Д. Р. Мясникова, М. А. Сторчевой

УДК 004.65:004.8:658.5

Разработка ETL-пайплайна сбора и верификации данных о компаниях для рейтинга ESG

В статье предложен подход к автоматизированному сбору и верификации данных об организациях для последующего ESG-рейтингования. Авторы представляют концепцию ETL-пайплайна, объединяющего первичный LLM-поиск и последующую проверку, и обогащение данных через Dadata, Яндекс.Геокодер и СПАРК. Также авторы представляют результаты тестирования пайплайна на примере компаний СЗФО: успешно сформирован массив из 2361 отраслевой карточки компаний, соответствующих 1953 уникальным ИНН. В завершение авторы обсуждают ограничения и направления дальнейшего развития исследования.

Ключевые слова: ESG-рейтинг, сбор данных, ETL-пайплайн, LLM, Dadata, СПАРК, геокодирование, нормализация данных, автоматизация, отраслевые данные, human-in-the-loop.

ВВЕДЕНИЕ

Рейтинги ESG являются важным инструментом измерения деятельности организаций, который позволяет инвесторам и регуляторам учитывать не только финансовые результаты, но и экологические, социальные и управленческие аспекты их функционирования. Российские провайдеры рейтингов ESG (RAEX, РСПП, Forbes, НРК-РБК и др.) используют в основном экспертный и анкетный методы для составления рейтингов, в то время как зарубежные провайдеры (MSCI, Sustainalytics, Refinitiv и др.) активно используют алгоритмы ИИ для обработки данных из открытых источников. Задачей нашего

проекта является разработка такой информационно-аналитической платформы, которая позволит собирать информацию и рассчитывать рейтинги ESG полность в автоматическом объеме с минимальным участием экспертов. В данной статье мы рассмотрим важнейшую часть этой платформы – *сбор базовой информации о компаниях в заданном регионе, которые должны будут войти в рейтинг ESG*. Данная базовая информация включает в себя сведения о юридическом статусе организаций, отраслевой принадлежности, регионе регистрации и деятельности, численности сотрудников, выручке, адресах, координатах, официальных сайтах и других характеристиках. Качество исходных данных напрямую влияет на корректность рейтингового расчета, сопоставимость организаций и практическую применимость результатов, поэтому вопросы качества данных и воспроизводимости их обработки всегда являются самостоятельной частью проектирования аналитических информационных систем [1–3].

Подготовка алгоритма для сбора базовой информации о компаниях представляет собой трудоемкую аналитико-техническую задачу. Проблема заключается в том, что данные об организациях распределены между различными государственными и коммерческими реестрами, открытыми справочниками, сайтами организаций, API-сервисами и отраслевыми перечнями. Эти источники отличаются по структуре, полноте, актуальности и форматам представления информации. Ни один источник не дает возможность сформировать сразу готовую информационную карточку компании для участия в рейтинге ESG. Попытки объединения информации из разных источников приводят к тому, что сразу возникают дубли, расхождения в названиях, отсутствие координат, неоднозначность отраслевой классификации, ошибки кодировки, неполные строки и устаревшие сведения о статусе организаций. Одним из перспективных инструментов решения проблемы сбора и объединения данных являются большие языковые модели, которые открывают много возможностей для автоматизации предварительного поиска и структурирования информации. В близких задачах уже показана перспективность объединения веб-сбора данных и LLM для извлечения структурированной информации из неструктурированных источников [4–6]. Вместе с тем LLM не являются универсальным источником достоверных данных: возможны неполные ответы, галлюцинации, ошибки в реквизитах и нестабильность формата вывода [6–8]. Поэтому для подготовки данных

к ESG-рейтингованию необходима гибридная архитектура, где генеративные модели сочетаются с верификацией, нормализацией, дедупликацией и обогащением данных.

В данной статье рассматривается разработка масштабируемого ETL-пайплайна на основе LLM для автоматизированного сбора, проверки и подготовки данных об организациях и тестирование этого пайплайна на примере компаний Северо-Западного федерального округа. Подход объединяет генерацию отраслевых запросов, LLM-поиск организаций, динамическое обновление генерального списка ИНН, объединение и очистку CSV-файлов, валидацию через Dadata, нормализацию наименований, двойное геокодирование, частичное обогащение из СПАРК и обработку поврежденных CSV-файлов.

Цель работы состоит в демонстрации пригодности гибридной архитектуры LLM + Dadata + ETL-процедуры для масштабируемого формирования проверенного датасета организаций для ESG-аналитики. Научно-прикладной результат заключается в разработке воспроизводимого пайплайна, который переводит разрозненный сбор корпоративных сведений в последовательный процесс подготовки данных. Предлагаемое технологическое решение выполнено в концепции «одной кнопки», которая предполагает максимальную автоматизацию операций сбора, проверки, нормализации, обогащения и выгрузки данных в едином управляемом сценарии.

ПОСТАНОВКА ЗАДАЧИ И ИССЛЕДОВАТЕЛЬСКИЙ ПРОБЕЛ

Рейтинг ESG является формализованной оценкой экологической и социальной ответственности компании и происходит от названия трех укрупненных областей, которые подлежат оценке — environmental, social и governance. Рейтинг ESG измеряет степень добросовестности отношения компании ко всем ее заинтересованным сторонам: потребителям, работникам, поставщикам, инвесторам, местному сообществу и т. д. В дальнейшим эти ESG-оценки используют инвесторы, банки и регулирующие органы для определения степени добросовестности компаний и расставления приоритетов в работе с ними.

В настоящее время в России только около 300 крупнейших компаний имеют рейтинги ESG, а десятки тысяч остальных крупных и средних предприятий не охвачены этим измерением, хотя именно они составляют основной экономический потенциал.

Наш проект направлен на создание такого рейтинга и одной из первостепенных задач здесь является подготовка *базы данных компаний*, которые подходят для целей рейтинга и классифицированы по видам деятельности и по регионам таким образом, который позволяет корректное и удобное сопоставление их между собой. В идеальной классификации рядом должны оказываться максимально однородные компании из одного региона, например, “все производители хлеба в Мурманской области” или “все операторы ТБО в Ленинградской области”. В готовом виде такой классификации не существует, поэтому возникает задача для разработки алгоритмов ее формирования. Заметим, что при этом необходимо не просто найти организации, но и собрать сведения об организациях, подтвердить их юридический статус, исключить дубли, нормализовать наименования, определить отраслевую принадлежность, получить адресно-географические данные и по возможности дополнить карточки финансовыми и кадровыми показателями. Подготовка такой базы данных является сложной исследовательской и технологической задачей [18-19].

Таким образом, целью нашего исследования является создание автоматизированного пайплайна сбора и структурирования данных о компаниях Северо-Западного федерального округа (СЗФО) для последующего расчета ESG-рейтинга и визуализации на интерактивной карте. Требовалось охватить 11 субъектов округа, при этом компания включалась в выборку при наличии регистрации, юридического адреса или операционной деятельности на его территории. Отраслевая структура была разработана в рамках нашего проекта и включала пять крупных секторов (услуги, промышленность, энергетика, инфраструктура, агропромышленность) и 105 подотраслей.

В исследовании мы использовали авторскую отраслевую классификацию, специально разработанную для нашего проекта рейтинга ESG (см. Приложение 1). Она не совпадает с классификатором ОКВЭД, поскольку ОКВЭД имеет иную логику построения: он ориентирован на формальную регистрацию видов экономической деятельности, содержит избыточную для ESG-аналитики детализацию и не всегда отражает фактический профиль деятельности региональных компаний. По этой причине ОКВЭД использовался не как основа классификации, а как вспомогательный проверочный признак. Основой отнесения организаций к секторам, отраслям и подотраслям служила заданная в

проекте отраслевая структура, а данные ОКВЭД применялись для дополнительной валидации и выявления спорных случаев.

Для каждой организации необходимо было сформировать унифицированную карточку, включающую идентификационные, описательные, отраслевые, адресно-географические и, по возможности, финансово-кадровые поля. Основным идентификатором выступал ИНН, через который осуществлялась дедупликация, а также обогащение данных из разных источников. Помимо ИНН и нормализованных названий, карточка должна была содержать юридический адрес, регион СЗФО, координаты (широту и долготу), отраслевую принадлежность по проектной структуре, код ОКВЭД (при наличии), описание деятельности, веб-сайт, а также выручку и численность сотрудников. При формировании реестра применялся размерный фильтр. В приоритетном порядке отбирались компании с численностью не менее 100 сотрудников. Вместе с тем из этого правила допускались исключения: если компания имела высокую выручку (от 100 млн рублей), являлась системообразующей, занимала заметное положение в отрасли или была широко известна на региональном рынке, она могла быть включена в датасет даже при меньшей численности персонала. На этапе постобработки основное правило включения формализовалось следующим образом: *CompanyEmployees >= 100 ИЛИ CompanyRevenueRawRUB >= 100 000 000.* Выручка учитывалась в рублях и использовалась как дополнительный признак значимости организации.

Отдельным требованием являлась подготовка данных к загрузке в интерактивную карту с функциями кластеризации, фильтрации по отраслям и регионам, отображения детальной карточки компании при клике и сравнения нескольких организаций. Следовательно, каждая запись в итоговом датасете должна была содержать достоверные географические координаты.

Для решения задачи использовались несколько типов источников. Большая языковая модель с веб-поиском применялась для первичной генерации списков компаний по подотраслям. API Dadata служило для валидации по ИНН, проверки актуального статуса, уточнения адреса, региона и координат. В случаях, когда адрес был известен, а

координаты отсутствовали, подключался Яндекс.Геокодер как fallback-механизм. Система СПАРК (по договору университета) обеспечивала обогащение данных выручкой и численностью сотрудников. Открытые справочники (Rusprofile, Audit-it, Checko, List-Org) использовались для нормализации названий и уточнения описаний, а официальные сайты компаний — как дополнительный канал информации. Ни один из перечисленных источников не давал полных и готовых к употреблению данных, поэтому требовалась гибридная схема с верификацией на каждом этапе и ручным контролем проблемных записей.

Разработанный пайплайн в первую очередь отрабатывался на примере компаний СЗФО, однако архитектура решения изначально проектировалась для масштабирования. Подразумевается, что после отладки на СЗФО конвейер будет запущен для всех остальных федеральных округов России. Более того, система должна обеспечивать регулярный перезапуск, чтобы отслеживать изменения в популяции компаний — появление новых организаций, ликвидацию старых, смену реквизитов и показателей. Таким образом, задача заключалась не в разовом формировании списка организаций, а в создании воспроизводимого, масштабируемого и ритмично обновляемого пайплайна, который превращает разрозненные неполные данные в структурированный, верифицированный и геопривязанный массив, пригодный для ESG-аналитики и картографической визуализации.

АРХИТЕКТУРА ПАЙПЛАЙНА

В первой версии пайплайна использовалась линейная схема: генерация поисковых запросов, сбор первичных списков через LLM, валидация через Dadata, нормализация, геокодирование, обогащение данными СПАРК, дедупликация и передача на карту. Такой подход позволил получить первый массив компаний, но почти сразу проявилось ограничение: ответы LLM в формате CSV далеко не всегда отличались стабильной структурой. В результате система приостанавливала работу, нарушала целостность строк с данными и не выполняла необходимую задачу. В качестве решения было принято перестроить систему. В результате линейная модель уступила место модульному ETL-пайплайну с раздельными контурами для корректных, проблемных и повторно ис-

правленных данных. В основу обновленной архитектуры легла новая файловая структура, включающая в себя специализированные каталоги: incoming/, retry/, clean/, dirty/, enriched/, geo/, final/, а также backup/, archive/, logs/ и cache/, благодаря более четкому разделению данных по каталогам в зависимости от их состояния, принцип “No Data Loss” проявляется более существенно.

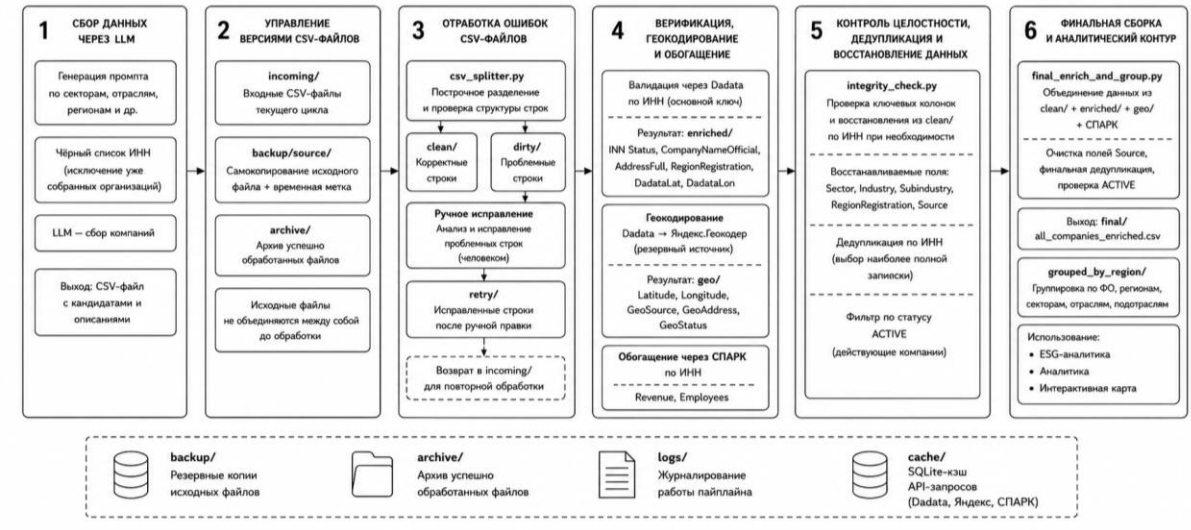

Рис. 1. Полная схема ETL-пайплайна и структуры хранения данных

Этап 1. Сбор данных через LLM и контроль полноты

На первом этапе формировался первичный список организаций-кандидатов для последующей верификации и обогащения. Поиск выполнялся большой языковой моделью с веб-доступом на основе параметризованного задания, в котором указывались сектор, отрасль, подотрасль, регионы СЗФО, требования к действующему статусу организации, наличию ИНН и минимальному масштабу деятельности. Для каждой из 105 подотраслей использовалась единая структура задания с автоматической подстановкой отраслевых параметров.

Модель должна была вернуть результат в машиночитаемом CSV-формате с фиксированной последовательностью полей. В первоначальную карточку включались сектор,

отрасль, подотрасль, официальное и нормализованное название компании, ИНН, регионы регистрации, расположения головного офиса и деятельности, краткое описание, официальный сайт и источник информации. На этом этапе данные рассматривались как предварительные: сведения об ИНН, статусе, адресе, размере и отраслевой принадлежности подлежали дальнейшей проверке по независимым источникам (см. пример в таблице 1). Полный шаблон поискового промпта приведен в Приложении 2.

Таблица 1. Пример собранных данных.

Поле	Значение
Sector	Агропродовольственный сектор
Industry	Сельское хозяйство
Subindustry	Животноводство
CompanyNameOfficial	ООО «Белгранкорм - Великий Новгород»
CompanyDescription	ООО «Белгранкорм - Великий Новгород» основано в 2024 в Новгородской области и специализируется на животноводстве и сельскохозяйственном производстве. Компания обслуживает Новгородскую область. Основная база находится в Новгородской области. Средняя численность работников составила 681 человек в 2024 году. *Показатели требуют уточнения.*
INN	5305006239
RegionRegistration	Новгородская область
Address	175450, Новгородская обл, Крестецкий р-н, деревня Новое Рахино, д 106А
CompanyEmployees	954
CompanyRevenue	6 149 490 000
Lat	58.138155
Lon	32.742388
URL	https://jasnzori.ru/
Source	проверено по rusprofile.ru

Для уменьшения числа повторов использовался динамически обновляемый контрольный список ИНН. После каждого запуска ИНН найденных организаций извлека-

лись из полученного файла и добавлялись в перечень исключений. При следующем обращении модель получала то же поисковое задание, но с прямым запретом повторно включать ранее найденные организации. Таким образом, последовательные запуски были направлены на расширение списка кандидатов, а не на повторное извлечение уже известных записей.

Один пустой ответ LLM не считался достаточным основанием для завершения поиска, поскольку при следующем запуске модель могла выявить дополнительные организации. Рабочим критерием остановки стали три последовательных ответа без новых компаний в режиме веб-поиска. После этого выполнялся дополнительный запуск в режиме углубленного исследования с теми же критериями и обновленным списком исключений. Этот режим использовался для расширения поисковых гипотез за счет обращения к смежным формулировкам, региональным и отраслевым источникам, кодам ОКВЭД и косвенным упоминаниям организаций.

Операциональный критерий завершения этапа включал три последовательных пустых ответа в режиме веб-поиска и один пустой ответ в режиме углубленного исследования. Такая процедура не доказывает абсолютную полноту реестра, поскольку эталонный перечень всех компаний заданной подотрасли отсутствует. Вместе с тем она задает воспроизводимое правило насыщения результатов в пределах использованных источников и режимов поиска.

В результате первого этапа формировался набор CSV-файлов с компаниями-кандидатами и накопительный реестр обработанных ИНН. Полученные записи передавались на следующий этап пайплайна для сохранения исходных версий, структурной проверки и разделения корректных и поврежденных строк.

Этап 2. Организация управления версиями CSV-файлов

Задачей второго этапа является сохранение исходных и промежуточных версий CSV-файлов на всех стадиях обработки. Первоначально применялась линейная схема, при которой региональные и отраслевые выгрузки последовательно объединялись Python-скриптом в единый файл. При увеличении числа выгрузок такой подход создавал риски

перезаписи данных, потери исходных записей, повторного включения строк и нарушения структуры таблицы.

В модульной архитектуре каждый необработанный CSV-файл сохраняется как самостоятельный входной объект в каталоге incoming/. Перед началом обработки его исходная версия автоматически копируется в каталог backup/source/ с временной меткой. Это позволяет восстановить первоначальную выгрузку при ошибке синтаксического разбора, сбое внешнего API или некорректном результате последующих преобразований. После успешной обработки файл переносится в archive/, что отделяет завершенные циклы от новых входных данных. Таким образом, incoming/ используется для файлов текущего цикла, backup/source/ — для неизменяемых резервных копий, а archive/ — для обработанных выгрузок.

Этап 3. Отработка ошибок в структуре и разметке файлов

Задача этого этапа заключается в обработке всех поступивших CSV-файлов так, чтобы случайные синтаксические ошибки не вели к пропаданию записей и не нарушали работу последующих модулей. В предыдущей версии пайплайна был реализован трехуровневый fallback механизм (механизм отказоустойчивости). На первом уровне использовался встроенный в pandas анализатор CSV со строгим режимом чтения: он обеспечивал высокую скорость, но при столкновении с синтаксическими ошибками, прекращал работу. Второй уровень использовал более гибкий, но медленный режим чтения на python, он был способен пропускать некорректные строки и продолжать разбор данных. Этот подход был устойчивее, однако он терял поврежденные записи, что противоречило принципу No Data Loss. Третьим уровнем было ручное построчное исправление, которое оставалось самым надежным, но крайне медленным на больших объемах данных. В обновленной версии пайплайна этот механизм был заменен на построчное разделением потока на корректные и проблемные записи с сохранением исходной строки и причины ошибки. Когда файлы попадают во входной каталог (incoming/), модуль csv_splitter.py разделяет записи на два потока: полностью корректные строки отправляются в clean/ и идут дальше по пайплайну, а строки с нарушениями структуры попадают в dirty/. Для

каждой проблемной записи фиксируются два служебных поля: raw_line (исходная строка без изменений) и dirty_reason (причина ошибки парсинга). Если позже такая строка исправляется вручную, она переносится в retry/ и повторно проходит весь цикл обработки наравне с новыми данными. Благодаря этому единичный дефект больше не блокирует разбор всего массива, а проблемные записи не удаляются, а уходят на отдельный контур для правки и повторного запуска.

Рис. 2. Схема отработки ошибок в структуре и разметке

Этап 4. Верификация, геокодирование и обогащение данных

Целью этапа является преобразование предварительного списка компаний-кандидатов в проверенные карточки организаций. Последовательно выполняются три операции: 1) валидация через Dadata, 2) геокодирование и 3) обогащение данных через СПАРК.

1. *Валидация через Dadat*а. Dadata — это справочно-информационный сервис, который по ИНН или названию возвращается актуальные регистрационные данные, юридический адрес, статус и координаты организации. Основным идентификатором организации служит ИНН. Поиск по названию рассматривался как вспомогательная мера: он менее надежен из-за сокращений, вариантов написания и возможных совпадений. На этом шаге с помощью API-ключа отправляется запрос на сервер Dadata и по собранным организациям уточняется официальный статус, полное наименование, юридический адрес, регион регистрации и, при наличии, координаты. Результаты сохраняются в каталог enriched/. Итоговый

профиль после Dadata может включать поля: INN, Status, CompanyNameOfficial, AddressFull, RegionRegistration, DadataLat, DadataLon.

2. *Геокодирование*. Основным источником координат выступает та же Dadata. Если координаты отсутствуют или требуют уточнения, подключается Яндекс.Геокодер — сервис для преобразования текстовых адресов в географические координаты, в системе, используемый в качестве запасного варианта. Это повышает полноту геопривязки итогового массива. Результаты сохраняются в каталог geo/ с полями: Latitude, Longitude, GeoSource, GeoAddress, GeoStatus.
3. *Обогащение через СПАРК*. СПАРК — это система профессионального анализа рынков и компаний, которая агрегирует данные из официальных государственных реестров (ФНС, Росстат) и предоставляет финансовые и организационные показатели через API-ключ. Ключом сопоставления снова служит ИНН. Для дальнейшего ESG-анализа наиболее важны выручка и численность сотрудников (поля Revenue, Employees). Если финансовые сведения по конкретной организации отсутствуют, запись не удаляется – соответствующие поля остаются незаполненными и могут быть дополнены в одном из следующих циклов обработки.
4. *Распределение по критериям included/review/excluded.* После обогащения данных через СПАРК компании распределялись по трем статусам: included, review и excluded. Это распределение выполнялось не на этапе первичного LLM-поиска, а на этапе постобработки и проверки уже собранных записей. Included — компания включалась в основной датасет, если она была действующей, имела надежно установленный ИНН, соответствовала заявленной отрасли или подотрасли и проходила размерный фильтр. Основным размерным критерием была численность сотрудников не менее 100 человек. Дополнительно учитывалась выручка: компания могла быть сохранена в основном датасете при меньшей или неизвестной численности, если ее выручка составляла не менее 100 000 000 рублей и были подтверждены остальные признаки значимости и отраслевого соответствия. Формально правило включения можно представить следующим образом: активный статус + надежно установленный ИНН + подтвержденная отраслевая принадлежность +(CompanyEmployees >= 100 ИЛИ CompanyRevenueRawRUB >= 100 000 000)=> included. Review — компания направлялась на дополнительную проверку, если данных было недостаточно для надежного автоматического решения. В этот

статус попадали случаи, когда отсутствовали или были неполными сведения о численности и выручке, была спорной отраслевая принадлежность, не хватало подтверждений деятельности, возникали проблемы сопоставления по ИНН или требовалось ручное решение. Excluded — компания исключалась из финальной included-выгрузки только при наличии формального основания: неактивный статус юридического лица, надежно установленное несоответствие размерному фильтру, отсутствие подтверждения принадлежности к нужной отрасли или невозможность надежно идентифицировать компанию по ИНН. Общий принцип обработки был следующим: подтверждено соответствие критериям — included; данных недостаточно или есть спорность — review; подтверждено несоответствие — excluded. Спорные записи не удалялись безвозвратно, а сохранялись в контуре review для последующей проверки.

5. *Работа над описаниями.* Для генерации описаний компаний использовался шаблонный промпт, автоматически формируемый пайплайном для каждой строки датасета. Этот этап выполнялся не на стадии первичного поиска организаций, а после структурного обогащения данных. Источником для генерации служил файл final/all_sectors_final_active_spark_included.csv, содержащий уже собранные и проверенные сведения о компаниях. В промпт передавались только структурированные поля компании: CompanyNameOfficial, INN, Sector, Industry, Subindustry, RegionName / RegionRegistration, Address_dadata / SparkAddress, SparkOKVEDMainCode, SparkOKVEDMainName, SparkActivity, CompanyEmployees, CompanyRevenueRawRUB, Website / SparkWebsite. На основе этих данных модуль description_builder.py формировал запрос к YandexGPT через yandexgpt_client.py. Модель не выполняла самостоятельный поиск описаний в интернете. Ее задача заключалась в том, чтобы на основе уже имеющихся структурированных данных сформировать краткое фактическое описание компании. Шаблонный промпт содержал ограничения: использовать только предоставленные поля, не добавлять неподтвержденные сведения, не использовать рекламные или оценочные формулировки, писать в нейтральном фактическом стиле, учитывать отрасль, регион, масштаб деятельности, выручку и численность сотрудников при наличии этих данных. После генерации описание проходило автоматическую проверку качества. Результаты проверки фиксировались в технических полях description_status,

description_quality_score, description_issues, description_model, description_cached и description_input_hash. Таким образом, генерация описаний была не ручным процессом, а частью автоматизированного пайплайна: для каждой компании формировался индивидуальный шаблонный запрос на основе ее структурированных данных, а затем результат проходил дополнительный контроль качества.

Этап 5. Контроль целостности, дедупликация и восстановление данных

Задача этапа заключается в проверке сохранности ключевых полей после обогащения данных, устранении дублирующихся записей и формировании массива действующих организаций. Поскольку данные проходят через несколько промежуточных файлов и внешних сервисов, в пайплайн встроен дополнительный контроль целостности, позволяющий убедиться, что исходные классификационные поля сохранены и корректно сопоставлены с обогащенными данными.

Чтобы гарантировать сохранность ключевых полей, в пайплайн встроен модуль integrity_check.py. Он проверяет наличие необходимых колонок после этапов Dadata и геокодирования и при необходимости восстанавливает их по ИНН из данных, сохраненных в clean/. Восстанавливаемые поля могут включать: Sector, Industry, Subindustry, RegionRegistration, Source.

После контроля целостности проводится дедупликация. Наиболее надежным идентификатором признан ИНН; сопоставление по названию или адресу используется лишь как вспомогательный инструмент. Если для одного ИНН обнаруживается несколько записей, из них выбирается наиболее полная карточка – та, в которой заполнены статус, адрес, координаты, отраслевые реквизиты и финансовые показатели. Дополнительно применяется фильтр по статусу ACTIVE — этот статус означает, что организация является действующей, а не ликвидированной или находящейся в процессе реорганизации. Такой фильтр позволяет сформировать рабочий массив действующих организаций для последующего анализа.

Этап 6. Финальная сборка и передача готовых данных

Финальный этап объединяет результаты всех предыдущих операций в единый датасет, готовый для ESG-анализа, группировки и визуализации. Сборка выполняется модулем final_enrich_and_group.py, который сводит воедино данные из каталогов clean/, enriched/, geo/ и обогащенные показатели СПАРК. На выходе формируется файл final/all_companies_enriched.csv. В процессе сборки очищается поле Source, проводится финальная дедупликация, повторная проверка активного статуса и группировка. Результаты могут быть разложены, например, по каталогу grouped_by_region/, где данные группируются по федеральному округу, сектору, отрасли и подотрасли. Корректность определения федерального округа напрямую зависит от качества заполнения поля RegionRegistration и полноты справочника регионов, поэтому этот компонент рассматривается, как зону для дальнейших улучшений.

В отличие от первоначальной схемы последовательного слияния CSV-файлов, где любой дефект CSV грозил остановкой или потерей данных, обновленный пайплайн разделяет потоки: корректные строки напрямую попадают в clean/, проблемные оседают в dirty/ с возможностью ручной правки и повторного ввода через retry/. Контрольный список ИНН снижает повторный сбор одних и тех же организаций через LLM, а набор защитных механизмов — от резервного копирования исходников, локального кэширования ответов от внешних API (чтобы избежать повторных запросов и при этом не терять данные при сбоях сети) и подробной записи каждого шага до промежуточных сохранений и восстановления утраченных колонок по ИНН из clean/ — придает системе дополнительную устойчивость. В совокупности эти решения превратили инструмент, изначально предназначавшийся для разовой выгрузки, в пайплайн, способный к регулярному обновлению базы и масштабированию на новые отрасли и регионы.

ОТБОР И НАСТРОЙКА МОДЕЛЕЙ LLM

Для первичного формирования отраслевых списков (этап 1) требовался инструмент, способный искать организации одновременно по территории, отрасли, подотрасли, юридическому статусу и масштабу деятельности и возвращать результат в машиночитаемом формате. Предварительно были протестированы DeepSeek, Claude Sonnet, GPT API, ChatGPT с веб-поиском, YandexGPT API и прямое обращение к Dadata.

Всем моделям передавались задания, построенные по единой логике: сектор, отрасль и подотрасль, территориальные границы СЗФО, требования к действующему статусу и численности организации, наличие ИНН, состав обязательных полей и фиксированный формат CSV или JSON. Оценивались полнота выдачи, наличие идентификаторов юридических лиц, соответствие региональным и отраслевым ограничениям, стабильность структуры ответа, корректность обработки кириллицы, возможность масштабирования на 105 подотраслей, стоимость и совместимость с последующей верификацией.

Таблица 2. Результаты предварительного тестирования инструментов

Инструмент	Результат	Основные ограничения	Роль в итоговом пайплайне
DeepSeek	Формировал отдельные списки кандидатов	Нестабильность поиска, отказы, непроверяемые сведения и частые галлюцинации	Не использовался
Claude Sonnet	Давал структурированные ответы и широкий список кандидатов	Качество зависело от подотрасли; сведения требовали обязательной проверки	Использовался только в сравнительном пилоте
OpenAI API	Позволял автоматически формировать и сохранять CSV-файлы	Ограниченная полнота отдельных запросов и рост стоимости при масштабировании	Проверка технической реализуемости автоматического контура
ChatGPT	Обеспечивал наиболее приемлемое сочетание полноты, структурированности и доступа к источникам	Требовал ручного запуска и передачи результатов в ETL-контур	Основной канал первичного поиска
YandexGPT API	В отдельных запусках возвращал структурированные данные	Нестабильный JSON, тайм-ауты и ошибки обработки кириллицы	Не использовался для первичного поиска; применялся для генерации описаний
Dadata	Надежно проверял известные организации по ИНН	Не формирует исходный отраслевой список и требует точного идентификатора	Основной справочный канал верификации

Результаты показали, что LLM и справочные системы выполняют разные функции. Языковая модель расширяет поисковое пространство и формирует список кандидатов,

но не гарантирует достоверность реквизитов. Dadata, напротив, проверяет уже известный ИНН, однако не может самостоятельно выявить все организации заданной подотрасли. Поэтому эти инструменты использовались последовательно, а не как взаимозаменяемые источники.

Эксперимент с OpenAI API

В эксперименте с OpenAI API было сформировано 17 CSV-файлов, содержащих 177 исходных отраслевых записей. После структурной обработки, проверки по ИНН и фильтрации сохранилось 170 записей, соответствующих 166 уникальным юридическим лицам. Все записи итогового массива относились к действующим организациям и имели формально корректные ИНН.

Сохранение 96,0% исходных строк характеризует результат всей последовательности «генерация — обработка — верификация — фильтрация», а не исходную точность GPT API. Эксперимент подтвердил техническую реализуемость автоматизированного сбора, но выявил недостаточную полноту отдельных запросов и рост стоимости при масштабировании на 105 подотраслей. Поэтому OpenAI API рассматривался как перспективный компонент автоматизации, но не использовался как единственный канал формирования основного массива.

Пилотное сравнение GPT-5 и Claude

Для анализа характерных ошибок исходных LLM-ответов был проведен пилот на двух подотраслях: «Кондитерская промышленность» и «Разработка ПО и ИТ-услуги». Первая характеризуется относительно устойчивой отраслевой и территориальной принадлежностью компаний. Для ИТ-организаций место регистрации, головной офис, филиалы и территория деятельности часто не совпадают.

Анализировались необработанные CSV-файлы до проверки через Dadata и другие справочные источники. Поэтому результаты отражают непосредственное качество модельной генерации, а не характеристики итогового верифицированного массива.

Таблица 3. Результаты пилотного сравнения необработанных ответов GPT-5 и Claude

Модель	Подотрасль	Записей	Регионы регистрации	Регионы деятельности	Заполнено URL	ИНН с кодом региона вне СЗФО	ИНН, прошедшие проверку контрольного разряда
GPT-5	Кондитерская промышленность	7	2	2	4 из 7 (57,1%)	0	0 из 7
GPT-5	Разработка ПО и ИТ-услуги	14	1	1	14 из 14 (100%)	2	2 из 14
GPT-5	Всего	21	2 уникальных региона	2 уникальных региона	18 из 21 (85,7%)	2 из 21 (9,5%)	2 из 21 (9,5%)
Claude	Кондитерская промышленность	18	9	10	3 из 18 (16,7%)	0	2 из 18
Claude	Разработка ПО и ИТ-услуги	45	8	11	32 из 45 (71,1%)	10	9 из 45
Claude	Всего	63	до 11 регионов	все 11 субъектов СЗФО	35 из 63 (55,6%)	10 из 63 (15,9%)	11 из 63 (17,5%)

Проверка контрольного разряда подтверждает только формальную структуру ИНН. Она не доказывает принадлежность идентификатора указанной организации, действующий статус или соответствие заданной подотрасли. Аналогично заявленные моделью регионы не считались подтвержденными до внешней верификации.

GPT-5 возвращал более компактные результаты и чаще заполнял URL, но демонстрировал ограниченный региональный охват. Claude формировал значительно больше кандидатов и шире охватывал регионы, однако чаще смешивал регион регистрации, расположение офиса и территорию деятельности. В обеих моделях встречались правдоподобные, но неверные реквизиты.

Таблица 4. Интерпретация результатов пилота

Модель	Основное преимущество	Основной риск	Обоснованная роль
GPT-5	Высокая заполненность URL и компактный формат результатов	Ограниченный географический охват и неверные реквизиты	Формирование основного или дополнительного списка кандидатов с последующей проверкой
Claude	Большое число кандидатов и широкий региональный охват	Смешивание регистрации, офиса и операционного присутствия; рост числа непроверенных записей	Расширение поискового пространства с обязательной отбраковкой

Пилот не предназначался для статистически полного ранжирования языковых моделей, поскольку охватывал только две подотрасли и проводился без эталонного перечня компаний. Его задачей было выявление характерных ошибок, которые необходимо учитывать при проектировании верификационного контура.

По совокупности практических критериев основным инструментом первичного поиска был выбран ChatGPT с веб-поиском. Этот выбор не означает его абсолютного превосходства над другими моделями. Полученные результаты обосновывают прежде всего гибридную архитектуру: LLM используется для обнаружения и предварительного структурирования компаний-кандидатов, а достоверность итогового массива обеспечивается независимой проверкой ИНН, юридического статуса, региона, отраслевой принадлежности и размера организации.

РЕЗУЛЬТАТЫ

В результате апробации пайплайна на 105 подотраслях СЗФО после структурной обработки, проверки юридического статуса и дедупликации был сформирован активный массив из 2361 отраслевой карточки, соответствующей 1953 уникальным ИНН. Под отраслевой карточкой понимается включение юридического лица в конкретную отрасль или подотрасль. Поэтому одна организация могла быть представлена несколькими карточками, если ее деятельность относилась к нескольким функциональным направлениям.

Активный массив включал 836 карточек промышленных компаний, 817 карточек организаций сферы услуг, 354 карточки инфраструктурных организаций без учета энергетики, 166 карточек энергетических компаний, 185 карточек агропродовольственного сектора и 3 записи без указанного укрупненного направления.

После обогащения данными СПАРК, применения размерного фильтра и итогового контроля качества 1890 карточек были отнесены к категории included, а 471 карточка — к категории review. В категорию included вошли 1511 уникальных юридических лиц. Разница между числом карточек и числом уникальных ИНН составила 379 записей и была обусловлена сохранением подтвержденных межотраслевых и межподотраслевых пересечений. Категория excluded в итоговом распределении отсутствовала, поскольку

записи с подтвержденным несоответствием критериям были удалены на предыдущих этапах, а случаи с недостаточными или противоречивыми данными направлялись в review.

Таблица 5. Распределение карточек по укрупненным направлениям

Направление	Активные карточки	Included	Review	Доля review, %
Промышленность	836	817	19	2,3
Услуги	817	369	448	54,8
Инфраструктура без энергетики	354	354	0	0
Энергетика	166	163	3	1,8
Агропродовольственный сектор	185	185	0	0
Не указано	3	2	1	33,3
Итого	2361	1890	471	19,9

Наибольшее число спорных случаев было выявлено в сфере услуг: из 817 активных карточек 448, или 54,8%, были направлены на дополнительную проверку. В промышленности доля review составила 2,3%, в энергетике — 1,8%. Для инфраструктурного и агропродовольственного блоков все активные карточки прошли критерии включения.

Итоговый included-массив содержал идентификационные, отраслевые, описательные, адресно-географические и финансово-кадровые сведения. Все 1890 карточек имели ИНН, название организации, адрес, регион, координаты, выручку, юридический статус и указание источника. Сайт или иной URL был заполнен для 1826 карточек, что составляет 96,6% массива. Данные о численности работников имелись для 1535 карточек, или 81,2%.

Таблица 6. Полнота полей итогового массива

Показатель	Значение
Отраслевые карточки	1890
Уникальные ИНН	1511

Показатель	Значение
Межотраслевые и межподотраслевые пересечения	379
Карточки с ИНН	1890 из 1890, 100%
Карточки с названием организации	1890 из 1890, 100%
Карточки с адресом	1890 из 1890, 100%
Карточки с регионом	1890 из 1890, 100%
Карточки с координатами	1890 из 1890, 100%
Карточки с выручкой	1890 из 1890, 100%
Карточки с юридическим статусом	1890 из 1890, 100%
Карточки с указанием источника	1890 из 1890, 100%
Карточки с сайтом или URL	1826 из 1890, 96,6%
Карточки с численностью работников	1535 из 1890, 81,2%

Финальный контроль качества завершился со статусом QUALITY_GATE=PASS. Число записей в итоговом экспорте совпало с числом карточек категории included, пропуски в обязательных полях record_id и INN отсутствовали, а все включенные записи прошли установленный размерный фильтр.

Наличие нормализованных адресов и координат позволило загрузить итоговый массив в тестовый картографический интерфейс на базе Leaflet и OpenStreetMap. Проверка подтвердила пригодность данных для фильтрации по регионам и отраслям, отображения организаций по координатам и формирования карточек компаний. Разработка полноценного интерфейса ESG-аналитики выходит за рамки настоящего исследования.

НАПРАВЛЕНИЯ ДАЛЬНЕЙШЕГО РАЗВИТИЯ

Первое направление связано с переходом от дискретного запуска отдельных модулей к единому оркестрируемому сценарию (концепция «одной кнопки»). В текущей реализации каждый модуль автоматически выполняет соответствующий этап обработки и сохраняет промежуточные результаты, однако последовательность модулей запускается аналитиком. Разработка оркестратора позволит объединить LLM-поиск, структурную

проверку, верификацию, обогащение, дедупликацию, контроль качества и экспорт данных в единый управляемый процесс. Экспертное участие при этом должно сохраняться для записей с конфликтующими идентификаторами, неопределенной отраслевой принадлежностью, несовпадением регионов регистрации и деятельности, а также неполными финансовыми и кадровыми сведениями.

Второе направление предусматривает интеграцию подготовленного массива с картографическим и аналитическим интерфейсом ESG-рейтинга. Наличие нормализованных адресов, регионов и координат уже позволило проверить отображение компаний на интерактивной карте на базе Leaflet и OpenStreetMap (см. рис. 3). Дальнейшая разработка может включать фильтрацию по регионам, секторам и подотраслям, кластеризацию организаций, отображение карточек компаний, сопоставление нескольких организаций и визуализацию результатов ESG-оценки. Картографический интерфейс при этом рассматривается не как самостоятельный источник данных, а как средство анализа и контроля результатов пайплайна.

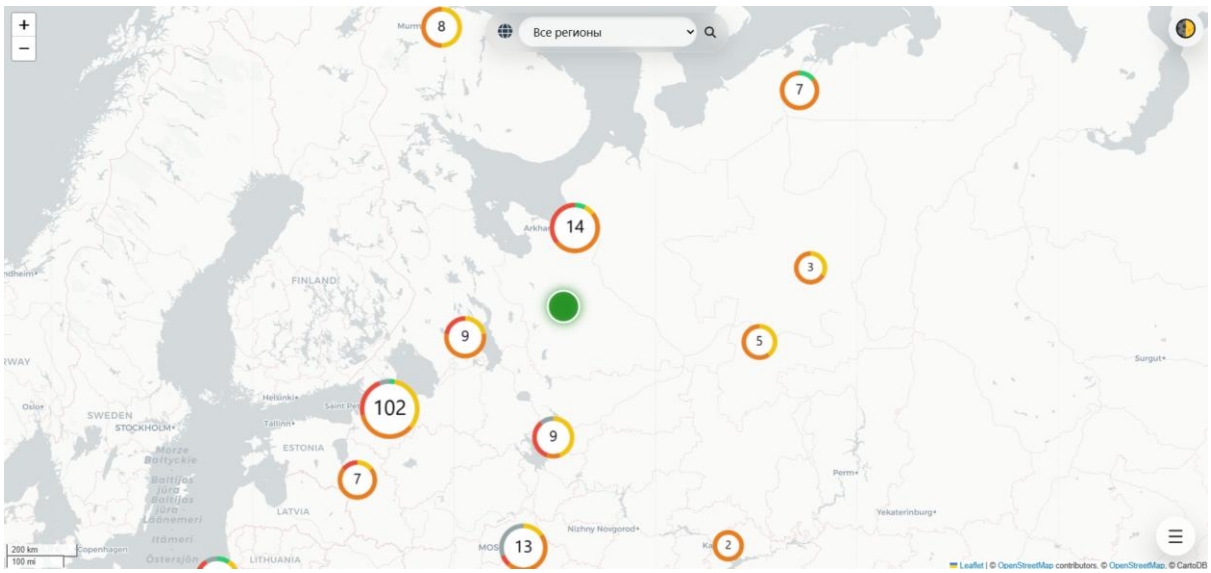

Рис. 3. Карта с геолокацией найденных компаний

Третье направление состоит в повышении полноты, воспроизводимости и актуальности формируемого реестра. Для этого необходимы автоматическая фиксация параметров и результатов каждого LLM-запуска, оценка прироста новых компаний при повторных запросах, сопоставление результатов нескольких моделей и поиск независимых

контрольных перечней организаций по отдельным отраслям и регионам. Регулярный перезапуск пайплайна позволит отслеживать появление и ликвидацию организаций, изменение их реквизитов, численности и выручки. Дополнительными задачами являются развитие правил автоматической проверки отраслевой принадлежности, сокращение доли записей review и масштабирование решения с СЗФО на другие федеральные округа.

ДИСКУССИЯ И ОГРАНИЧЕНИЯ

Основным ограничением исследования является отсутствие эталонного перечня всех организаций, соответствующих заданной комбинации региона, масштаба и подотрасли. Поэтому примененный критерий остановки — три последовательных пустых ответа веб-поиска и один пустой ответ в режиме углубленного исследования — следует рассматривать как операциональный показатель насыщения поиска, а не как доказательство абсолютной полноты реестра.

Дополнительное ограничение связано с авторским отраслевым классификатором. Он разработан для сопоставления компаний в ESG-анализе и не полностью совпадает с ОКВЭД, поэтому отнесение организаций к отдельным подотраслям может зависеть от экспертных решений. ОКВЭД использовался как вспомогательный признак, однако он не всегда отражает фактический профиль деятельности компании. Для спорных случаев требуется сопоставление нескольких источников и сохранение экспертной проверки.

Пилотное сравнение GPT-5 и Claude также имеет ограниченный характер. Оно проводилось только на двух подотраслях и без эталонной совокупности компаний, поэтому не позволяет построить статистически обоснованный рейтинг моделей или оценить полноту каждой из них. Его результаты следует интерпретировать как выявление характерных типов ошибок: ограниченного географического охвата, смешения региона регистрации и деятельности, а также генерации правдоподобных, но неверных реквизитов.

Наконец, итоговый результат зависит от полноты и доступности внешних информационных сервисов. Dadata, СПАРК и геокодирующие системы различаются по покрытию данных, имеют ограничения API и могут изменять структуру или условия доступа. Наименее полным показателем в итоговом массиве оказалась численность работников.

Для промышленного использования пайплайна необходимы журналирование запросов, кэширование ответов, повторная обработка ошибок, фиксация версий источников и регулярное обновление сведений о компаниях.

ЗАКЛЮЧЕНИЕ

В работе разработан и апробирован гибридный ETL-пайплайн формирования верифицированного реестра компаний для ESG-аналитики. В предлагаемой архитектуре LLM используется для обнаружения и предварительного структурирования организаций-кандидатов, а достоверность итогового массива обеспечивается последующей проверкой по ИНН, юридическому статусу, региону, отраслевой принадлежности и размерным критериям с использованием независимых источников и процедур контроля качества.

Апробация на 105 подотраслях СЗФО позволила сформировать 2361 активную отраслевую карточку, соответствующую 1953 уникальным ИНН; после верификации и фильтрации в итоговый массив вошли 1890 карточек, соответствующих 1511 юридическим лицам. Полученные результаты подтверждают практическую применимость гибридного подхода, но не позволяют считать сформированный реестр абсолютно полным из-за отсутствия эталонного перечня компаний. Основными направлениями дальнейшей работы являются повышение воспроизводимости оценки полноты, развитие автоматической отраслевой проверки и объединение модулей пайплайна в единый оркестрируемый процесс.

СПИСОК ЛИТЕРАТУРЫ

1. Batini C., Scannapieco M. Data and Information Quality: Dimensions, Principles and Techniques. — Cham : Springer, 2016. — 500 p.

2. Khatri V., Brown C. V. Designing data governance // Communications of the ACM. — 2010. — Vol. 53, № 1. — P. 148–152.

3. Christen P. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. — Berlin ; Heidelberg : Springer, 2012. — 270 p.

4. Грозовский Ф. М., Логинова И. В. О разработке подхода к автоматизированному сбору и интеллектуальной обработке данных с применением методов веб-скрейпинга и больших языковых моделей (на примере задачи по извлечению оценок уровней готовности технологий) // Научно-техническая информация. Сер. 2. Информационные процессы и системы. — 2025. — № 8. — С. 27–36.

5. Dagdelen J. et al. Structured information extraction from scientific text with large language models // Nature Communications. — 2024. — Vol. 15, № 1. — Article 1418.

6. Han R. et al. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors [Электрон. ресурс]. — 2023. — URL: https://arxiv.org/abs/2305.14450 (дата обращения: 06.05.2026).

7. Schwitter N. Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences // Methodological Innovations. — 2025. — Vol. 18, № 1. — P. 61–65.

8. Li H. et al. Extracting financial data from unstructured sources: Leveraging large language models // Journal of Information Systems. — 2025. — Vol. 39, № 1. — P. 135–156.

9. OpenAI API. Web search in the Responses API [Электрон. ресурс]. — URL: https://developers.openai.com/api/docs/guides/tools-web-search (дата обращения: 06.05.2026).

10. Dadata API: сервис проверки и обогащения данных о юридических лицах [Электрон. ресурс]. — URL: https://dadata.ru/api/ (дата обращения: 06.05.2026).

11. Яндекс.Геокодер API: сервис определения координат по адресу [Электрон. ресурс]. — URL: https://developer.tech.yandex.ru/services (дата обращения: 06.05.2026).

12. СПАРК-Интерфакс: система профессионального анализа рынков и компаний [Электрон. ресурс]. — URL: https://spark-interfax.ru/ (дата обращения: 06.05.2026).

13. Python Software Foundation. Python Language Reference [Электрон. ресурс]. — URL: https://www.python.org/ (дата обращения: 06.05.2026).

14. McKinney W. Data structures for statistical computing in Python // Proceedings of the 9th Python in Science Conference. — Austin, 2010. — P. 56–61.

15. Budibase: low-code platform for internal tools and web applications [Электрон. ресурс]. — URL: https://budibase.com/ (дата обращения: 06.05.2026).

16. Leaflet: an open-source JavaScript library for interactive maps [Электрон. ресурс]. — URL: https://leafletjs.com/ (дата обращения: 06.05.2026).

17. PostgreSQL: The World’s Most Advanced Open Source Relational Database [Электрон. ресурс]. — URL: https://www.postgresql.org/ (дата обращения: 06.05.2026).

18. Ефимова О. В. ESG-рейтинги: актуальные проблемы формирования и использования [Электрон. ресурс] // Экономические науки. — 2023. — № 4 (221). — С. 104–109. — URL: https://cyberleninka.ru/article/n/esg-reytingi-aktualnye-problemy-formirovaniya-i-ispolzovaniya (дата обращения: 28.05.2026).

19. Kotsantonis S., Serafeim G. Four Things No One Will Tell You About ESG Data [Электрон. ресурс] // Journal of Applied Corporate Finance. — 2019. — Vol. 31, No. 2. — P. 50–58. — URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3420297 (дата обращения: 28.05.2026).

Приложение 1. Отраслевая структура.

Сектор	Отрасль	Подотрасль
Агропродовольственный сектор	Сельское хозяйство	Растениеводство
		Животноводство
		Птицеводство и яичное производство
		Комбикорма, семеноводство, тепличные комплексы
	Пищевая промышленность	Мясная и мясоперерабатывающая
		Молочная промышленность

		Рыбная и морепродуктов
		Хлебопекарная и макаронная
		Кондитерская промышленность
		Масложировая промышленность
		Сахарная и крахмалопаточная
		Консервы
		Табак
		Напитки
Инфраструктура	Электроэнергетика	Сетевые компании
		Генерирующие компании
		Сбыт и городская инфраструктура
		Инжиниринг и оборудование
	Коммунальные услуги	Водоснабжение и водоотведение
		Теплоснабжение
		Обращение с отходами
		Эксплуатация городской инфраструктуры
		Газораспределение
	Транспорт	Публичный городской транспорт
		Шеринговая мобильность
		Пассажирская авиация
		Пассажирский водный транспорт
		Грузовые перевозки и логистика
		Транспортная инфраструктура и терминалы

		Проектирование и строительство транспортной инфраструктуры
	Телекоммуникации	Универсальные операторы связи
		Провайдеры фиксированного ШПД и кабельной инфраструктуры
		Корпоративная и магистральная связь
		Инфраструктура связи
		Спутниковая и удаленная связь
Обрабатывающая промышленность	Машиностроение	Судостроение и судоремонт
		Тракторостроение и тяжелая техника
		Приборостроение, промышленная электроника и автоматизация
		Научные исследования и разработки
		Оборудование для добывающей промышленности
		Энергетическое машиностроение и инфраструктура
		Вооружение и оружейные системы
		Авиастроение
		Автомобилестроение
		Железнодорожное машиностроение
		Арматуростроение и трубопроводная арматура
	Фармацевтика	Фармацевтика
	Легкая промышленность	Текстильная промышленность
		Производство одежды и трикотажа
		Обувная промышленность
		Кожевенная промышленность

		Прочие изделия легкой промышленности
	Мебель и товары для дома	Корпусная мебель
		Мягкая мебель
		Матрасы и товары для сна
Сырьевые отрасли	Лесная промышленность	Лесозаготовка и деревообработка
		Целлюлозно-бумажная промышленность
	Металлургия и добыча металлических руд	Черная металлургия
		Цветные металлы и сплавы
		Трубопроводная продукция
		Драгоценные металлы и камни
	Химическая промышленность	Потребительская химия и личная гигиена
		Химия для промышленности и материалов
		Агрохимия и удобрения
	Строительные материалы	Цемент и бетон
		Нерудные материалы
		Стеновые, фасадные и стекольные материалы
		Гипс, известь и отделочные материалы
		Кровельные и изоляционные материалы
	Тара и упаковка	Тара из металла, стекла и пластика
		Бумажная, картонная и пластиковая упаковка
	Топливная отрасль	Нефтегазовый сервис и геологоразведка
		Нефтедобыча и газодобыча
		Переработка нефти и газа

		Нефтепроводы и газопроводы
		Хранение и терминалы
		Оптовая торговля и АЗС
		Угольная промышленность
		Разработка, производство и продажа ядерного топлива
		Нефтехимическая промышленность
Услуги	Гостиницы и общественное питание	Гостиницы
		Кафе и рестораны
	Здравоохранение	Здравоохранение (пока не разделяем)
	Образование	Образование (пока не разделяем)
	Оптовая торговля	Оптовая торговля потребительскими товарами и техникой
		Оптовая торговля сырьем и промышленными товарами
		Фармацевтическая дистрибуция и медтовары
	Профессиональные услуги	Архитектурное и инженерно-техническое проектирование
		Разработка ПО и ИТ-услуги
		Реклама и исследование конъюнктуры рынка
		Туристические услуги
	Розничная торговля	Автозапчасти
		Аптеки и медицинские товары
		Маркетплейсы
		Непродовольственная розничная торговля
		Продовольственный ритейл

	Финансовый сектор	Банки
		Деятельность биржевых посредников и биржевых брокеров
		Инвестиционные фонды
		Кредитные бюро
		Платежные системы
		Специализированные финансовые услуги
		Страхование
		Управление холдинговыми компаниями
		Ценные бумаги

Приложение 2. Промпт для первичного сбора данных

Ты — аналитик, собирающий данные о реально существующих компаниях Северо-Западного федерального округа (СЗФО) для ESG-карты.
Твоя задача: найти компании, работающие в подотрасли «[подотрасль]» (сектор «[сектор]», отрасль «[отрасль]»), с численностью сотрудников более 100 человек. Используй веб-поиск, чтобы найти актуальную информацию на официальных сайтах компаний, в справочниках и других открытых источниках.
Для каждой компании предоставь информацию строго в формате CSV со следующими полями:
Sector, Industry, Subindustry, CompanyName, CompanyNameOfficial, INN, RegionRegistration, RegionHeadOffice, RegionOperation, Description, URL, Source
Правила:
1. Регионы — только субъекты СЗФО: Республика Карелия, Республика Коми, Архангельская область, Вологодская область, Калининградская область, Ленинградская область, Мурманская область, Новгородская область, Псковская область, г. Санкт-Петербург, Ненецкий автономный округ.
2. Компания должна быть действующей на момент поиска: не ликвидирована и не находится в стадии банкротства.
3. ИНН должен быть реальным и состоять строго из 10 или 12 цифр. Если ИНН отсутствует в источнике или вызывает сомнения, компанию не включать.
4. RegionRegistration обязательно указывается по источнику.
5. RegionHeadOffice может совпадать с регионом регистрации или отличаться, но должен относиться к СЗФО.
6. RegionOperation указывается как список регионов деятельности. Если информация отсутствует, поле остается пустым.
7. Description — краткое фактическое описание деятельности компании на основе открытых источников. Рекламные эпитеты и неподтвержденные утверждения не используются.

8. URL — официальный сайт компании, если он найден. Если сайта нет, поле остается пустым.
9. Source — конкретный источник проверки, например официальный сайт, rusprofile.ru, ЕГРЮЛ или другой открытый справочник.
10. Названия компаний приводятся к единому формату:
- CompanyNameOfficial: «ОПФ «Название»»;
- CompanyName: «Название, ОПФ»;
- ОПФ указывается аббревиатурой: ООО, АО, ПАО, МУП, ГУП и т.д.
11. CSV должен содержать заголовок, разделитель — запятая, все поля — в двойных кавычках.
12. Нужно постараться найти все реальные компании в заданной подотрасли в СЗФО. Если подходящих компаний меньше трех, вернуть столько, сколько найдено. Если подходящих компаний нет, вернуть только строку заголовков.
13. Не добавлять пояснений вне CSV.
Теперь выполни задание для:
Сектор: [сектор]
Отрасль: [отрасль]
Подотрасль: [подотрасль]
ВАЖНО: Исключи компании со следующими ИНН, они уже есть в базе данных: [контрольный список ИНН]
Не включай их в ответ, даже если они подходят под подотрасль.

СВЕДЕНИЯ ОБ АВТОРАХ (ДЛЯ РЕДАКЦИИ)

Фамилия, имя, отчество: Мясникова Дарья Романовна

Ученая степень, звание: отсутствуют

Место учебы: Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург.

Статус: студентка 1 курса; исполнитель проекта НУГ (уточнить при подаче)

Почтовый адрес: 190121, г. Санкт-Петербург, Адмиралтейский район, ул. Витебская, д. 14.

Телефон для связи: +7 911 149 46 00

Адрес электронной почты: daromiasnikova@edu.hse.ru

Фамилия, имя, отчество: Сторчевой Максим Анатольевич

Ученая степень, звание: кандидат экономических наук

Место работы: Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург.

Доцент департамента менеджмента НИУ ВШЭ в Санкт-Петербурге

Почтовый адрес: Санкт-Петербург, ул. Весельная 7/10, кв. 102.

Телефон для связи: +7 921 353 95 20

Адрес электронной почты: mstorchevoy@hse.ru