs (for validation)

Tracking Methods:

- API logs and token counters
- Provider billing statements
- Internal usage analytics

6.2 Step 2: Model Classification

Default Classification:

- **Commercial APIs (ChatGPT, Claude, Gemini)** → Medium Models
- **Specialised/coding models** → Large Models
- **Edge/mobile applications** → Small Models

6.3 Step 3: Calculate Operational Emissions

```
Operational = Token Usage × Emissions Factor × Location Multiplier

Default: Tokens × 0.4 gCO2e/1K tokens × 1.0
```

6.4 Step 4: Add Training Attribution

```
Training = Operational × 0.2
Total = Operational + Training = Operational × 1.2
```

6.5 Step 5: Validation Check

Spend-Based Validation:

- Annual LLM costs × 0.4-0.6 kg CO2e/USD
- Compare with token-based calculation
- Use higher value for conservative reporting

7 Worked Examples

7.1 Example 1: Medium Enterprise Usage

Scenario:

- 50 million tokens annually
- Mix of GPT-3.5 and Claude-3 Sonnet
- Standard commercial regions

Calculation:

Operational = 50M tokens × 0.4 gCO2e/1K tokens = 20,000 gCO2e = 20 kg CO2e
Training = 20 kg × 0.2 = 4 kg CO2e
Total = 20 + 4 = 24 kg CO2e

7.2 Example 2: Heavy AI-First Company

Scenario:

- 500 million tokens annually
- Primarily GPT-4 and Claude-3 Opus
- Some usage in low-carbon regions

Calculation:

Large Model Usage (80%): 400M × 0.8 gCO2e/1K = 320 kg CO2e
Medium Model Usage (20%): 100M × 0.4 gCO2e/1K = 40 kg CO2e
Regional Adjustment (20% low-carbon): Total × 0.9 = 324 kg CO2e
Training Attribution: 324 × 0.2 = 65 kg CO2e
Total = 324 + 65 = 389 kg CO2e

7.3 Example 3: Cost-Optimised Deployment

Scenario:

- 200 million tokens annually
- Primarily smaller models (Gemma, quantised)
- Edge and on-premise deployment

Calculation:

Operational = 200M × 0.1 gCO2e/1K = 20 kg CO2e
Training = 20 × 0.2 = 4 kg CO2e
Total = 20 + 4 = 24 kg CO2e

8 Conclusion: Emission Ranges by Usage Profile

8.1 Summary Table

Usage Profile	Annual Tokens	Model Type	Total Emissions Range	Default Calculation
Light User	<10M	Medium	2-8 kg CO2e	**5 kg CO2e**
Medium User	10-100M	Medium	5-50 kg CO2e	**25 kg CO2e**
Heavy User	100M-1B	Mixed	50-500 kg CO2e	**250 kg CO2e**
AI-First Company	1B+	Large	500+ kg CO2e	**1,000 kg CO2e**

8.2 Quick Reference Factors

For Rapid Estimation:

- **Conservative Default:** 0.5 gCO2e per 1,000 tokens (includes all components)
- **Optimised Deployment:** 0.15 gCO2e per 1,000 tokens
- **Large Model Heavy Usage:** 1.0 gCO2e per 1,000 tokens

8.3 Annual Spend Validation

Alternative Calculation:

- **Annual LLM Costs × 0.5 kg CO2e/USD** (conservative baseline)
- Use for validation against token-based calculations
- Report higher value for conservative Scope 3 accounting

9 Recommendations

9.1 For Scope 3 Reporting

1. **Use token-based calculation** with default factors as primary method
2. **Validate with spend-based calculation** using EPA USEEIO factors
3. **Document methodology clearly** including assumptions and data limitations
4. **Plan annual updates** as provider transparency improves
5. **Consider geographic optimisation** for large users

9.2 For Reducing Emissions

1. **Optimise model selection** - use smallest effective model size
2. **Implement geographic routing** - choose low-carbon regions when possible
3. **Batch processing** - combine requests to improve efficiency

4. **Quantisation and optimisation** - use compressed models where appropriate
5. **Monitor usage patterns** - identify and reduce unnecessary queries

9.3 Data Quality Improvements

- Request provider disclosure of region-specific emissions factors
- Track model-specific usage rather than aggregated metrics
- Implement real-time monitoring for large-scale deployments
- Collaborate with providers on sustainability initiatives

10 Appendix: Uncertainty and Limitations

10.1 Key Uncertainties

- Provider transparency gaps - most don't disclose detailed emissions
- Regional allocation - many providers don't specify data center locations
- Training cost attribution - no standard methodology exists
- Technology improvements - efficiency gains may not be reflected in static factors
- Users may switch to newer models sooner than the lifetime of the model, so the 20% flat addition may be more appropriate

10.2 Conservative Approach

This framework errs on the side of higher estimates to ensure adequate Scope 3 coverage. Actual emissions may be 20-50% lower with optimised configurations and efficient providers.

10.3 Annual Review

Update factors annually as:

- Provider transparency improves
- Technology efficiency advances
- Regional energy mixes evolve
- Industry standards mature

Document Version: 0.01
Last Updated: June 2025

alan@goalwrangler.com
Framework based on GHG Protocol, EPA USEEIO, and current academic research

11 Appendix A: Token Estimates

Typical token consumption patterns for text-based LLM usage (excluding code development):

11.1 Token Consumption Patterns by Task Type

Basic Text Generation & Short Queries

- **Prompt:** 10-50 tokens (simple questions, basic requests)
- **Response:** 50-200 tokens (brief answers, explanations)
- **Total per interaction: 75-250 tokens**
- **Examples:** "Explain X", "What is Y?", "Summarise this briefly"

Research & Analysis Tasks

- **Prompt:** 50-200 tokens (detailed questions with context)
- **Response:** 200-800 tokens (comprehensive explanations)
- **Total per interaction:** 300-1,000 tokens
- **Examples:** Research questions, comparative analysis, detailed explanations

Document Summarisation

- **Prompt:** 1,000-3,000 tokens (includes source text)
- **Response:** 100-500 tokens (summary output)
- **Total per interaction: 1,200-3,500 tokens**
- **Examples:** A pretty short question with a one-paragraph answer may cost 3-4 thousand tokens when using context-heavy approaches

Writing Assistance & Editing

- **Prompt:** 200-800 tokens (instructions + content to edit)
- **Response:** 300-1,000 tokens (revised content)
- **Total per interaction: 500-1,800 tokens**
- **Examples:** Proofreading, rewriting, style improvements

11.2 Key Conversion References

Token-to-Text Ratios:

- 1 token ~= 4 chars in English
- 100 tokens ~= 75 words
- Wayne Gretzky's quote "You miss 100% of the shots you don't take" contains 11 tokens

Practical Example:

- The total token count for the prompt and essays reached 3,747, with over 2,600 words and 17,258 characters

11.3 Recommended Defaults for Scope 3 Calculations

Conservative Estimates by Usage Type:

Task Type	Typical Tokens per Interaction	Recommended Default
Simple Queries	75-250 tokens	**150 tokens**
Research Tasks	300-1,000 tokens	**500 tokens**
Document Work	1,200-3,500 tokens	**2,000 tokens**
Writing Tasks	500-1,800 tokens	**1,000 tokens**

Mixed Usage Default

For organisations with varied text-based LLM usage, use **750 tokens per interaction** as a conservative baseline.

Annual Token Estimation

Calculate your annual usage:

```
Annual Tokens = (Interactions per day × Average tokens per interaction × 365 days)
```

Example calculations:

- **Light user:** `10 interactions/day × 500 tokens = 1.8M tokens/year`
- **Medium user:** `25 interactions/day × 750 tokens = 6.9M tokens/year`
- **Heavy user:** `50 interactions/day × 1,000 tokens = 18.3M tokens/year`

11.4 Practical Tracking Recommendations

- **Monitor actual usage** through API logs or provider dashboards
- **Track by task type** to refine your estimates
- **Account for conversation context** - each subsequent follow-up message in the same conversation will send the entire chat history (within the context window) to the LLM
- **Consider efficiency improvements** like batching requests and optimising prompt length

Key Insight: Most text-based business usage falls in the **500-1,000 tokens per interaction** range, making this a reasonable default for Scope 3 calculations when detailed usage data isn't available.