

A Appendix

A.1 Related Work

A wide variety of works seek to examine the mechanisms of generative AI systems. Many initially focused on attributions in input space [8, 9], with later approaches focused on discovering circuit and graph structures [3, 4, 5], as well as self-explanations generated by the model [10, 11]. To extract relevant representations and model features, a variety of works have explored interpretability for analyzing concepts [68], representations [69], and clustering [70] in the context of attributions of sentence summaries [71] and concept discovery in LLMs [72]. In the following, we will focus specifically on studies and methods centered on feature descriptions.

A.2 Feature Description Methods

One of the earliest works on feature description in language models is SASC (Summarize and Score) [73], which generates natural language descriptions of neurons in a pre-trained BERT model. Shortly thereafter, an automated interpretability method for describing all neurons in GPT-2 XL was proposed [20]. This approach analyzes the textual patterns that cause a neuron to activate, and uses GPT-4 as *explainer model* to generate a description of the neuron’s function. Given a set of token-activation pairs derived from text excerpts and corresponding neuron activations, the *explainer model* identifies common patterns, based on which it generates a textual description of the neuron’s role. This method has since been widely adopted and further developed, forming the basis for many subsequent methods targeting both individual neurons [22, 24] and SAE features [25, 18, 21, 27, 26, 42, 28, 24]. An overview of representative feature description methods is provided in Table 2.

Table 2: Representative feature description methods for language models, listed in chronological order, including the model and feature types they target.

Method	Target Model	Feature Type
SASC [73]	BERT	neuron
GPT-Explain [20]	GPT-2 XL	neuron
Pythia SAE [25]	Pythia-70M and Pythia 410-M	SAE feature
Anthropic SAE [18]	one-layer transformer	SAE feature
GPT-2 SAE [21]	GPT-2 Small	SAE feature
GPT-4 SAE [21]	GPT-4	SAE feature
EleutherAI SAE [27]	Llama 3.1 7B & Gemma 2 9B	SAE feature
Transluce-Explain [22]	Llama 3.1-8B Instruct	neuron
Llama Scope [26]	Llama 3.1 8B Base	SAE feature
Gemma Scope [42]	Gemma 2	SAE feature
Goodfire SAE [28]	Llama 3.3 70B	SAE feature
Output-Centric neuron [24]	Llama 3.1 8B Instruct	neuron
Output-Centric SAE [24]	Gemma 2.2B, GPT-2 Small, Llama 3.1 8B	SAE feature

A.3 Description Scoring Details

Figure 6 illustrates the COSY evaluation procedure [38] as adapted for language models. As the control dataset \mathbb{X}_0 , we use a subset of 1,000 randomly sampled entries from Cosmopedia [74]. For each candidate description of a target feature, we use Gemini 1.5 Pro [44] to generate 10 concept-specific text samples, each with a maximum length of 512 tokens. These samples form the concept dataset \mathbb{X}_1 . The generation prompt is shown in Figure 7. We then pass both datasets through the model to extract activations corresponding to the target feature. We then use Average Pooling as aggregation function $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}$ to each activation vector to obtain scalar representations:

$$\begin{aligned}\mathbb{A}_0 &= \{\sigma(f_{\ell,i}(\mathbf{x}_1^0)), \dots, \sigma(f_{\ell,i}(\mathbf{x}_n^0))\} \in \mathbb{R}^n, \\ \mathbb{A}_1 &= \{\sigma(f_{\ell,i}(\mathbf{x}_1^1)), \dots, \sigma(f_{\ell,i}(\mathbf{x}_m^1))\} \in \mathbb{R}^m.\end{aligned}\tag{7}$$

The resulting activation distributions \mathbb{A}_0 and \mathbb{A}_1 are compared to compute the CoSY Score (see Equations 5 and 6), which quantifies how accurately a given description captures the target feature. Higher scores indicate clearer separation between concept and control samples, reflecting more precise and informative descriptions.

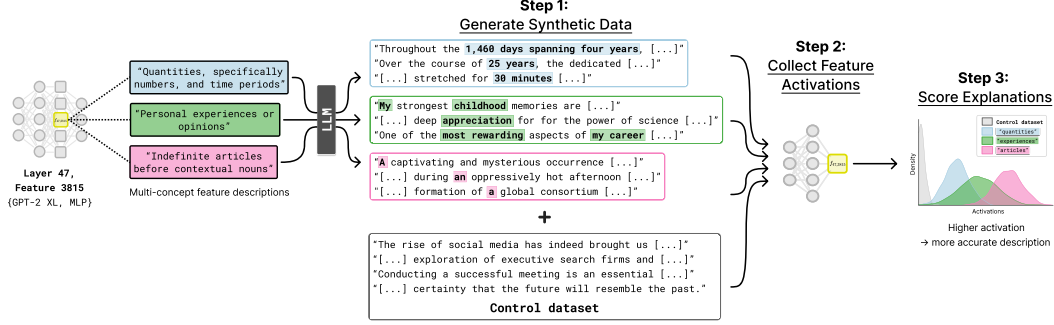


Figure 6: Score feature descriptions with COSY. First, we compile multiple candidate descriptions for a target feature. For each description, we prompt an LLM to generate 10 text samples including the described concept. These concept-specific samples, along with a control set of random text samples, are processed through the model to extract activations for the target feature. The COSY Score quantifies the separation between activation distributions of concept samples versus control samples, enabling objective comparison of different feature descriptions. Higher scores indicate descriptions that better capture the feature’s underlying concept.

Generate 10 sentences with a length of 512 words, one per line, with no additional formatting, introduction, or explanation. Each sentence should be a complete, standalone text sample that can be saved as an individual row in a text file. The sentences should include: {feature_description}

Figure 7: Prompt used to generate concept-specific text samples for evaluation. The placeholder {feature_description} is replaced with a candidate textual feature description before being passed to a large language model (Gemini 1.5 Pro). The model then generates 10 standalone text samples, which form the concept dataset \mathbb{X}_1 . These samples are used to evaluate how well the description aligns with the target feature.

923 A.4 Benchmark Experiment Details

924 **Reference Descriptions.** We use the publicly available GPT-Explain descriptions⁸ and the Output-
925 Centric feature descriptions⁹ as comparison.

926 **Model layers.** For GPT-2 XL¹⁰, we use layers 0, 20, and 40. For Llama 3.1 8B Instruct¹¹, we
927 sample from layers 0, 20, and 30. For GPT-2 Small SAE, we use the original implementation¹²,
928 specifically version 5 with a width of 32k. We select features from layers 0, 5, and 10. For Gemma
929 Scope¹³, we use the residual stream SAE with width 16, selecting features from layers 0, 10, and 20.

930 **Feature selection.** For each model, we randomly choose 60 features, 20 from each of three layers,
931 with available reference descriptions from prior work. The only exception is GPT-2 Small SAE,
932 where only 59 features are annotated in the Output-Centric benchmark.

933 **Prompt for generating descriptions.** To produce textual feature descriptions, we use a prompt that
934 instructs a large language model (Gemini 1.5 Pro) to identify shared concepts across high-activation
935 text excerpts within a cluster. The model receives the top $N_s = 20$ excerpts per cluster, with
936 high-activation token spans highlighted using square brackets. The full prompt is shown in Figure 8.

937 **AUC and MAD distributions.** To better understand the standard deviation observed in our bench-
938 mark results, we provide distribution plots of the evaluation metrics. Figure 9 shows the distribution

⁸<https://github.com/openai/automated-interpretability/tree/main?tab=readme-ov-file#public-datasets>

⁹<https://github.com/yoavgur/Feature-Descriptions>

¹⁰<https://huggingface.co/openai-community/gpt2-xl>

¹¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

¹²https://github.com/openai/sparse_autoencoder?tab=readme-ov-file

¹³<https://huggingface.co/google/gemma-scope>

You are a meticulous AI researcher conducting an important investigation into a specific neuron inside a language model that activates in response to text excerpts. Each text starts with ">" and has a header indicated by === Text #1234 ===, where #1234 can be any number and is the identifier of the text.

Neurons activate on a word-by-word basis. Also, neuron activations can only depend on words before the word it activates on, so the description cannot depend on words that come after, and should only depend on words that come before the activation.

Your task is to describe what the common pattern is within the following texts. From the provided list of text excerpts, identify the concepts that trigger the activation of a particular feature. If a recurring pattern or theme emerges where these concepts appear consistently, describe this pattern. Focus especially on the spans and tokens in each example that are inside a set of [delimiters] and consider the contexts they are in. The highlighted spans correspond to very important patterns.

At the beginning, before the list of texts, there will be a list of the highlighted tokens with their activation values.

At the end, following 'Description:', your task is to write the description that fits the above criteria the best.

Do NOT just list the highlighted words!

Do NOT cite any words from the texts using quotation marks, but try to find overarching concepts instead!

Do NOT write an entire sentence!

Do NOT finish the description with a full stop!

Do NOT mention the [delimiters] in the description!

Do NOT include phrases like 'highlighted spans', 'Concepts of', or 'Concepts related to', and instead only state the actual semantics!

Do NOT start with 'Description:' and instead only state the description itself!

Figure 8: Prompt used to generate textual descriptions of a feature based on its activation patterns. The language model (Gemini 1.5 Pro) is instructed to analyze a set of text excerpts, focusing on highlighted spans corresponding to high activations of a specific feature. The model is guided to identify consistent patterns or concepts that trigger the feature. The resulting output is a concept-level description used as textual feature description.

939 of AUC scores across all evaluated model features, while Figure 10 presents the distribution of MAD
940 scores.

941 **Compute Resources** All experiments were conducted using a single NVIDIA A100 80GB GPU.
942 The description procedure takes approximately 9 minutes per feature, including percentile sampling,
943 clustering, and the generation of 5 descriptions. For evaluation, the generation of 10 sentences per
944 feature requires roughly 3 minutes.

945 A.5 Metalabels

946 In Figure 11 and Figure 12, we provide additional examples of metalabels for GPT-2 XL and GPT-2
947 Small SAE. These resulted from clustering 300 sentence representations (embedder: GPT-2 XL,
948 last-token pooling) of identified feature descriptions for a given model and neurons. We show 20
949 randomly selected samples from a total of $k_m = 50$ meta-clusters that were computed using k-means,
950 along with up to three feature descriptions selected at random. Metalabel descriptions were generated
951 via Gemini 1.5 Pro. Clusters for which no concise label was generated, are labeled with 'N/A'.

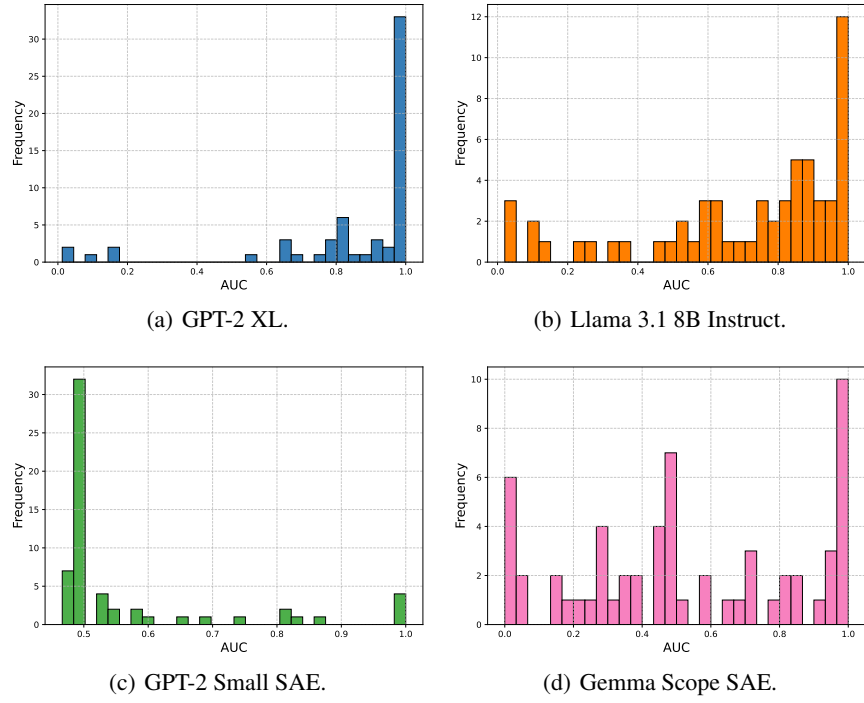


Figure 9: Distributions of PRISM (max) AUC scores across different models and layers.

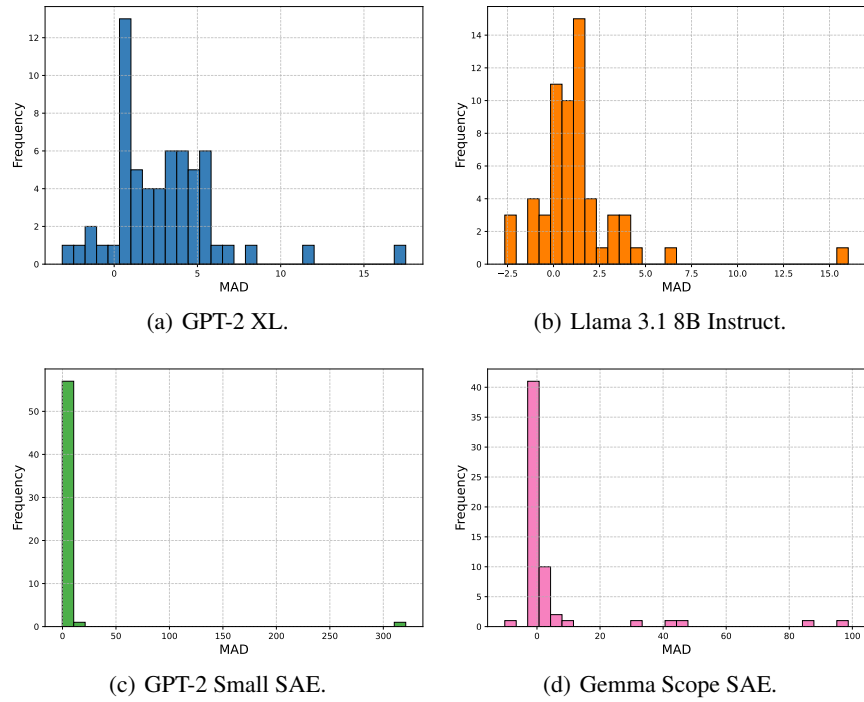


Figure 10: Distributions of PRISM (max) MAD scores across different models and layers.

id	Metalabel	Feature Descriptions
3	Technology and Specifications	<ul style="list-style-type: none"> - Settings, assignments, or actions related to software or applications - Textile material, food and beverage, medicinal substances - Qualities, characteristics, or specifications of animals or products
4	Online Discourse and New Experiences	<ul style="list-style-type: none"> - A first time experience, often with an element of surprise or anticipation - Social media, Donald Trump, Twitter, counsel, upbeat tweets - Apologies, add-ins, testaments, descendants, grades, versions, dialogue, honesty, downgrades, payments, relevance, sensibility, timelessness, credit, superiors, decency, hardened hearts, ageing, genuinely frightened by reality, reliability, shrouded, credited, ironically permitted, gigs, obsession, ...
6	Positive Experiences	<ul style="list-style-type: none"> - Expressions of excitement, sharing, or positive feedback - Expressions of gratitude, current time references, or positive descriptions - Experiences related to travel, leisure activities, meals, and events, particularly those with a temporal element (time, dates, or duration)
7	Commerce/ Finance	<ul style="list-style-type: none"> - Months, numbers, and second-hand collectibles, rentals, applications, retail spaces, or holiday gifts - Holiday or special occasion accessories/decorations - Food, tools/devices, and cosmetics/accessories
9	Access/ Acquisition	<ul style="list-style-type: none"> - Transfer, storage, or placement of objects or people - Consumption of food, beverages, or medications, sometimes for free or at a reduced price - Winning a prize or participating in a competition
13	Personal and Professional Experiences and Standards	<ul style="list-style-type: none"> - A discussion of customer service experiences, religious figures and texts, TV series, sporting events, community events, restaurants, international summits, golf rules, summer camps, company performance, and international relations - Proper nouns, often people's names, in contexts of competitions, scandals, or events, especially when related to games, sports, politics, or entertainment - Professional standards, requirements, and practices related to a variety of fields, including surveying, reviewing products/services, nutrition, data analysis, education, healthcare, music, and spirituality
14	Achievements, Lifestyle, Risks, News and Controversy, Conflicts, Corruption	<ul style="list-style-type: none"> - Events, particularly those related to conflict, competition, or problematic situations - Topics related to politics, sports, and current events, specifically focusing on major decisions, outcomes, and controversies - Government-related scandals and investigations, particularly those involving leaks, cover-ups, or accusations of wrongdoing
18	Structured Data	<ul style="list-style-type: none"> - Numerical or quantitative values, including years, measurements, and counts, in advert-like text excerpts - Titles, names, locations, and dates in bibliographic entries or citations - Academic degrees, professional roles, locations, and years related to education or employment history
19	Development and Improvement	<ul style="list-style-type: none"> - Mental and/or physical manipulation or transformation - Funding of projects and initiatives related to healthcare, social issues, and education - Methods, procedures, or sequences related to improvement, change, or progression
21	Personal Reflections	<ul style="list-style-type: none"> - First-person accounts, often expressing personal opinions, beliefs, or experiences - Personal updates - Experiences, actions, and feelings related to entertainment, media, and technology, along with personal anecdotes or opinions
26	Products/ Services and Medical Information	<ul style="list-style-type: none"> - Products or services with descriptions and/or characteristics - Products or services with details or instructions - Medical conditions, types of medical treatment, or medical professionals, sometimes involving a duration or repeating pattern
27	N/A	<ul style="list-style-type: none"> - Fitness, essays, digital skills training, product design and marketing, audits, internships, coaching, graphic design, academic assistance, data analytics, predictive modeling, computational chemistry, handwriting development, software documentation, development tools, intellectual property, ... - Business services, including career services, company recruitment, tours, and consulting, offered by organizations, for students and professionals, in various fields, such as marketing, technology, and healthcare - Rope access, locations in Arkansas, educational courses, probation terms, Nigerian aid, hospital communication improvement, South American airline travel, Nigerian profession improvement, admission to a school nursery, Florida ...
28	Business and Organization Information	<ul style="list-style-type: none"> - Locations of businesses or organizations - Geopolitical events and entities involved, particularly government actions and agencies - Events, services, or products offered by a business or organization
29	Events and Activities	<ul style="list-style-type: none"> - Achievements, awards, or special events, often including a specific person or group - Food, specific locations, and activities/routines, often involving a change in direction or state - Events, shows, or locations, often with a time or date, and sometimes including named people
35	Personal Development and Management	<ul style="list-style-type: none"> - Self-awareness, identity, and reality, often related to technology and its impact on the user - Discussions of financial, life, or career planning, resource allocation, and caregiving, often in the context of family or children - Actions or states of being, often ongoing or recently completed
38	Diverse Inquiries and Services"	<ul style="list-style-type: none"> - Promotional products or gifts relating to cuteness and popularity with customers - Locations, often neighborhoods or districts, and named entities associated with those locations - French language learning, professional certifications and qualifications, and company services related to specific industries
42	N/A	<ul style="list-style-type: none"> - First-person introspection, often related to mental and emotional states, self-awareness, and personal beliefs - First-person perspective related to identification, personal information, and objects - Conditional actions or situations and their potential outcomes, especially relating to rules, regulations, or personal choices
44	N/A	<ul style="list-style-type: none"> - Legal cases, particularly theft and court proceedings, and discussions of sports teams and players - International trade, diplomacy, and agreements between countries - Past events, especially from a year ago or more
45	Legal and Administrative Affairs	<ul style="list-style-type: none"> - Division of assets/property, medical procedures/treatments, legal disputes/court proceedings, and organizational activities/events - Ownership of property or membership status - Commercial transactions, legal proceedings, and financial obligations
48	N/A	<ul style="list-style-type: none"> - Health benefits of avocado, color testing tools, gene normalization for hPDL fibroblasts, teriyaki chicken wings, reptile hemoparasite identification, hemorrhoid treatments, omega-3 fatty acid supplements for fertility, baking trout, ... - Activities related to hobbies including airplane livery design, scrapbooking, SMART team coordination, journaling, early childhood sensory play and development, painting vegetables, taking consumer surveys, decorating with lampshades, ... - Infants, food, and crafting

Figure 11: Clustering of identified PRISM feature descriptions in **GPT-2 XL**. Shown are the $k = 50$ meta-clusters of feature descriptions, each labeled with a corresponding metalabel generated by Gemini 1.5 Pro, along with up to 3 randomly selected sample descriptions per cluster.

id	Metalabel	Feature Descriptions
3	Miscellaneous	<ul style="list-style-type: none"> - Japanese teriyaki chicken, a type of sedan, art pieces featuring wood, sugary food/drinks, gameplay mechanics, leggings, plumbing services, file names of furniture images - Hib vaccine, investor-owned utilities, couples counselling, weak economy/immune system, HARPO fellowship, outage - Energy-related industry or resource
4	Seeking and Managing Resources	<ul style="list-style-type: none"> - Data storage, transfer, or management, often in relation to websites, software, or online platforms - Ordering, requesting, or discussing types of services, accounts, or information, often related to online platforms, finances, or businesses - Requesting, searching for, or looking for something, especially services like legal or insurance, or items like quotes or properties
6	Linguistic Elements and Structures	<ul style="list-style-type: none"> - The conjunction "So" starting a sentence, often introducing a conclusion or consequence based on the preceding context - Conjunctions, prepositions, and occasionally other function words, appearing in descriptions of food and drink preparation, achievement announcements, or product descriptions - Commercial enterprises, locations of residence, textual works, family members, sports, and proper nouns
8	N/A	<ul style="list-style-type: none"> - Occupations or roles related to water - Locations (cities, states, or neighborhoods) and things found in homes or related to home maintenance/improvement - Products or services related to attire, beauty, or personal care
12	Products and Locations	<ul style="list-style-type: none"> - Clothing, accessories, or cosmetic products with descriptions of their materials, features, or benefits - Items or products and their descriptions including specifications, materials, and uses - Medical and/or chemical terms in the context of product descriptions or technical documentation
17	Competitive Analysis	<ul style="list-style-type: none"> - Business competitors/competition - Evaluation, academic/educational institutions, and certification/qualification - Publication details, often including author, title, date, and publisher/journal
19	N/A	<ul style="list-style-type: none"> - IndyCar series or races, often with the word "Indy" highlighted - Questions about processes, mostly using the auxiliary "does" - A person named Fei appearing in a conversational context
20	Spatiotemporal Descriptions and Personal Anecdotes	<ul style="list-style-type: none"> - Locations, proper nouns, and numbers related to places, events, or entities - Timestamps, specifically times of day - A short personal story often including a mention of a family member, sometimes in relation to a specific past time or recent event
22	Pattern Matching	<ul style="list-style-type: none"> - The letter G, capitalized or not, related to proper nouns in a list-like structure - A substring "ib" or "lbr", often within a proper noun, especially a person's name - Names, punctuation marks, and specifically the tokens "sh", "l", "mm", "j", and a newline character
26	Product/ Service Descriptions	<ul style="list-style-type: none"> - New service/product offerings or marketing/promotion of existing services/products - Belonging, origins, sources, or components - Product features or qualities
27	Ordinal Numbers	<ul style="list-style-type: none"> - Ordinal numbers in contextual information describing locations, groups, or ordered lists - Ordinal numbers, often within the context of lists or ordered items - Ordinal numbers in a numbered list
28	WordPress and Digital Business Skills	<ul style="list-style-type: none"> - Content related to the Wordpress platform, possibly focusing on its usage, features, and user groups - Financial/business topics, digital skills training, or software platforms and their features/benefits - Computer/IT skills, software, or computer programs, often in a business/professional/marketing/sales context
34	Digital Business and Technology	<ul style="list-style-type: none"> - Products and services related to pet care, home improvement, electronics, and computing - Software, tools, and resources for creating and managing websites and other digital content - Website/software development, marketing, and financial services/products
35	Publication and Distribution	<ul style="list-style-type: none"> - Relating to an edition or version of a book or relating to a card game - Giveaway, donation request, advertisement, or sharing information, related to a link or media - Relating to the beginning section of a piece of writing
38	N/A	<ul style="list-style-type: none"> - Medical studies of the effects of various factors or substances on different types of cancer - Energy sources, including renewable energy like tidal power as an alternative to fossil fuels, geological formations and processes like sediments, and motor neuron degeneration - International Business Machines (a technology company) and cystic fibrosis
42	N/A	<ul style="list-style-type: none"> - Locations offering services or events - Competition, playoff, or tournament sporting events - Discussions of computer hardware and software, website creation and management, recipes, personal anecdotes and hobbies, product reviews, and summaries of events
44	Product and Service Specifications	<ul style="list-style-type: none"> - Video file sharing, software, or online services related to media, including video resolution, file formats, platforms, and user experience - Attributes of products related to materials, sizes/dimensions, and/or color - Screen resolution or magnification
46	Listeria Contamination	<ul style="list-style-type: none"> - Food recalls due to bacterial contamination, specifically Listeria - Food contamination with Listeria
47	Taxes and Legal Obligations	<ul style="list-style-type: none"> - Ownership of creative digital content, specifically relating to Italian architecture or fashion accessories and their online availability - Tax obligations for non-resident sellers of real estate, especially focusing on the buyer's responsibility to withhold a percentage of the sale price for tax purposes - Instructions related to food preparation, especially baking or chilling in a refrigerator, sometimes followed by serving instructions
49	Instructional/ Explanatory	<ul style="list-style-type: none"> - Demonstrative pronouns (this, that) at the beginning of sentences, especially related to new information or summaries - Competitive situations, often sports or games, with emphasis on positions and actions taken by a team or individual player - Second-person pronouns in instructional or user-manual style texts, frequently appearing in contexts involving explanations of processes, tools, or options available to the reader

Figure 12: Clustering of identified PRISM feature descriptions in **GPT-2 Small SAE**. Shown are the $k = 50$ meta-clusters of feature descriptions, each labeled with a corresponding metalabel generated by Gemini 1.5 Pro, along with up to 3 randomly selected sample descriptions per cluster.